Int J Comput Vis (2017) 123:94–120
DOI 10.1007/s11263-016-0987-1

Movie Description

Anna Rohrbach[1] · Atousa Torabi[3] · Marcus Rohrbach[2] · Niket Tandon[1] · Christopher Pal[4] · Hugo Larochelle[5,6] · Aaron Courville[7] · Bernt Schiele[1]

Received: 10 May 2016 / Accepted: 23 December 2016 / Published online: 25 January 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Audio description (AD) provides linguistic descriptions of movies and allows visually impaired people to follow a movie along with their peers. Such descriptions are by design mainly visual and thus naturally form an interesting data source for computer vision and computational linguistics. In this work we propose a novel dataset which contains transcribed ADs, which are temporally aligned to full length movies. In addition we also collected and aligned movie scripts used in prior work and compare the two sources of descriptions. We introduce the *Large Scale Movie Description Challenge* (LSMDC) which contains a parallel corpus of 128,118 sentences aligned to video clips from 200 movies (around 150 h of video in total). The goal of the challenge is to automatically generate descriptions for the movie clips. First we characterize the dataset by benchmarking different approaches for generating video descriptions. Comparing ADs to scripts, we find that ADs are more visual and describe precisely what *is shown* rather than what *should happen* according to the scripts created prior to movie production. Furthermore, we present and compare the results of several teams who participated in the challenges organized in the context of two workshops at ICCV 2015 and ECCV 2016.

Keywords Movie description · Video description · Video captioning · Video understanding · Movie description dataset · Movie description challenge · Long short-term memory network · Audio description · LSMDC

Communicated by Margaret Mitchell, John Platt and Kate Saenko.

✉ Anna Rohrbach
arohrbach@mpi-inf.mpg.de

1 Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany

2 ICSI and EECS, UC Berkeley, Berkeley, CA, USA

3 Disney Research, Pittsburgh, PA, USA

4 École Polytechnique de Montréal, Montreal, Canada

5 Université de Sherbrooke, Sherbrooke, Canada

6 Twitter, Cambridge, ON, USA

7 Université de Montréal, Montreal, Canada

1 Introduction

Audio descriptions (ADs) make movies accessible to millions of blind or visually impaired people.[1] AD—sometimes also referred to as descriptive video service (DVS)—provides an audio narrative of the "most important aspects of the visual information" (Salway 2007), namely actions, gestures, scenes, and character appearance as can be seen in Figs. 1 and 2. AD is prepared by trained describers and read by professional narrators. While more and more movies are audio transcribed, it may take up to 60 person-hours to describe a 2-h movie (Lakritz and Salway 2006), resulting in the fact that today only a small subset of movies and TV programs are available for the blind. Consequently, automating this process has the potential to greatly increase accessibility to this media content.

In addition to the benefits for the blind, generating descriptions for video is an interesting task in itself, requiring the combination of core techniques from computer vision and computational linguistics. To understand the visual input one has to reliably recognize scenes, human activities, and participating objects. To generate a good description one has to

[1] In this work we refer for simplicity to "the blind" to account for all blind and visually impaired people which benefit from AD, knowing of the variety of visually impaired and that AD is not accessible to all.

AD: Abby gets in the basket.

Script: After a moment a frazzled Abby pops up in his place.

Mike leans over and sees how high they are.

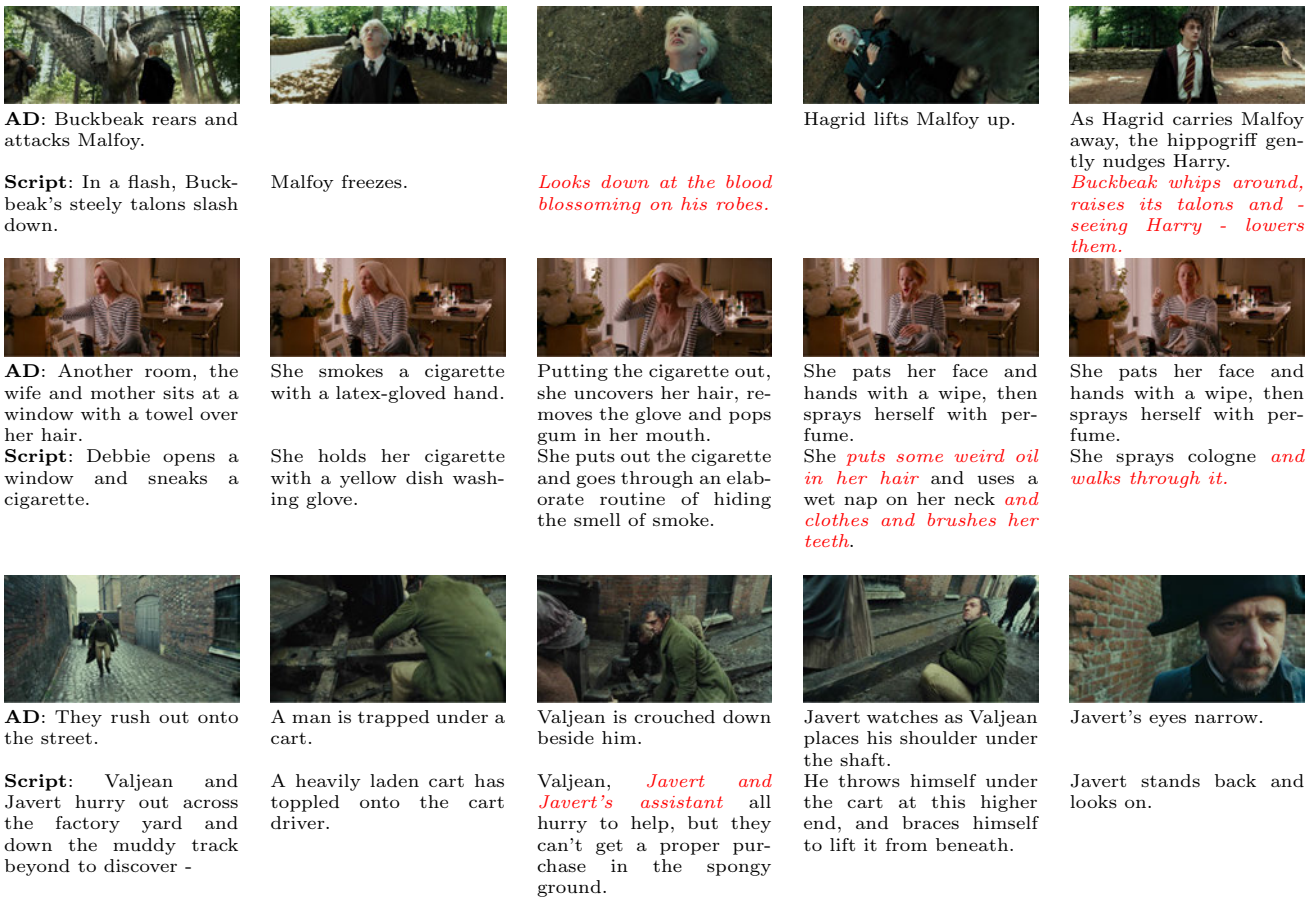
Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately.
For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

Fig. 1 Audio description (AD) and movie script samples from the movie "Ugly Truth"

decide what part of the visual information to verbalize, i.e. recognize what is salient.

Large datasets of objects (Deng et al. 2009) and scenes (Xiao et al. 2010; Zhou et al. 2014) have had an important impact in computer vision and have significantly improved our ability to recognize objects and scenes. The combination of large datasets and convolutional neural networks (CNNs) has been particularly potent (Krizhevsky et al. 2012). To be able to learn how to generate descriptions of visual content, parallel datasets of visual content paired with descriptions are indispensable (Rohrbach et al. 2013). While recently several large datasets have been released which provide images with descriptions (Young et al. 2014; Lin et al. 2014; Ordonez et al. 2011), video description datasets focus on short video clips with single sentence descriptions and have a limited number of video clips (Xu et al. 2016; Chen and Dolan 2011) or are not publicly available (Over et al. 2012). TACoS Multi-Level (Rohrbach et al. 2014) and YouCook (Das et al. 2013) are exceptions as they provide multiple sentence descriptions and longer videos. While these corpora pose challenges in terms of fine-grained recognition, they are restricted to the cooking scenario. In contrast, movies are open domain and realistic, even though, as any other video source (e.g. YouTube or surveillance videos), they have their specific characteristics. ADs and scripts associated with movies provide rich multiple sentence descriptions. They even go beyond this by telling a story which means they facilitate the study of how to extract plots, the understanding of long term semantic dependencies and human interactions from both visual and textual data.



AD: Buckbeak rears and attacks Malfoy.

Script: In a flash, Buckbeak's steely talons slash down.

Malfoy freezes.

Looks down at the blood blossoming on his robes.

Hagrid lifts Malfoy up.

As Hagrid carries Malfoy away, the hippogriff gently nudges Harry.
Buckbeak whips around, raises its talons and - seeing Harry - lowers them.

AD: Another room, the wife and mother sits at a window with a towel over her hair.
Script: Debbie opens a window and sneaks a cigarette.

She smokes a cigarette with a latex-gloved hand.

She holds her cigarette with a yellow dish washing glove.

Putting the cigarette out, she uncovers her hair, removes the glove and pops gum in her mouth.
She puts out the cigarette and goes through an elaborate routine of hiding the smell of smoke.

She pats her face and hands with a wipe, then sprays herself with perfume.
She *puts some weird oil in her hair* and uses a wet nap on her neck *and clothes and brushes her teeth*.

She pats her face and hands with a wipe, then sprays herself with perfume.
She sprays cologne *and walks through it.*

AD: They rush out onto the street.

Script: Valjean and Javert hurry out across the factory yard and down the muddy track beyond to discover -

A man is trapped under a cart.

A heavily laden cart has toppled onto the cart driver.

Valjean is crouched down beside him.

Valjean, *Javert and Javert's assistant* all hurry to help, but they can't get a proper purchase in the spongy ground.

Javert watches as Valjean places his shoulder under the shaft.
He throws himself under the cart at this higher end, and braces himself to lift it from beneath.

Javert's eyes narrow.

Javert stands back and looks on.

Fig. 2 Audio description (AD) and movie script samples from the movies "Harry Potter and the Prisoner of Azkaban", "This is 40", and "Les Miserables". Typical mistakes contained in scripts marked in *red italic*

Fig. 3 Some of the diverse verbs/actions present in our Large Scale Movie Description Challenge (LSMDC)

Figures 1 and 2 show examples of ADs and compare them to movie scripts. Scripts have been used for various tasks (Cour et al. 2008; Duchenne et al. 2009; Laptev et al. 2008; Liang et al. 2011; Marszalek et al. 2009), but so far not for video description. The main reason for this is that automatic alignment frequently fails due to the discrepancy between the movie and the script. As scripts are produced prior to the shooting of the movie they are frequently not as precise as the AD (Fig. 2 shows some typical mistakes marked in red italic). A common case is that part of the sentence is correct, while another part contains incorrect/irrelevant information. As can be seen in the examples, AD narrations describe key visual elements of the video such as changes in the scene, people's appearance, gestures, actions, and their interaction with each other and the scene's objects in concise and precise language. Figure 3 shows the variability of AD data w.r.t. to verbs (actions) and corresponding scenes from the movies.

In this work we present a dataset which provides transcribed ADs, aligned to full length movies. AD narrations are carefully positioned within movies to fit in the natural pauses in the dialogue and are mixed with the original movie soundtrack by professional post-production. To obtain ADs we retrieve audio streams from DVDs/Blu-ray disks, segment out the sections of the AD audio and transcribe them via a crowd-sourced transcription service. The ADs provide an initial temporal alignment, which however does not always cover the full activity in the video. We discuss a way to fully automate both audio-segmentation and temporal alignment, but also manually align each sentence to the movie for all the data. Therefore, in contrast to Salway (2007) and Salway et al. (2007), our dataset provides alignment to the actions in the video, rather than just to the audio track of the description.

In addition we also mine existing movie scripts, pre-align them automatically, similar to Cour et al. (2008) and Laptev et al. (2008), and then manually align the sentences to the movie.

As a first study on our dataset we benchmark several approaches for movie description. We first examine nearest neighbor retrieval using diverse visual features which do not require any additional labels, but retrieve sentences from the training data. Second, we adapt the translation approach of Rohrbach et al. (2013) by automatically extracting an intermediate semantic representation from the sentences using semantic parsing. Third, based on the success of long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber 1997) for the image captioning problem (Donahue et al. 2015; Karpathy and Fei-Fei 2015; Kiros et al. 2015; Vinyals et al. 2015) we propose our approach *Visual-Labels*. It first builds robust visual classifiers which distinguish verbs, objects, and places extracted from weak sentence annotations. Then the visual classifiers form the input to an LSTM for generating movie descriptions.

The main contribution of this work is the Large Scale Movie Description Challenge (LSMDC)[2] which provides transcribed and aligned AD and script data sentences. The LSMDC was first presented at the Workshop "Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC)", collocated with ICCV 2015. The second edition, LSMDC 2016, was presented at the "Joint Workshop on Storytelling with Images and Videos and Large Scale Movie Description and Understanding Challenge", collocated with ECCV 2016. Both challenges include the same public and blind test sets with an evaluation server[3] for automatic evaluation. LSMDC is based on the MPII Movie Description dataset (MPII-MD) and the Montreal Video Annotation Dataset (M-VAD) which were initially collected independently but are presented jointly in this work. We detail the data collection and dataset properties in Sect. 3, which includes our approach to automatically collect and align AD data. In Sect. 4 we present several benchmark approaches for movie description, including our

[2] https://sites.google.com/site/describingmovies/.

[3] https://competitions.codalab.org/competitions/6121.

Visual-Labels approach which learns robust visual classifiers and generates description using an LSTM. In Sect. 5 we present an evaluation of the benchmark approaches on the M-VAD and MPII-MD datasets, analyzing the influence of the different design choices. Using automatic and human evaluation, we also show that our Visual-Labels approach outperforms prior work on both datasets. In Sect. 5.5 we perform an analysis of prior work and our approach to understand the challenges of the movie description task. In Sect. 6 we present and discuss the results of the LSMDC 2015 and LSMDC 2016.

This work is partially based on the original publications from Rohrbach et al. (2015c, b) and the technical report from Torabi et al. (2015). Torabi et al. (2015) collected M-VAD, Rohrbach et al. (2015c) collected the MPII-MD dataset and presented the translation-based description approach. Rohrbach et al. (2015b) proposed the Visual-Labels approach.

2 Related Work

We discuss recent approaches to image and video description including existing work using movie scripts and ADs. We also discuss works which build on our dataset. We compare our proposed dataset to related video description datasets in Table 3 (Sect. 3.5).

2.1 Image Description

Prior work on image description includes Farhadi et al. (2010), Kulkarni et al. (2011), Kuznetsova et al. (2012, 2014), Li et al. (2011), Mitchell et al. (2012) and Socher et al. (2014). Recently image description has gained increased attention with work such as that of Chen and Zitnick (2015), Donahue et al. (2015), Fang et al. (2015), Karpathy and Fei-Fei (2015), Kiros et al. (2014, 2015), Mao et al. (2015), Vinyals et al. (2015) and Xu et al. (2015a). Much of the recent work has relied on Recurrent Neural Networks (RNNs) and in particular on long short-term memory networks (LSTMs). New datasets have been released, such as the Flickr30k (Young et al. 2014) and MS COCO Captions (Chen et al. 2015), where Chen et al. (2015) also presents a standardized protocol for image captioning evaluation. Other work has analyzed the performance of recent methods, e.g. Devlin et al. (2015) compare them with respect to the novelty of generated descriptions, while also exploring a nearest neighbor baseline that improves over recent methods.

2.2 Video Description

In the past video description has been addressed in controlled settings (Barbu et al. 2012; Kojima et al. 2002), on a small scale (Das et al. 2013; Guadarrama et al. 2013; Thomason et al. 2014) or in single domains like cooking (Rohrbach et al. 2014, 2013; Donahue et al. 2015). Donahue et al. (2015) first proposed to describe videos using an LSTM, relying on pre-computed CRF scores from Rohrbach et al. (2014). Later Venugopalan et al. (2015c) extended this work to extract CNN features from frames which are max-pooled over time. Pan et al. (2016b) propose a framework that consists of a 2-/3-D CNN and LSTM trained jointly with a visual-semantic embedding to ensure better coherence between video and text. Xu et al. (2015b) jointly address the language generation and video/language retrieval tasks by learning a joint embedding for a deep video model and a compositional semantic language model. Li et al. (2015) study the problem of summarizing a long video to a single concise description by using ranking based summarization of multiple generated candidate sentences.

Concurrent and Consequent Work To handle the challenging scenario of movie description, Yao et al. (2015) propose a soft-attention based model which selects the most relevant temporal segments in a video, incorporates 3-D CNN and generates a sentence using an LSTM. Venugopalan et al. (2015b) propose S2VT, an encoder–decoder framework, where a single LSTM encodes the input video frame by frame and decodes it into a sentence. Pan et al. (2016a) extend the video encoding idea by introducing a second LSTM layer which receives input of the first layer, but skips several frames, reducing its temporal depth. Venugopalan et al. (2016) explore the benefit of pre-trained word embeddings and language models for generation on large external text corpora. Shetty and Laaksonen (2015) evaluate different visual features as input for an LSTM generation frame-work. Specifically they use dense trajectory features (Wang et al. 2013) extracted for the clips and CNN features extracted at center frames of the clip. They find that training concept classifiers on MS COCO with the CNN features, combined with dense trajectories provides the best input for the LSTM. Ballas et al. (2016) leverages multiple convolutional maps from different CNN layers to improve the visual representation for activity and video description. To model multi-sentence description, Yu et al. (2016a) propose to use two stacked RNNs where the first one models words within a sentence and the second one, sentences within a paragraph. Yao et al. (2016) has conducted an interesting study on performance upper bounds for both image and video description tasks on available datasets, including the LSMDC dataset.

2.3 Movie Scripts and Audio Descriptions

Movie scripts have been used for automatic discovery and annotation of scenes and human actions in videos (Duchenne et al. 2009; Laptev et al. 2008; Marszalek et al. 2009),

as well as a resource to construct activity knowledge base (Tandon et al. 2015; de Melo and Tandon 2016). We rely on the approach presented by Laptev et al. (2008) to align movie scripts using subtitles.

Bojanowski et al. (2013) approach the problem of learning a joint model of actors and actions in movies using weak supervision provided by scripts. They rely on the semantic parser SEMAFOR (Das et al. 2012) trained on the FrameNet database (Baker et al. 1998), however, they limit the recognition only to two frames. Bojanowski et al. (2014) aim to localize individual short actions in longer clips by exploiting the ordering constraints as weak supervision. Bojanowski et al. (2013, 2014), Duchenne et al. (2009), Laptev et al. (2008), Marszalek et al. (2009) proposed datasets focused on extracting several activities from movies. Most of them are part of the "Hollywood2" dataset (Marszalek et al. 2009) which contains 69 movies and 3669 clips. Another line of work (Cour et al. 2009; Everingham et al. 2006; Ramanathan et al. 2014; Sivic et al. 2009; Tapaswi et al. 2012) proposed datasets for character identification targeting TV shows. All the mentioned datasets rely on alignments to movie/TV scripts and none uses ADs.

ADs have also been used to understand which characters interact with each other (Salway et al. 2007). Other prior work has looked at supporting AD production using scripts as an information source (Lakritz and Salway 2006) and automatically finding scene boundaries (Gagnon et al. 2010). Salway (2007) analyses the linguistic properties on a non-public corpus of ADs from 91 movies. Their corpus is based on the original sources to create the ADs and contains different kinds of artifacts not present in actual description, such as dialogs and production notes. In contrast, our text corpus is much cleaner as it consists only of the actual ADs.

2.4 Works Building on Our Dataset

Interestingly, other works, datasets, and challenges are already building upon our data. Zhu et al. (2015b) learn a visual-semantic embedding from our clips and ADs to relate movies to books. Bruni et al. (2016) also learn a joint embedding of videos and descriptions and use this representation to improve activity recognition on the Hollywood 2 dataset Marszalek et al. (2009). Tapaswi et al. (2016) use our AD transcripts for building their MovieQA dataset, which asks natural language questions about movies, requiring an understanding of visual and textual information, such as dialogue and AD, to answer the question. Zhu et al. (2015a) present a fill-in-the-blank challenge for audio description of the current, previous, and next sentence description for a given clip, requiring to understand the temporal context of the clips.

3 Datasets for Movie Description

In the following, we present how we collect our data for movie description and discuss its properties. The Large Scale Movie Description Challenge (LSMDC) is based on two datasets which were originally collected independently. The MPII Movie Description Dataset (MPII-MD), initially presented by Rohrbach et al. (2015c), was collected from Blu-ray movie data. It consists of AD and script data and uses sentence-level manual alignment of transcribed audio to the actions in the video (Sect. 3.1). In Sect. 3.2 we discuss how to fully automate AD audio segmentation and alignment for the Montreal Video Annotation Dataset (M-VAD), initially presented by Torabi et al. (2015). M-VAD was collected with DVD data quality and only relies on AD. Section 3.3 details the Large Scale Movie Description Challenge (LSMDC) which is based on M-VAD and MPII-MD, but also contains additional movies, and was set up as a challenge. It includes a submission server for evaluation on public and blind test sets. In Sect. 3.4 we present the detailed statistics of our datasets, also see Table 1. In Sect. 3.5 we compare our movie description data to other video description datasets.

3.1 The MPII Movie Description (MPII-MD) Dataset

In the following we describe our approach behind the collection of ADs (Sect. 3.1.1) and script data (Sect. 3.1.2). Then we discuss how to manually align them to the video (Sect. 3.1.3) and which visual features we extracted from the video (Sect. 3.1.4).

3.1.1 Collection of ADs

We search for Blu-ray movies with ADs in the "Audio Description" section of the British Amazon[4] and select 55 movies of diverse genres (e.g. drama, comedy, action). As ADs are only available in audio format, we first retrieve the audio stream from the Blu-ray HD disks. We use MakeMKV[5] to extract a Blu-ray in the .mkv file format, and then XMediaRecode[6] to select and extract the audio streams from it. Then we semi-automatically segment out the sections of the AD audio (which is mixed with the original audio stream) with the approach described below. The audio segments are then transcribed by a crowd-sourced transcription service[7] that also provides us the time-stamps for each spoken sentence.

[4] www.amazon.co.uk.

[5] https://www.makemkv.com/.

[6] https://www.xmedia-recode.de/.

[7] CastingWords transcription service, http://castingwords.com/.

Table 1 Movie description dataset statistics, see discussion in Sect. 3.4; for average/total length we report the "2-seconds-expanded" alignment, used in this work, and an actual manual alignment in brackets

	Unique movies	Words	Sentences	Clips	Average length (s)	Total length (h)
MPII-MD (AD)	55	330,086	37,272	37,266	4.2 (4.1)	44.0 (42.5)
MPII-MD (movie script)	50	317,728	31,103	31,071	3.9 (3.6)	33.8 (31.1)
MPII-MD (total)	94	647,814	68,375	68,337	4.1 (3.9)	77.8 (73.6)
M-VAD (AD)	92	502,926	55,904	46,589	6.2	84.6
LSMDC 15 training	153	914.327	91,941	91,908	4.9 (4.8)	124.9 (121.4)
LSMDC 15 validation	12	63,789	6542	6542	5.3 (5.2)	9.6 (9.4)
LSMDC 15 and 16 public test	17	87,150	10,053	10,053	4.2 (4.1)	11.7 (11.3)
LSMDC 15 and 16 blind test	20	83,766	9578	9578	4.5 (4.4)	12.0 (11.8)
LSMDC 15 (total)	200	1,149,032	118,114	118,081	4.8 (4.7)	158.1 (153.9)
LSMDC 16 training	153	922,918	101,079	101,046	4.1 (3.9)	114.9 (109.7)
LSMDC 16 validation	12	63,321	7408	7408	4.1 (3.9)	8.4 (8.0)
LSMDC 15 and 16 public test	17	87,150	10,053	10,053	4.2 (4.1)	11.7 (11.3)
LSMDC 15 and 16 blind test	20	83,766	9578	9578	4.5 (4.4)	12.0 (11.8)
LSMDC 16 (Total)	200	1,157,155	128,118	128,085	4.1 (4.0)	147.0 (140.8)

Semi-automatic Segmentation of ADs We are given two audio streams: the original audio and the one mixed with the AD. We first estimate the temporal alignment between the two as there might be a few time frames difference. The precise alignment is important to compute the similarity of both streams. Both steps (alignment and similarity) are estimated using the spectograms of the audio stream, which is computed using a Fast Fourier Transform (FFT). If the difference between the two audio streams is larger than a given threshold we assume the mixed stream contains AD at that point in time. We smooth this decision over time using a minimum segment length of 1 s. The threshold was picked on a few sample movies, but had to be adjusted for each movie due to different mixing of the AD stream, different narrator voice level, and movie sound. While we found this semi-automatic approach sufficient when using a further manual alignment, we describe a fully automatic procedure in Sect. 3.2.

3.1.2 Collection of Script Data

In addition to the ADs we mine script web resources[8] and select 39 movie scripts. As starting point we use the movie scripts from "Hollywood2" (Marszalek et al. 2009) that have highest alignment scores to their movie. We are also interested in comparing the two sources (movie scripts and ADs), so we are looking for the scripts labeled as "Final", "Shooting", or "Production Draft" where ADs are also available. We found that the "overlap" is quite narrow, so we analyze 11 such movies in our dataset. This way we end up with 50 movie scripts in total. We follow existing approaches (Cour et al. 2008; Laptev et al. 2008) to automatically align scripts to movies. First we parse the scripts, extending the method of Laptev et al. (2008) to handle scripts which deviate from the default format. Second, we extract the subtitles from the Blu-ray disks with SubtitleEdit.[9] It also allows for subtitle alignment and spellchecking. Then we use the dynamic programming method of Laptev et al. (2008) to align scripts to subtitles and infer the time-stamps for the description sentences. We select the sentences with a reliable alignment score (the ratio of matched words in the near-by monologues) of at least 0.5. The obtained sentences are then manually aligned to video in-house.

3.1.3 Manual Sentence-Video Alignment

As the AD is added to the original audio stream between the dialogs, there might be a small misalignment between the time of speech and the corresponding visual content. Therefore, we manually align each sentence from ADs and scripts to the movie in-house. During the manual alignment we also filter out: (a) sentences describing movie introduction/ending (production logo, cast, etc); (b) texts read from the screen; (c) irrelevant sentences describing something not present in the video; (d) sentences related to audio/sounds/music. For the movie scripts, the reduction in number of words is about 19%, while for ADs it is under 4%. In the case of ADs, filtering mainly happens due to initial/ending movie intervals and transcribed dialogs (when shown as text). For the scripts, it is mainly attributed to irrelevant sentences. Note that we retain the sentences that are "alignable" but contain minor

[8] http://www.weeklyscript.com, http://www.simplyscripts.com, http://www.dailyscript.com, http://www.imsdb.com.

[9] www.nikse.dk/SubtitleEdit/.

mistakes. If the manually aligned video clip is shorter than 2 s, we symmetrically expand it (from beginning and end) to be exactly 2 s long. In the following we refer to the obtained alignment as a "2-seconds-expanded" alignment.

3.1.4 Visual Features

We extract video clips from the full movie based on the aligned sentence intervals. We also uniformly extract 10 frames from each video clip. As discussed earlier, ADs and scripts describe activities, objects and scenes (as well as emotions which we do not explicitly handle with these features, but they might still be captured, e.g. by the context or activities). In the following we briefly introduce the visual features computed on our data which are publicly available.[10]

IDT We extract the improved dense trajectories compensated for camera motion (Wang and Schmid 2013). For each feature (Trajectory, HOG, HOF, MBH) we create a codebook with 4,000 clusters and compute the corresponding histograms. We apply L1 normalization to the obtained histograms and use them as features.

LSDA We use the recent large scale object detection CNN (Hoffman et al. 2014) which distinguishes 7604 ImageNet (Deng et al. 2009) classes. We run the detector on every second extracted frame (due to computational constraints). Within each frame we max-pool the network responses for all classes, then do mean-pooling over the frames within a video clip and use the result as a feature.

PLACES and HYBRID Finally, we use the recent scene classification CNNs (Zhou et al. 2014) featuring 205 scene classes. We use both available networks, *Places-CNN* and *Hybrid-CNN*, where the first is trained on the Places dataset (Zhou et al. 2014) only, while the second is additionally trained on the 1.2 million images of ImageNet (ILSVRC 2012) (Russakovsky et al. 2015). We run the classifiers on all the extracted frames of our dataset. We mean-pool over the frames of each video clip, using the result as a feature.

3.2 The Montreal Video Annotation Dataset (M-VAD)

One of the main challenges in automating the construction of a video annotation dataset derived from AD audio is accurately segmenting the AD output, which is mixed with the original movie soundtrack. In Sect. 3.1.1 we have introduced a way of semi-automatic AD segmentation. In this section we describe a fully automatic method for AD narration isolation and video alignment. AD narrations are typically carefully placed within key locations of a movie and edited by a post-production supervisor for continuity. For example, when a scene changes rapidly, the narrator will speak multiple sentences without pauses. Such content should be kept together when describing that part of the movie. If a scene changes slowly, the narrator will instead describe the scene in one sentence, then pause for a moment, and later continue the description. By detecting those short pauses, we are able to align a movie with video descriptions automatically.

In the following we describe how we select the movies with AD for our dataset (Sect. 3.2.1) and detail our automatic approach to AD segmentation (Sect. 3.2.2). In Sect. 3.2.3 we discuss how to align AD to the video and obtain high quality AD transcripts.

3.2.1 Collection of ADs

To search for movies with AD we use the movie lists provided in "An Initiative of the American Council of the Blind" [11] and "Media Access Group at WGBH"[12] websites, and buy them based on their availability and price. To extract video and audio from the DVDs we use the DVDfab[13] software.

3.2.2 AD Narrations Segmentation Using Vocal Isolation

Despite the advantages offered by AD, creating a completely automated approach for extracting the relevant narration or annotation from the audio track and refining the alignment of the annotation with the video still poses some challenges. In the following, we discuss our automatic solution for AD narrations segmentation. We use two audio tracks included in DVDs: (1) the standard movie audio signal and (2) the standard movie audio mixed with AD narrations signal.

Vocal isolation techniques boost vocals, including dialogues and AD narrations while suppressing background movie sound in stereo signals. This technique is used widely in karaoke machines for stereo signals to remove the vocal track by reversing the phase of one channel to cancel out any signal perceived to come from the center while leaving the signals that are perceived as coming from the left or the right. The main reason for using vocal isolation for AD segmentation is based on the fact that AD narration is mixed in natural pauses in the dialogue. Hence, AD narration can only be present when there is no dialogue. In vocal isolated signals, whenever the narrator speaks, the movie signal is almost a flat line relative to the AD signal, allowing us to cleanly separate the narration by comparing the two signals. Figure 4 illustrates an example from the movie "Life of Pi", where in the original movie soundtrack there are sounds of ocean waves in the background.

Our approach has three main steps. First we isolate vocals, including dialogues and AD narrations. Second, we separate

[10] mpii.de/movie-description.

[11] http://www.acb.org/adp/movies.html.

[12] http://main.wgbh.org/wgbh/pages/mag/dvsondvd.html.

[13] http://www.dvdfab.cn/.

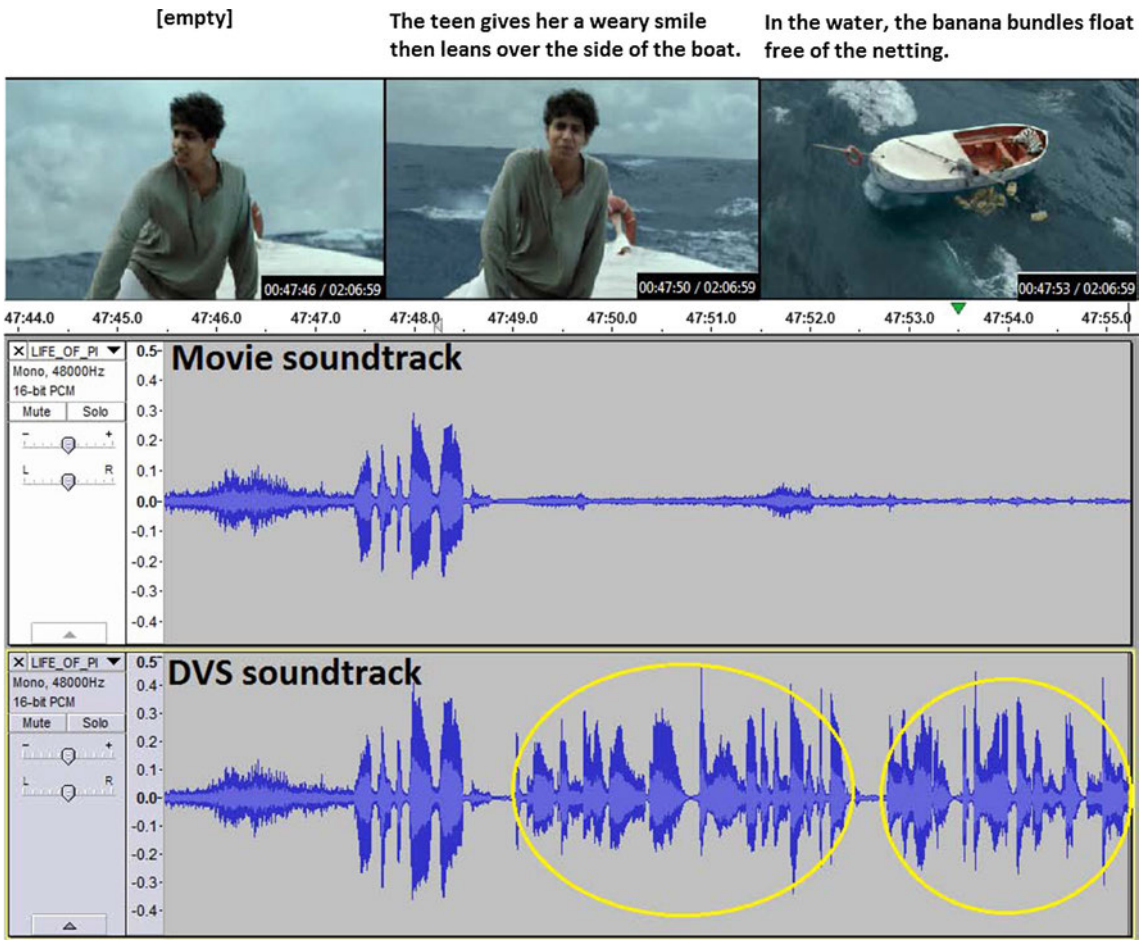


Fig. 4 AD dataset collection. From the movie "Life of Pi". Line 2 and 3: Vocal isolation of movie and AD soundtrack. Second and third rows shows movie and AD audio signals after voice isolation. The *two circles* show the AD segments on the AD mono channel track. A pause (flat signal) between two AD narration parts shows the natural AD narration segmentation while the narrator stops and then continues describing the movie. We automatically segment AD audio based on these natural pauses. At first row, you can also see the transcription related to first and second AD narration parts on top of second and third image shots

the AD narrations from dialogues. Finally, we apply a simple thresholding method to extract AD segment audio tracks.

We isolate vocals using Adobe Audition's center channel extractor[14] implementation to boost AD narrations and movie dialogues while suppressing movie background sounds on both AD and movie audio signals. We align the movie and AD audio signals by taking an FFT of the two audio signals, compute the cross-correlation, measure similarity for different offsets and select the offset which corresponds to peak cross-correlation. After alignment, we apply Least Mean Square (LMS) noise cancellation and subtract the AD mono squared signal from the movie mono squared signal in order to suppress dialogue in the AD signal. For the majority of movies on the market (among the 104 movies that we purchased, 12 movies have been mixed to the center of the audio signal, therefore we were not able to automatically align them), applying LMS results in cleaned AD narrations for the AD audio signal. Even in cases where the shapes of the standard movie audio signal and standard movie audio mixed with AD signal are very different—due to the AD mixing process—our procedure is sufficient for the automatic segmentation of AD narration.

Finally, we extract the AD audio tracks by detecting the beginning and end of AD narration segments in the AD audio signal (i.e. where the narrator starts and stops speaking) using a simple thresholding method that we applied to all DVDs without changing the threshold value. This is in contrast to the semi-automatic approach presented in Sect. 3.1.1, which requires individual adjustment of a threshold for each movie.

3.2.3 Movie/AD Alignment and Professional Transcription

AD audio narration segments are time-stamped based on our automatic AD narration segmentation. In order to compen-

[14] creative.adobe.com/products/audition.

Table 2 Vocabulary and POS statistics (after word stemming) for our movie description datasets, see discussion in Sect. 3.4

Dataset	Vocab. size	Nouns	Verbs	Adjectives-	Adverbs
MPII-MD	18,871	10,558	2933	4239	1141
M-VAD	17,609	9512	2571	3560	857
LSMDC 15	22,886	12,427	3461	5710	1288
LSMDC 16	22,500	12,181	3394	5633	1292

sate for the potential 1–2 s misalignment between the AD narrator speaking and the corresponding scene in the movie, we automatically add 2 s to the end of each video clip. Also we discard all the transcriptions related to movie introduction/ending which are located at the beginning and the end of movies.

In order to obtain high quality text descriptions, the AD audio segments were transcribed with more than 98% transcription accuracy, using a professional transcription service.[15] These services use a combination of automatic speech recognition techniques and human transcription to produce a high quality transcription. Our audio narration isolation technique allows us to process the audio into small, well defined time segments and reduce the overall transcription effort and cost.

3.3 The Large Scale Movie Description Challenge (LSMDC)

To build our Large Scale Movie Description Challenge (LSMDC), we combine the M-VAD and MPII-MD datasets. We first identify the overlap between the two, so that the same movie does not appear in the training and test set of the joined dataset. We also exclude script-based movie alignments from the validation and test sets of MPII-MD. The datasets are then joined by combining the corresponding training, validation and test sets, see Table 1 for detailed statistics. The combined test set is used as a *public* test set of the challenge. We additionally acquired 20 more movies where we only release the video clips, but not the aligned sentences. They form the *blind* test set of the challenge and are only used for evaluation. We rely on the respective best aspects of M-VAD and MPII-MD for the public and blind test sets: we provide Blu-ray quality for them, use the automatic alignment/ transcription described in Sect. 3.2 and clean them using a manual alignment as in Sect. 3.1.3. For the second edition of our challenge, LSMDC 2016, we also manually align the M-VAD validation and training sets and release them with Blu-ray quality. The manual alignment results in many multi-sentences descriptions to be split. Also the more precise alignment reduces the average clip length.

We set up the evaluation server[3] for the challenge using the Codalab[16] platform. The challenge data is available online[2]. We provide more information about the challenge setup and results in Sect. 6.

In addition to the description task, LSMDC 2016 includes three additional tracks, not discussed in this work. There is a movie annotation track which asks to select the correct sentence out of five in a multiple-choice test, a retrieval track which asks to retrieve the correct test clip for a given sentence, and a fill-in-the-blank track which requires to predict a missing word in a given description and the corresponding clip. The data and more details can be found on our web site[2]; Torabi et al. (2016) provide more details about the annotation and the retrieval tasks.

3.4 Movie Description Dataset Statistics

Table 1 presents statistics for the number of words, sentences and clips in our movie description corpora. We also report the average/total length of the annotated time intervals. We report both, the "2-seconds-expanded" clip alignment (see Sect. 3.1.3) and the actual clip alignment in brackets. In total MPII-MD contains 68,337 clips and 68,375 sentences (rarely multiple sentences might refer to the same video clip), while M-VAD includes 46,589 clips and 55,904 sentences.

Our combined LSMDC 2015 dataset contains over 118 K sentence-clips pairs and 158 h of video. The training/validation/public-/blind-test sets contain 91,908, 6542, 10,053 and 9578 video clips respectively. This split balances movie genres within each set, which is motivated by the fact that the vocabulary used to describe, say, an action movie could be very different from the vocabulary used in a comedy movie. After manual alignment of the training/validation sets, the new LSMDC 2016 contains 101,046 training clips, 7408 validation clips and 128 K clips in total.

Table 2 illustrates the vocabulary size, number of nouns, verbs, adjectives, and adverbs in each respective dataset. To compute the part of speech statistics for our corpora we tag and stem all words in the datasets with the Standford Part-Of-Speech (POS) tagger and stemmer toolbox (Toutanova et al.

[15] TranscribeMe professional transcription, http://transcribeme.com.

[16] https://codalab.org/.

Table 3 Comparison of video description datasets; for discussion see Sect. 3.5

Dataset	Multisentence	Domain	Sentence source	Videos	Clips	Sentences	Length (h)
YouCook (Das et al. 2013)	x	Cooking	Crowd	88	–	2668	2.3
TACoS (Regneri et al. 2013)	x	Cooking	Crowd	127	7206	18, 227	10.1
TACoS Multi-Level (Rohrbach et al. 2014)	x	Cooking	crowd	185	24,764	74, 828	15.8
MSVD (Chen and Dolan 2011)		Open	Crowd	–	1970	70, 028	5.3
TGIF (Li et al. 2016)		Open	Crowd	–	100,000	125, 781	≈86.1
MSR-VTT (Xu et al. 2016)		Open	Crowd	7180	10,000	200, 000	41.2
VTW (Zeng et al. 2016)	x	Open	Crowd/profess.	18,100	–	44, 613	213.2
M-VAD (ours)	x	Open	Professional	92	46,589	55, 904	84.6
MPII-MD (ours)	x	Open	Professional	94	68,337	68, 375	77.8
LSMDC 15 (ours)	x	Open	Professional	200	118,081	118, 114	158.1
LSMDC 16 (ours)	x	Open	Professional	200	128,085	128, 118	147.0

2003), then we compute the frequency of stemmed words in the corpora. It is important to notice that in our computation each word and its variations in corpora is counted once since we applied stemmer. Interesting observation on statistics is that e.g. the number of adjectives is larger than the number of verbs, which shows that the AD is describing the characteristics of visual elements in the movie in high detail.

3.5 Comparison to Other Video Description Datasets

We compare our corpus to other existing parallel video corpora in Table 3. We look at the following properties: availability of multi-sentence descriptions (long videos described continuously with multiple sentences), data domain, source of descriptions and dataset size. The main limitations of prior datasets include the coverage of a single domain (Das et al. 2013; Regneri et al. 2013; Rohrbach et al. 2014) and having a limited number of video clips (Chen and Dolan 2011). Recently, a few video description datasets have been proposed, namely MSR-VTT (Xu et al. 2016), TGIF (Li et al. 2016) and VTW (Zeng et al. 2016). Similar to MSVD dataset (Chen and Dolan 2011), MSR-VTT is based on YouTube clips. While it has a large number of sentence descriptions (200K) it is still rather small in terms of the number of video clips (10 K). TGIF is a large dataset of 100 k image sequences (GIFs) with associated descriptions. VTW is a dataset which focuses on longer YouTube videos (1.5 min on average) and aims to generate concise video titles from user provided descriptions as well as editor provided titles. All these datasets are similar in that they contain web-videos, while our proposed dataset focuses on movies. Similar to e.g. VTW, our dataset has a "multi-sentence" property, making it possible to study multi-sentence description or understanding stories and plots.

4 Approaches for Movie Description

Given a training corpus of aligned videos and sentences we want to describe a new unseen test video. In this section we discuss two approaches to the video description task that we benchmark on our proposed datasets. Our first approach in Sect. 4.1 is based on the statistical machine translation (SMT) approach of Rohrbach et al. (2013). Our second approach (Sect. 4.2) learns to generate descriptions using long short-term memory network (LSTM). For the first step both approaches rely on visual classifiers learned on annotations (labels) extracted from natural language descriptions using our semantic parser (Sect. 4.1.1). While the first approach does not differentiate which features to use for different labels, our second approach defines different semantic groups of labels and uses most relevant visual features for each group. For this reason we refer to this approach as *Visual-Labels*. Next, the first approach uses the classifier scores as input to a CRF to predict a semantic representation (SR) (SUBJECT, VERB, OBJECT, LOCATION), and then translates it into a sentence with SMT. On the other hand, our second approach directly provides the classifier scores as input to an LSTM which generates a sentence based on them. Figure 5 shows an overview of the two discussed approaches.

4.1 Semantic Parsing + SMT

As our first approach we adapt the two-step translation approach of Rohrbach et al. (2013). As a first step it trains the visual classifiers based on manually annotated tuples e.g. $\langle cut, knife, tomato \rangle$ provided with the video. Then it trains a CRF which aims to predict such tuple, or semantic representation (SR), from a video clip. At a second step, the Statistical Machine Translation (SMT) (Koehn et al. 2007) is used to translate the obtained SR into a natural language sentence, e.g. "The person cuts a tomato with a knife", see

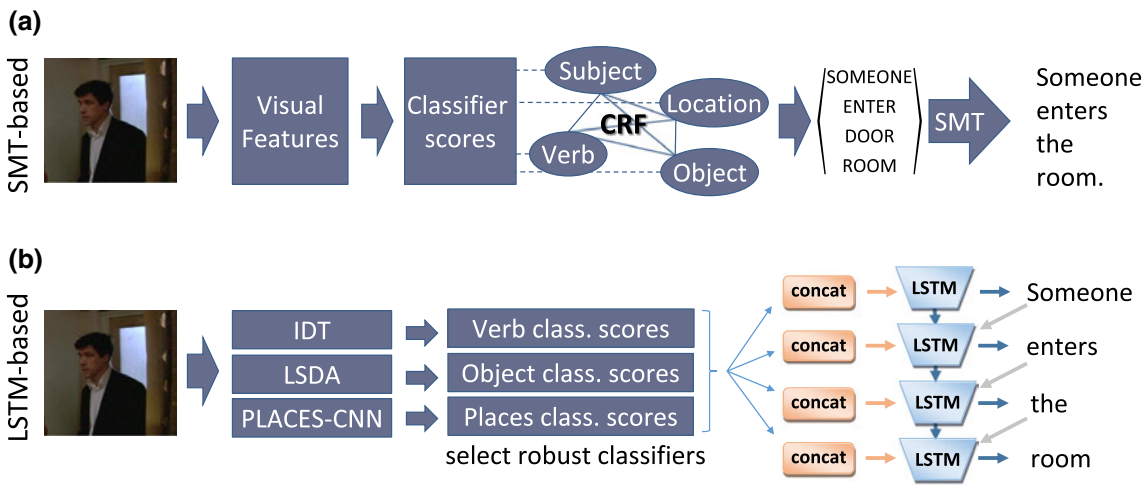


Fig. 5 Overview of our movie description approaches: **a** SMT-based approach, adapted from Rohrbach et al. (2013), **b** our proposed LSTM-based approach

Fig. 5a. While we cannot rely on a manually annotated SR as in Rohrbach et al. (2013), we automatically mine the SR from sentences using semantic parsing which we introduce in this section.

4.1.1 Semantic Parsing

Learning from a parallel corpus of videos and natural language sentences is challenging when no annotated intermediate representation is available. In this section we introduce our approach to exploit the sentences using semantic parsing. The proposed method automatically extracts intermediate semantic representations (SRs) from the natural sentences.

Approach We lift the words in a sentence to a semantic space of roles and WordNet (Fellbaum 1998) senses by performing SRL (Semantic Role Labeling) and WSD (Word Sense Disambiguation). For an example, refer to Table 4 where the desired outcome of SRL and WSD on the input sentence "*He shot a video in the moving bus*" is "$\texttt{Agent}$: $\texttt{man}_n^1$, $\texttt{Action}$: $\texttt{shoot}_v^4$, $\texttt{Patient}$: $\texttt{video}_n^2$, $\texttt{Location}$: $\texttt{bus}_n^1$". Here, e.g. $\texttt{shoot}_v^4$ refers to the fourth verb sense of $\texttt{shoot}$ in WordNet.[17] This is similar to the semantic representation of Rohrbach et al. (2013), except that those semantic frames were constructed manually while we construct them automatically and our role fillers are additionally sense disambiguated. As verbs are known to have high ambiguity, the disambiguation step will provide clearer representations (corresponding WordNet sense) of a large set of verbs present in movie descriptions.

We start by decomposing the typically long sentences present in movie descriptions into smaller clauses using the ClausIE tool (Del Corro and Gemulla 2013). For example, "*he shot and modified the video*" is split into two clauses "*he shot the video*" and "*he modified the video*"). We then use the OpenNLP tool suite[18] to chunk every clause into phrases. These chunks are disambiguated to their WordNet senses[17] by enabling a state-of-the-art WSD system called IMS (Zhong and Ng 2010), to additionally disambiguate phrases that are not present in WordNet and thus, out of reach for IMS. We identify and disambiguate the head word of an out of WordNet phrase, e.g. $\texttt{the moving bus}$ to the proper WordNet sense $\texttt{bus}_n^1$ via IMS. In this way we make an extension to IMS so it works for phrases and not just words. We link verb phrases to the proper sense of its head word in WordNet (e.g. $\texttt{begin to shoot}$ to $\texttt{shoot}_v^4$). The phrasal verbs such as e.g. *"pick up"* or *"turn off"* are preserved as long as they exist in WordNet.

Having estimated WordNet senses for the words and phrases, we need to assign semantic role labels to them. Typical SRL systems require large amounts of training data, which we do not possess for the movie domain. Therefore, we propose leveraging VerbNet (Kipper et al. 2006; Schuler et al. 2009), a manually curated high-quality linguistic resource for English verbs that supplements WordNet verb senses with syntactic frames and semantic roles, as a distant signal to assign role labels. Every VerbNet verb sense comes with a syntactic frame e.g. for $\texttt{shoot}_v^4$, the syntactic frame is $\texttt{NP V NP}$. VerbNet also provides a role restriction on the arguments of the roles e.g. for

[17] The WordNet senses for *shoot* and *video* are:

- $\texttt{shoot}_v^1$: hit with missile … $\texttt{video}_n^1$: picture in TV
- $\texttt{shoot}_v^2$: kill by missile … $\texttt{video}_n^2$: a recording …
- … …
- $\texttt{shoot}_v^4$: make a film … $\texttt{video}_n^4$: broadcasting …

where, $\texttt{shoot}_v^1$ refers to the first verb (v) sense of $\texttt{shoot}$.

[18] OpenNLP tool suite: http://opennlp.sourceforge.net/.

Table 4 Semantic parse for "*He began to shoot a video in the moving bus*"; for discussion, see Sect. 4.1.1

Phrase	WordNet Mapping	VerbNet Mapping	Desired Frame
the man	$\texttt{man}_n^1$	`Agent`.animate	Agent: $\texttt{man}_n^1$
begin to shoot	$\texttt{shoot}_v^4$	$\texttt{shoot}_v^4$	Action: $\texttt{shoot}_v^4$
a video	$\texttt{video}_n^2$	`Patient`.inanimate	Patient: $\texttt{video}_n^2$
in	in	PP.in	
the moving bus	$\texttt{bus}_n^1$	NP.`Location`. solid	Location: $\texttt{moving bus}_n^1$

$\texttt{shoot}_v^3$ (sense killing), the role restriction is `Agent.animate V Patient`.**animate** `PP Instrument.solid`. For another sense, $\texttt{shoot}_v^4$ (sense film), the semantic restriction is `Agent.animate V Patient`.**inanimate**. We ensure that the selected WordNet verb sense adheres to both the syntactic frame and the semantic role restriction provided by VerbNet. For example, in Table 4, because $\texttt{video}_n^2$ is a type of inanimate object (inferred through WordNet noun taxonomy), this sense correctly adheres to the VerbNet role restriction. We can now simply apply the VerbNet suggested role `Patient` to $\texttt{video}_n^2$.

Semantic Representation Although VerbNet is helpful as a distant signal to disambiguate and perform semantic role labeling, VerbNet contains over 20 roles and not all of them are general or can be recognized reliably. Therefore, for simplicity, we generalize and group them to get the SUBJECT, VERB, OBJECT, LOCATION roles. For example, the roles `patient`, `recepient`, and, `benefeciary` are generalized to OBJECT. We explore two approaches to obtain the labels based on the output of the semantic parser. First is to use the extracted text chunks directly as labels. Second is to use the corresponding senses as labels (and therefore group multiple text labels). In the following we refer to these as *text-* and *sense-labels*. Thus from each sentence we extract a semantic representation in a form of (SUBJECT, VERB, OBJECT, LOCATION).

4.1.2 SMT

For the sentence generation we build on the two-step translation approach of Rohrbach et al. (2013). As the first step it learns a mapping from the visual input to the semantic representation (SR), modeling pairwise dependencies in a CRF using visual classifiers as unaries. The unaries are trained using an SVM on dense trajectories (Wang and Schmid 2013). In the second step it translates the SR to a sentence using Statistical Machine Translation (SMT) (Koehn et al. 2007). For this the approach uses a concatenated SR as input language, e.g. *cut knife tomato*, and natural sentence as output language, e.g. *The person slices the tomato.* We obtain the SR automatically from the semantic parser, as described above, Sect. 4.1.1. In addition to dense trajectories we use the features described in Sect. 3.1.4.

4.2 Visual Labels + LSTM

Next we present our two-step LSTM-based approach. The first step performs visual recognition using the visual classifiers which we train according to labels' semantics and "visuality". The second step generates textual descriptions using an LSTM network (see Fig. 5b). We explore various design choices for building and training the LSTM.

4.2.1 Robust Visual Classifiers

For training we rely on a parallel corpus of videos and weak sentence annotations. As before (see Sect. 4.1) we parse the sentences to obtain a set of labels (single words or short phrases, e.g. *look up*) to train visual classifiers. However, this time we aim to select the most *visual* labels which can be robustly recognized. In order to do that we take three steps.

Avoiding Parser Failure Not all sentences can be parsed successfully, as e.g. some sentences are incomplete or grammatically incorrect. To avoid loosing the potential labels in these sentences, we match our set of initial labels to the sentences which the parser failed to process. Specifically, we do a simple word matching, i.e. if the label is found in the sentence, we consider this sentence as a positive for the label.

Semantic Groups Our labels correspond to different semantic groups. In this work we consider three most important groups: verbs, objects and places. We propose to treat each label group independently. First, we rely on a different representation for each semantic group, which is targeted to the specific group. Namely we use the activity recognition features Improved Dense Trajectories (DT) for verbs, LSDA scores for objects and PLACES-CNN scores for places. Second, we train one-vs-all SVM classifiers for each group separately. The intuition behind this is to avoid "wrong negatives" (e.g. using *object* "bed" as negative for *place* "bedroom").

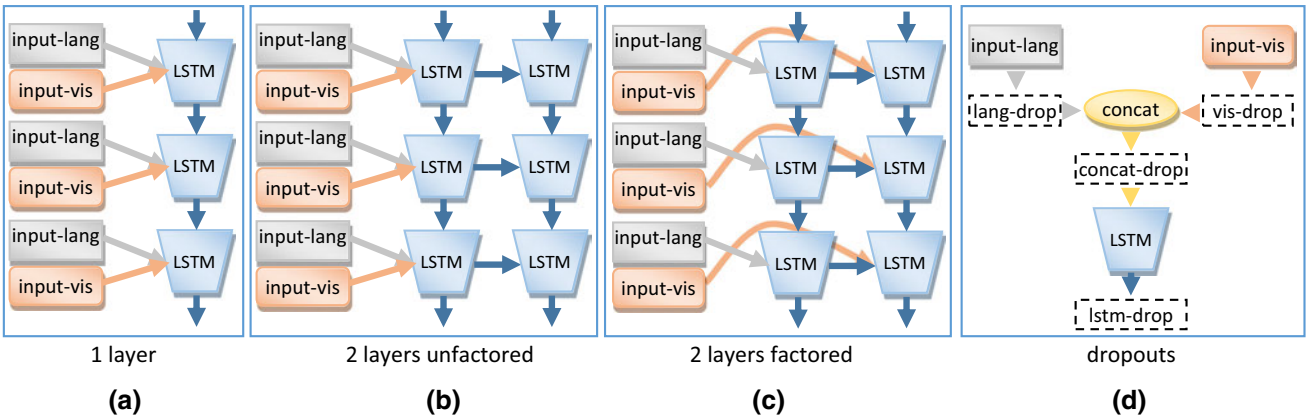


Fig. 6 **a–c** LSTM architectures, **d** variants of placing the dropout layer

Visual Labels Now, how do we select *visual* labels for our semantic groups? In order to find the verbs among the labels we rely on our semantic parser (Sect. 4.1.1). Next, we look up the list of "places" used in Zhou et al. (2014) and search for corresponding words among our labels. We look up the object classes used in Hoffman et al. (2014) and search for these "objects", as well as their base forms (e.g. "domestic cat" and "cat"). We discard all the labels that do not belong to any of our three groups of interest as we assume that they are likely not visual and thus are difficult to recognize. Finally, we discard labels which the classifiers could not learn reliably, as these are likely noisy or not visual. For this we require the classifiers to have certain minimum area under the ROC-curve (Receiver Operating Characteristic). We estimate a threshold for the ROC values on a validation set. We empirically evaluate this as well as all other design choices of our approach in Sect. 5.4.2.

4.2.2 LSTM for Sentence Generation

We rely on the basic LSTM architecture proposed in Donahue et al. (2015) for video description. At each time step an LSTM generates a word and receives the visual classifiers (*input-vis*) as well as as the previous generated word (*input-lang*) as input (see Fig. 6a). We encode each word with a one-hot-vector according to its index in a dictionary and project it in a lower dimensional embedding. The embedding is jointly learned during training of the LSTM. We feed in the classifier scores as input to the LSTM which is equivalent to the best variant proposed in Donahue et al. (2015). We analyze the following aspects for this architecture:

Layer Structure We compare a 1-layer architecture with a 2-layer architecture. In the 2-layer architecture, the output of the first layer is used as input for the second layer (Fig. 6b) and was used by Donahue et al. (2015) for video description. Additionally we also compare to a 2-layer factored architecture of Donahue et al. (2015), where the first layer only gets the language as input and the second layer gets the output of the first as well as the visual input.

Dropout Placement To learn a more robust network which is less likely to overfit we rely on a dropout (Hinton et al. 2012), i.e. a ratio r of randomly selected units is set to 0 during training (while all others are multiplied with $1/r$). We explore different ways to place dropout in the network, i.e. either for language input (*lang-drop*) or visual (*vis-drop*) input only, for both inputs (*concat-drop*) or for the LSTM output (*lstm-drop*), see Fig. 6d.

5 Evaluation on MPII-MD and M-VAD

In this section we evaluate and provide more insights about our movie description datasets MPII-MD and M-VAD. We compare ADs to movie scripts (Sect. 5.1), present a short evaluation of our semantic parser (Sect. 5.2), present the automatic and human evaluation metrics for description (Sect. 5.3) and then benchmark the approaches to video description introduced in Sect. 4 as well as other related work. We conclude this section with an analysis of the different approaches (Sect. 5.5).

In Sect. 6 we will extend this discussion to the results of the Large Scale Movie Description Challenge.

5.1 Comparison of AD Versus Script Data

We compare the AD and script data using 11 movies from the MPII-MD dataset where both are available (see Sect. 3.1.2). For these movies we select the overlapping time intervals with an intersection over union overlap of at least 75%, which results in 279 sentence pairs, we remove 2 pairs which have idendical sentences. We ask humans via Amazon Mechanical Turk (AMT) to compare the sentences with respect to their correctness and relevance to the video, using both video inter-

Table 5 Human evaluation of movie scripts and ADs: which sentence is more correct/relevant with respect to the video (forced choice); majority vote of 5 judges in %. In brackets: at least 4 out of 5 judges agree; see also Sect. 5.1

	Correctness	Relevance
Movie scripts	33.9 (11.2)	33.4 (16.8)
ADs	66.1 (35.7)	66.6 (44.9)

vals as a reference (one at a time). Each task was completed by 5 different human subjects, covering 2770 tasks done in total. Table 5 presents the results of this evaluation. AD is ranked as more correct and relevant in about 2/3 of the cases (i.e. there is margin of about 33%). Looking at the more strict evaluation where at least 4 out of 5 judges agree (in brackets in Table 5) there is still a significant margin of 24.5% between ADs and movie scripts for Correctness, and 28.1% for Relevance. One can assume that in the cases of lower agreement the descriptions are probably of similar quality. This evaluation supports our intuition that scrips contain mistakes and irrelevant content even after being cleaned up and manually aligned.

5.2 Semantic Parser Evaluation

We empirically evaluate the various components of the semantic parsing pipeline, namely, clause splitting (Clause), POS tagging and chunking (NLP), semantic role labeling (Roles), and, word sense disambiguation (WSD). We randomly sample 101 sentences from the MPII-MD dataset over which we perform semantic parsing and log the outputs at various stages of the pipeline (similar to Table 4). We let three human judges evaluate the results for every token in the clause (similar to evaluating every row in Table 4) with a correct/ incorrect label. From this data, we consider the majority vote for every token in the sentence (i.e. at least 2 out of 3 judges must agree). For a given clause, we assign a score of 1 to a component if the component made no mistake for the entire clause. For example, "Roles" gets a score of 1 if, according to majority vote from the judges, we correctly estimate all semantic roles in the clause. Table 6 reports the average accuracy of the components over 130 clauses (generated from 101 sentences).

It is evident that the poorest performing parts are the NLP and the WSD components. Some of the NLP mistakes arise due to incorrect POS tagging. WSD is considered a hard problem and when the dataset contains rare words, the performance is severely affected.

5.3 Evaluation Metrics for Description

In this section we describe how we evaluate the generated descriptions using automatic and human evaluation.

Table 6 Semantic parser accuracy on MPII-MD; discussion in Sect. 5.2

Corpus	Clause	NLP	Roles	WSD
MPII-MD	0.89	0.62	0.86	0.7

5.3.1 Automatic Metrics

For automatic evaluation we rely on the MS COCO Caption Evaluation API.[19] The automatic evaluation measures include BLEU-1,-2,-3,-4 (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin 2004), and CIDEr (Vedantam et al. 2015). We also use the recently proposed evaluation measure SPICE (Anderson et al. 2016), which aims to compare the semantic content of two descriptions, by matching the information contained in dependency parse trees for both descriptions. While we report all measures for the final evaluation in the LSMDC (Sect. 6), we focus our discussion on METEOR and CIDEr scores in the preliminary evaluations in this section. According to Elliott and Keller (2013) and Vedantam et al. (2015), METEOR/CIDEr supersede previously used measures in terms of agreement with human judgments.

5.3.2 Human Evaluation

For the human evaluation we rely on a ranking approach, i.e. human judges are given multiple descriptions from different systems, and are asked to rank them with respect to the following criteria: correctness, relevance, and grammar, motivated by prior work Rohrbach et al. (2013) and on the other hand we asked human judges to rank sentences for "how helpful they would be for a blind person to understand what is happening in the movie". The AMT workers are given randomized sentences, and, in addition to some general instruction, the following definitions:

Grammar "Rank grammatical correctness of sentences: Judge the fluency and readability of the sentence (independently of the correctness with respect to the video)."

Correctness "Rank correctness of sentences: For which sentence is the content more correct with respect to the video (independent if it is complete, i.e. describes everything), independent of the grammatical correctness."

Relevance "Rank relevance of sentences: Which sentence contains the more salient (i.e. relevant, important) events/ objects of the video?"

[19] https://github.com/tylin/coco-caption.

Table 7 Video description performance of different SMT versions on MPII-MD; discussion in Sect. 5.4.1

METEOR	
SMT with our sense-labels	
IDT 30	4.93
IDT 100	5.12
Combi 100	5.19
SMT with our text-labels	
IDT 30	5.59
IDT 100	5.51
Combi 100	5.42

Helpful for the Blind In the LSMDC evaluation we introduce a new measure, which should capture how useful a description would be for blind people: "Rank the sentences according to how useful they would be for a blind person which would like to understand/follow the movie without seeing it."

5.4 Movie Description Evaluation

As the collected text data comes from the movie context, it contains a lot of information specific to the plot, such as names of the characters. We pre-process each sentence in the corpus, transforming the names to "Someone" or "people" (in case of plural).

We first analyze the performance of the proposed approaches on the MPII-MD dataset, and then evaluate the best version on the M-VAD dataset. For MPII-MD we split the 11 movies with associated scripts and ADs (in total 22 alignments, see Sect. 3.1.2) into validation set (8) and test set (14). The other 83 movies are used for training. On M-VAD we use 10 movies for testing, 10 for validation and 72 for training.

5.4.1 Semantic Parsing + SMT

Table 7 summarizes results of multiple variants of the SMT approach when using the SR from our semantic parser. "Combi" refers to combining IDT, HYBRID, and PLACES as unaries in the CRF. We did not add LSDA as we found that it reduces the performance of the CRF. After extracting the labels we select the ones which appear at least 30 or 100 times as our visual attributes. Overall, we observe similar performance in all cases, with slightly better results for text-labels than sense-labels. This can be attributed to sense disambiguation errors of the semantic parser. In the following we use the "IDT 30" model, which achieves the highest score of 5.59, and denote it as "SMT-Best".[20]

5.4.2 Visual Labels + LSTM

We start with exploring different design choices of our approach. We build on the labels discovered by the semantic parser. To learn classifiers we select the labels that appear at least 30 times, resulting in 1263 labels. The parser additionally tells us whether the label is a verb. The LSTM output/hidden unit as well as memory cell have each 500 dimensions.

Robust Visual Classifiers We first analyze our proposal to consider groups of labels to learn different classifiers and also to use different visual representations for these groups (see Sect. 4.2). In Table 8 we evaluate our generated sentences using different input features to the LSTM on the validation set of MPII-MD. In our baseline, in the top part of Table 8, we use the same visual descriptors for all labels. The PLACES feature is best with 7.10 METEOR. Combination by stacking all features (IDT + LSDA + PLACES) improves further to 7.24 METEOR. The second part of the table demonstrates the effect of introducing different semantic label groups. We first split the labels into "Verbs" and all others. Given that some labels appear in both roles, the total number of labels increases to 1328 (line 5). We compare two settings of training the classifiers: "Retrieved" (we retrieve the classifier scores from the classifiers trained in the previous step), "Trained" (we train the SVMs specifically for each label type, e.g. "Verbs"). Next, we further divide the non-"Verb" labels into "Places" and "Others"(line 6), and finally into "Places" and "Objects"(line 7). We discard the unused labels and end up with 913 labels. Out of these labels, we select the labels where the classifier obtains a ROC higher or equal to 0.7 (threshold selected experimentally). After this we obtain 263 labels and the best performance in the "Trained" setting (line 8). To support our intuition about the importance of the label discrimination (i.e. using different features for different semantic groups of labels), we propose another baseline (line 9). Here we use the same set of 263 labels but provide the same feature for all of them, namely the best performing combination IDT + LSDA + PLACES. As we see, this results in an inferior performance.

We make several observations from Table 8 which lead to robust visual classifiers from the weak sentence annotations. (a) It is beneficial to select features based on the label semantics. (b) Training one-vs-all SVMs for specific label groups consistently improves the performance as it avoids "wrong" negatives. (c) Focusing on more "visual" labels helps: we reduce the LSTM input dimensionality to 263 while improving the performance.

[20] We also evaluated the "Semantic parsing+SMT" approach on a corpus where annotated SRs are available, namely TACoS Multi-Level (Rohrbach et al. 2014), and showed the comparable performance to manually annotated SRs, see Rohrbach et al. (2015c).

Table 8 Comparison of different choices of labels and visual classifiers; all results reported on the validation set of MPII-MD; for discussion see Sect. 5.4.2

Approach	Labels	Classifiers (METEOR in %)	
		Retrieved	Trained
Baseline: all labels treated the same way			
(1) IDT	1263	–	6.73
(2) LSDA	1263	–	7.07
(3) PLACES	1263	–	7.10
(4) IDT+LSDA+PLACES	1263	–	7.24
Visual labels			
(5) Verbs(IDT), Others(LSDA)	1328	7.08	7.27
(6) Verbs(IDT), Places(PLACES), Others(LSDA)	1328	7.09	7.39
(7) Verbs(IDT), Places(PLACES), Objects(LSDA)	913	7.10	7.48
(8) + restriction to labels with $ROC \geq 0.7$	263	7.41	**7.54**
Baseline: all labels treated the same way, labels from (8)			
(9) IDT+LSDA+PLACES	263	7.16	7.20

Bold value indicates the best performing variant in the table

Table 9 LSTM architectures (fixed parameters: LSTM-drop, dropout 0.5), MPII-MD val set; labels, classifiers as Table 8, line (8); for discussion see Sect. 5.4.2

Architecture	METEOR
1 layer	**7.54**
2 layers unfact.	**7.54**
2 layers fact.	7.41

Bold value indicates the best performing variant in the table

Table 10 Dropout strategies (fixed parameters: 1-layer, dropout 0.5), MPII-MD val set; labels, classifiers as Table 8, line (8); for discussion see Sect. 5.4.2

Dropout	METEOR
No dropout	7.19
Lang-drop	7.13
Vis-drop	7.34
Concat-drop	7.29
LSTM-drop	**7.54**

Bold value indicates the best performing variant in the table

Table 11 Dropout ratios (fixed parameters: 1-layer, LSTM-drop), MPII-MD val set; labels, classifiers as Table 8, line (8); for discussion see Sect. 5.4.2

Dropout ratio	METEOR
r = 0.1	7.22
r = 0.25	7.42
r = 0.5	**7.54**
r = 0.75	7.46

Bold value indicates the best performing variant in the table

LSTM Architectures Now, as described in Sect. 4.2.2, we look at different LSTM architectures and training configurations. In the following we use the best performing "Visual Labels" approach, Table 8, line (8).

We start with examining the architecture, where we explore different configurations of LSTM and dropout layers. Table 9 shows the performance of three different networks: "1 layer", "2 layers unfactored" and "2 layers factored" introduced in Sect. 4.2.2. As we see, the "1 layer" and "2 layers unfactored" perform equally well, while "2 layers factored" is inferior to them. In the following experiments we use the simpler "1 layer" network. We then compare different dropout placements as illustrated in Table 10. We obtain the best result when applying dropout after the LSTM layer ("lstm-drop"), while having no dropout or applying it only to language leads to stronger over-fitting to the visual features. Putting dropout after the LSTM (and prior to a final prediction layer) makes the entire system more robust. As for the best dropout ratio, we find that 0.5 works best with lstm-dropout (Table 11).

In most of the experiments we trained our networks for 25,000 iterations. After looking at the METEOR scores for intermediate iterations we found that at iteration 15,000 we achieve best performance overall. Additionally we train multiple LSTMs with different random orderings of the training data. In our experiments we combine three in an ensemble, averaging the resulting word predictions.

To summarize, the most important aspects that decrease over-fitting and lead to better sentence generation are: (a) a correct learning rate and step size, (b) dropout after the LSTM layer, (c) choosing the training iteration based on METEOR score as opposed to only looking at the LSTM accuracy/loss which can be misleading, and (d) building ensembles of multiple networks with different random initializations.[21]

[21] More details can be found in our corresponding arXiv version (Rohrbach et al. 2015a).

Table 12 Test Set of MPII-MD: Comparison of our proposed methods to baselines and prior work: S2VT (Venugopalan et al. 2015a), Temporal Attention (Yao et al. 2015); human eval ranked 1–3, lower is better; for discussion see Sect. 5.4.3

Approach	METEOR in %	CIDEr in %	Human evaluation: rank		
			Correct.	Grammar	Relev.
NN baselines					
IDT	4.87	2.77	–	–	–
LSDA	4.45	2.84	–	–	–
PLACES	4.28	2.73	–	–	–
HYBRID	4.34	3.29	–	–	–
SMT-Best (ours)	5.59	8.14	2.11	2.39	2.08
S2VT	6.27	9.00	2.02	**1.67**	2.06
Visual-Labels (ours)	**7.03**	**9.98**	**1.87**	1.94	**1.86**
NN METEOR upperbound	19.43	–	–	–	–

Bold values indicate the best performing variant per measure/column

Table 13 Test set of M-VAD: Comparison of our proposed methods to prior work: S2VT (Venugopalan et al. 2015a), Temporal Attention (Yao et al. 2015); human eval ranked 1–3, lower is better; for discussion see Sect. 5.4.3

Approach	METEOR in %	CIDEr in %
Temporal Attention	4.33	5.55
S2VT	5.62	7.22
Visual-Labels (ours)	**6.36**	**7.48**

Bold values indicate the best performing variant per measure/column

5.4.3 Comparison to Related Work

Experimental Setup In this section we perform the evaluation on the test set of the MPII-MD dataset (6578 clips) and M-VAD dataset (4951 clips). We use METEOR and CIDEr for automatic evaluation and we perform a human evaluation on a random subset of 1300 video clips, see Sect. 5.3 for details. For M-VAD experiments we train our method on M-VAD and use the same LSTM architecture and parameters as for MPII-MD, but select the number of iterations on the M-VAD validation set.

Results on MPII-MD Table 12 summarizes the results on the test set of MPII-MD. Here we additionally include the results from a nearest neighbor baseline, i.e. we retrieve the closest sentence from the training corpus using L1-normalized visual features and the intersection distance. Our SMT-Best approach clearly improves over the nearest neighbor baselines. With our Visual-Labels approach we significantly improve the performance, specifically by 1.44 METEOR points and 1.84 CIDEr points. Moreover, we improve over the recent approach of (Venugopalan et al. 2015a), which also uses an LSTM to generate video descriptions. Exploring different strategies to label selection and classifier training, as well as various LSTM configurations allows to obtain better result than prior work on the MPII-MD dataset. Human evaluation mainly agrees with the automatic measure. Visual-Labels outperforms both other methods in terms of Correctness and Relevance, however it loses to S2VT in terms of Grammar. This is due to the fact that S2VT produces overall shorter (7.4 vs. 8.7 words per sentence) and simpler sentences, while our system generates longer sentences and therefore has higher chances to make mistakes. We also propose a retrieval upperbound. For every test sentence we retrieve the closest training sentence according to the METEOR score. The rather low METEOR score of 19.43 reflects the difficulty of the dataset. We show some qualitative results in Fig. 7.

Results on M-VAD Table 13 shows the results on the test set of M-VAD dataset. Our Visual-Labels method outperforms S2VT (Venugopalan et al. 2015a) and Temporal Attention (Yao et al. 2015) in METEOR and CIDEr score. As we see, the results agree with Table 12, but are consistently lower, suggesting that M-VAD is more challenging than MPII-MD. We attribute this to a more precise manual alignment of the MPII-MD dataset.

5.5 Movie Description Analysis

Despite the recent advances in the video description task, the performance on the movie description datasets (MPII-MD and M-VAD) remains rather low. In this section we want to look closer at three methods, SMT-Best, S2VT and Visual-Labels, in order to understand where these methods succeed and where they fail. In the following we evaluate all three methods on the MPII-MD test set.

5.5.1 Difficulty Versus Performance

As the first study we suggest to sort the test reference sentences by difficulty, where difficulty is defined in multiple ways[21].

	Approach	Sentence
	SMT-Best (ours)	Someone is a man, someone is a man.
	S2VT	Someone looks at him, someone turns to someone.
	Visual-Labels (ours)	Someone is standing in the crowd, a little man with a little smile.
	Reference	Someone, back in elf guise, is trying to calm the kids.
	SMT-Best (ours)	The car is a water of the water.
	S2VT	On the door, opens the door opens.
	Visual-Labels (ours)	The fellowship are in the courtyard.
	Reference	They cross the quadrangle below and run along the cloister.
	SMT-Best (ours)	Someone is down the door, someone is a back of the door, and someone is a door.
	S2VT	Someone shakes his head and looks at someone.
	Visual-Labels (ours)	Someone takes a drink and pours it into the water.
	Reference	Someone grabs a vodka bottle standing open on the counter and liberally pours some on the hand.

Fig. 7 Qualitative comparison of our proposed methods to prior work: S2VT (Venugopalan et al. 2015a). Examples from the test set of MPII-MD. Visual-Labels identifies activities, objects, and places better than the other two methods. See Sect. 5.4.3

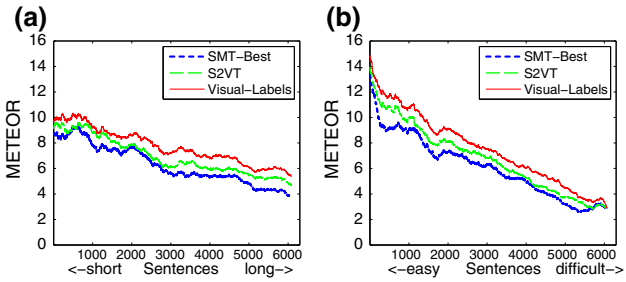


Fig. 8 *Y-axis* METEOR score per sentence. *X-axis* MPII-MD test sentences 1–6578 sorted by **a** length (increasing); **b** word frequency (decreasing). Shown values are smoothed with a mean filter of size 500. For discussion see Sect. 5.5.1

Sentence Length and Word Frequency Some of the intuitive sentence difficulty measures are its length and average frequency of its words. When sorting the data by difficulty (increasing sentence length or decreasing average word frequency), we find that all three methods have the same tendency to obtain lower METEOR score as the difficulty increases. Fig. 8a) shows the performance of compared methods w.r.t. the sentence length. For the word frequency the correlation is even stronger, see Fig. 8b. Visual-Labels consistently outperforms the other two methods, most notable as the difficulty increases.

5.5.2 Semantic Analysis

WordNet Verb Topics Next we analyze the test reference sentences w.r.t. verb semantics. We rely on WordNet Topics (high level entries in the WordNet ontology), e.g. "motion", "perception", defined for most synsets in WordNet (Fellbaum 1998). Sense information comes from our automatic semantic parser, thus it might be noisy. We showcase the 3 most frequent verbs for each Topic in Table 14. We select sentences with a single verb, group them according to the verb Topic and compute an average METEOR score for each Topic, see Fig. 9. We find that Visual-Labels is best for all Topics except "communication", where SMT-Best wins. The most frequent verbs there are "look up" and "nod", which are also frequent in the dataset and in the sentences produced by SMT-Best. The best performing Topic, "cognition", is highly biased to "look at" verb. The most frequent Topics,

Table 14 Entropy and top 3 frequent verbs of each WordNet topic; for discussion see Sect. 5.5.2

Topic	Entropy	Top-1	Top-2	Top-3
Motion	7.05	Turn	Walk	Shake
Contact	7.10	Open	Sit	Stand
Perception	4.83	Look	Stare	See
Stative	4.84	Be	Follow	Stop
Change	6.92	Reveal	Start	Emerge
Communication	6.73	Look up	Nod	Face
Body	5.04	Smile	Wear	Dress
Social	6.11	Watch	Join	Do
Cognition	5.21	Look at	See	Read
Possession	5.29	Give	Take	Have
None	5.04	Throw	Hold	Fly
Creation	5.69	Hit	Make	Do
Competition	5.19	Drive	Walk over	Point
Consumption	4.52	Use	Drink	Eat
Emotion	6.19	Draw	Startle	Feel
Weather	3.93	Shine	Blaze	Light up

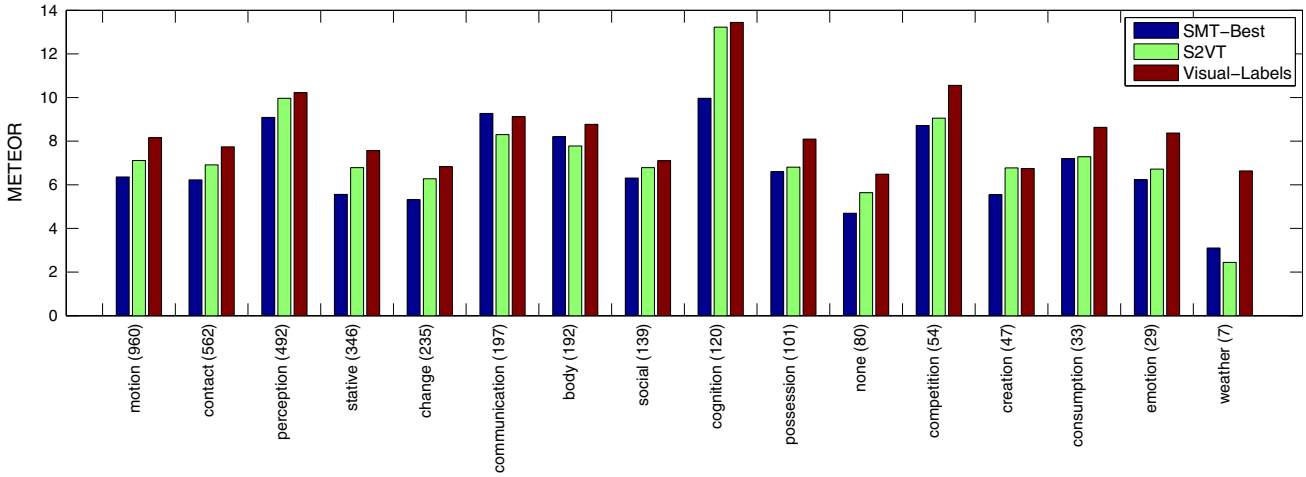


Fig. 9 Average METEOR score for WordNet verb Topics. Selected sentences with single verb, number of sentences in *brackets*. For discussion see Sect. 5.5.2

"motion" and "contact", which are also visual (e.g. "turn", "walk", "sit"), are nevertheless quite challenging, which we attribute to their high diversity (see their entropy w.r.t. different verbs and their frequencies in Table 14). Topics with more abstract verbs (e.g. "be", "have", "start") get lower scores.

Top 100 Best and Worst Sentences We look at 100 test reference sentences, where Visual-Labels obtains highest and lowest METEOR scores. Out of 100 best sentences 44 contain the verb "look" (including phrases such as "look at"). The other frequent verbs are "walk", "turn", "smile", "nod", "shake", i.e. mainly visual verbs. Overall the sentences are simple. Among the worst 100 sentences we observe more diversity: 12 contain no verb, 10 mention unusual words (specific to the movie), 24 have no subject, 29 have a non-human subject. This leads to a lower performance, in particular, as most training sentences contain "Someone" as subject and generated sentences are biased towards it.

Summary (a) The test reference sentences that mention verbs like "look" get higher scores due to their high frequency in the dataset. (b) The sentences with more "visual" verbs tend to get higher scores. (c) The sentences without verbs (e.g. describing a scene), without subjects or with non-human subjects get lower scores, which can be explained by dataset biases.

6 The Large Scale Movie Description Challenge

The Large Scale Movie Description Challenge (LSMDC) was held twice, first in conjunction with ICCV 2015 (LSMDC 15) and then at ECCV 2016 (LSMDC 16). For the automatic evaluation we set up an evaluation server[3]. During the first phase of the challenge the participants could evaluate the outputs of their system on the public test set. In the second phase of the challenge the participants were provided with the videos from the blind test set (without textual descriptions). These were used for the final evaluation. To measure performance of the competing approaches we performed both automatic and human evaluation. The submission format was similar to the MS COCO Challenge (Chen et al. 2015) and we also used the identical automatic evaluation protocol. The challenge winner was determined based on the human evaluation. In the following we review the participants and their results for both LSMDC 15 and LSMDC 16. As they share the same public and blind test sets, as described in Sect. 3.3, we can also compare the submissions to both challenges with each other.

6.1 LSMDC Participants

We received 4 submissions to LSMDC 15, including our Visual-Labels approach. The other submissions are S2VT (Venugopalan et al. 2015b), Temporal Attention (Yao et al. 2015) and Frame-Video-Concept Fusion (Shetty and Laaksonen 2015). For LSMDC 16 we received 6 new submissions. As the blind test set is not changed between LSMDC 2015 to LSMDC 2016, we look at all the submitted results jointly. In the following we summarize the submissions based on the (sometimes very limited) information provided by the authors.

6.1.1 LSMDC 15 Submissions

S2VT (Venugopalan et al. 2015b) Venugopalan et al. (2015b) propose S2VT, an encoder–decoder framework, where a single LSTM encodes the input video, frame by frame, and decodes it into a sentence. We note that the results to LSMDC

were obtained with a different set of hyper-parameters then the results discussed in the previous section. Specifically, S2VT was optimized w.r.t. METEOR on the validation set, which resulted in significantly longer but also nosier sentences.

Frame-Video-Concept Fusion (Shetty and Laaksonen 2015) Shetty and Laaksonen (2015) evaluate diverse visual features as input for an LSTM generation frame-work. Specifically they use dense trajectory features (Wang et al. 2013) extracted for the entire clip and VGG (Simonyan and Zisserman 2015) and GoogleNet (Szegedy et al. 2015) CNN features extracted at the center frame of each clip. They find that training 80 concept classifiers on MS COCO with the CNN features, combined with dense trajectories provides the best input for the LSTM.

Temporal Attention (Yao et al. 2015) Yao et al. (2015) propose a soft-attention model based on Xu et al. (2015a) which selects the most relevant temporal segments in a video, incorporates 3-D CNN and generates a sentence using an LSTM.

6.1.2 LSMDC 16 Submissions

Tel Aviv University This submission retrieves a nearest neighbor from the training set, learning a unified space using Canonical Correlation Analysis (CCA) over textual and visual features. For the textual representation it relies on the Word2Vec representation using a Fisher Vector encoding with a Hybrid Gaussian-Laplacian Mixture Model (Klein et al. 2015) and for the visual representation it uses RNN Fisher Vector (Lev et al. 2015), encoding video frames with the 19-layer VGG.

Aalto University (Shetty and Laaksonen 2016) Shetty and Laaksonen (2016) rely on an ensemble of four models which were trained on the MSR-VTT dataset (Xu et al. 2016) without additional training on the LSMDC dataset. The four models were trained with different combinations of key-frame based GoogleLeNet features and segment based dense trajectory and C3D features. A separately trained evaluator network was used to predict the result of the ensemble.

Seoul NU This work relies on temporal and attribute attention.

SNUVL (Yu et al. 2016b) Yu et al. (2016b) first learn a set of semantic attribute classifiers. To generate a description for a video clip, they rely on Temporal Attention and attention over semantic attributes.

IIT Kanpur This submission uses an encoder–decoder framework with 2 LSTMs, one LSTM used to encode the frame sequence of the video and another to decode it into a sentence.

VD-ivt (BUPT CIST AI lab) According to the authors, their VD-ivt model consists of three parallel channels: a basic video description channel, a sentence to sentence channel for language learning, and a channel to fuse visual and textual information.

6.2 LSMDC Quantitative Results

We first discuss the submissions w.r.t. to automatic measures and then discuss the human evaluations, which determined the winner for the challenges.

6.2.1 Automatic Evaluation

We first look at the results of the automatic evaluation on the blind test set of LSMDC in Table 15. In the first edition of the challenge, LSMDC 15, our Visual-Labels approach obtains highest scores in all evaluation measures except BLEU-1,-2, where S2VT wins. One reason for lower scores for Frame-Video-Concept Fusion and Temporal Attention appears to be the generated sentence length, which is much smaller compared to the reference sentences, as we discuss below (see also Table 16). When extended to LSMDC 16 submissions, we observe that most approaches perform below S2VT/Visual-Labels, except for VD-ivt, which achieves METEOR 8.0. Surprisingly, but confirmed with the authors, VD-ivt predicts only a single sentence "*Someone is in the front of the room.*", which seems to be optimized w.r.t. the METEOR score, while e.g. CIDEr score shows that this sentence is not good for most video clips. While most approaches are generating novel descriptions, Tel Aviv University is the only retrieval-based approached among the submissions. It takes a second place w.r.t. the CIDEr score, while not achieving particularly high scores in other measures.

We closer analyze the outputs of the compared approaches in Table 16, providing detailed statistics over the generated descriptions. Among the LSMDC 15 submissions, with respect to the sentence length, Visual-Labels and S2VT demonstrate similar properties to the reference descriptions, while the approaches Frame-Video-Concept Fusion and Temporal Attention generate much shorter sentences (5.16 and 3.63 words on average vs. 8.74 of the references). In terms of vocabulary size all approaches fall far below the reference descriptions. This large gap indicates a problem in that all the compared approaches focus on a rather small set of visual and language concepts, ignoring a long tail in the distribution. The number of unique sentences confirms the previous finding, showing slightly higher numbers for Visual-

Table 15 Automatic evaluation on the blind test set of the LSMDC, in %; for discussion see Sect. 6.2

Approach	BLEU				METEOR	ROUGE	CIDEr	SPICE
	1	2	3	4				
Submissions to LSMDC 15								
Visual-Labels (ours)	16.1	5.2	**2.1**	**0.9**	**7.1**	**16.4**	**11.2**	13.2
S2VT (Venugopalan et al. 2015b)	**17.4**	**5.3**	1.8	0.7	7.0	16.1	9.1	11.4
Frame-Video-Concept Fusion (Shetty and Laaksonen 2015)	11.0	3.4	1.3	0.6	6.1	15.6	9.0	13.4
Temporal Attention (Yao et al. 2015)	5.6	1.5	0.6	0.3	5.2	13.4	6.2	**14.3**
Submissions to LSMDC 16								
Tel Aviv University	14.5	4.1	1.4	0.6	5.8	13.4	10.1	7.7
Aalto University (Shetty and Laaksonen 2016)	6.9	1.6	0.5	0.2	3.4	7.0	3.5	2.6
Seoul NU	9.2	2.9	1.0	0.4	4.0	9.6	7.6	4.8
SNUVL (Yu et al. 2016b)	15.6	4.4	1.4	0.4	7.1	14.7	7.0	11.5
IIT Kanpur	11.8	3.6	1.3	0.5	7.4	14.2	4.7	7.2
VD-ivt (BUPT CIST AI lab)	15.9	4.3	1.0	0.3	**8.0**	15.0	4.8	10.6

Bold values indicate the best performing approach per measure/column for LSMDC 2015, and LSMDC 2016, if it improved over LSMDC 2015

Table 16 Description statistics for different methods and reference sentences on the blind test set of the LSMDC; for discussion see Sect. 6.2

Approach	Avg. sentence length	Vocabulary size	% Unique sentences	%Novel sentences
Submissions to LSMDC 15				
Visual-Labels (ours)	7.47	525	45.11	66.76
S2VT (Venugopalan et al. 2015b)	8.77	663	30.17	72.10
Frame-Video-Concept Fusion (Shetty and Laaksonen 2015)	5.16	401	9.09	30.81
Temporal Attention (Yao et al. 2015)	3.63	117	1.39	6.48
Submissions to LSMDC 16				
Tel Aviv University	9.34	5530	58.35	0.00
Aalto University (Shetty and Laaksonen 2016)	6.83	651	24.39	94.09
Seoul NU	6.16	459	24.26	52.78
SNUVL (Yu et al. 2016b)	8.53	756	41.54	76.03
IIT Kanpur	16.2	1172	39.37	100.00
VD-ivt (BUPT CIST AI lab)	8.00	7	0.01	100.00
Reference	8.75	6820	97.19	92.63

Labels and S2VT, while the other two tend to frequently generate the same description for different clips. Finally, the percentage of novel sentences (not present among the training descriptions) highlights another aspect, namely the amount of novel vs. retrieved descriptions. As we see, all the methods "retrieve" some amount of descriptions from training data, while the approach Temporal Attention produces only 7.36% novel sentences. Looking at the LSMDC 16 submissions, we, not surprisingly, see that Tel Aviv University retrieval approach achieves highest diversity among all approaches. Most other submissions have similar statistics to LSMDC 15 submissions. Interestingly, Shetty and Laaksonen (2016) generate many novel sentences, as they are not trained on LSMDC, but on the MSR-VTT dataset. Two outliers are IIT Kanpur, which generates very long and noisy descriptions, and VD-ivt, which, as mentioned above, generates the same sentence for all video clips.

6.2.2 Human Evaluation

We performed separate human evaluations for LSMDC 15 and LSMDC 16.

LSMDC 15 The results of the human evaluation are shown in Table 17. The human evaluation was performed over 1,200 randomly selected clips from the blind test set of LSMDC. We follow the evaluation protocol defined in Sect. 5.3.2. As known from literature (Chen et al. 2015; Elliott and Keller 2013; Vedantam et al. 2015), automatic evaluation measures do not always agree with the human evaluation.

Table 17 Human evaluation on the blind test set of the LSMDC; human eval ranked 1–5, lower is better; for discussion see Sect. 6.2

Approach	Correctness	Grammar	Relevance	Helpful for blind
Visual-Labels (ours)	3.32	3.37	3.32	**3.26**
S2VT (Venugopalan et al. 2015a)	3.55	3.09	3.53	3.42
Frame-Video-Concept Fusion (Shetty and Laaksonen 2015)	**3.10**	**2.70**	**3.29**	3.29
Temporal Attention (Yao et al. 2015)	3.14	2.71	3.31	3.36
Reference	1.88	3.13	1.56	1.57

Bold values indicate the best performing approach per measure/column

Table 18 LSMDC 16; human evaluation; ratio of sentences which are judged better or equal compared to the reference description, with at least two out of three judges agreeing (in %); for discussion see Sect. 6.2

Approach	Better or equal than reference
Submissions to LSMDC 15	
Visual-Labels (ours)	18.8
S2VT (Venugopalan et al. 2015b)	15.6
Frame-Video-Concept Fusion (Shetty and Laaksonen 2015)	15.2
Temporal Attention (Yao et al. 2015)	16.8
Submissions to LSMDC 16	
Tel Aviv University	**22.4**
Aalto University (Shetty and Laaksonen 2016)	16.4
Seoul NU	14.4
SNUVL (Yu et al. 2016b)	8.8
IIT Kanpur	7.2
VD-ivt (BUPT CIST AI lab)	1.6

Bold value indicates the best performing approach in the table

Here we see that human judges prefer the descriptions from Frame-Video-Concept Fusion approach in terms of correctness, grammar and relevance. In our alternative evaluation, in terms of being helpful for the blind, Visual-Labels wins. Possible explanation for it is that in this evaluation criteria human judges penalized less the errors in the descriptions but rather looked at their overall informativeness. In general, the gap between different approaches is not large. Based on the human evaluation the winner of the LSMDC 15 challenge is Frame-Video-Concept Fusion approach of Shetty and Laaksonen (2015).

LSMDC 16 For the LSMDC 16 the evaluation protocol is different from the one above. As we have to compare more approaches the ranking becomes unfeasible. Additionally we would like to capture the human agreement in this evaluation. This leads us to the following evaluation protocol which is inspired by the human evaluation metric "M1" in the MS COCO Challenge (Chen et al. 2015). The humans are provided with randomized pairs (reference, generated sentence) from each system and asked to decide in terms of being helpful for the blind person (a) if sentence 1 is better (b) both are similar (c) sentence 2 is better. Each pair is judged by 3 humans. For an approach to get a point at least 2 out of 3 humans should agree that a generated sentence is better or equal to a reference. The results of the human evaluation on 250 randomly selected sentence pairs are presented in Table 18. Tel Aviv University is ranked best by the human judges and thus it wins the LSMDC 16 challenge. Visual-Labels gets the second place, next are Temporal Attention and Aalto University. The VD-ivt submission with identical descriptions is ranked worst. Additionally we measure the correlation between the automatic and human evaluation in Fig. 10. We compare BLEU@4, METEOR, CIDEr and SPICE and find that CIDEr score provides the highest and reasonable (0.61) correlation with human judgments. SPICE shows no correlation, METEOR demonstrates negative correlation. We attribute this to the fact that the approaches generate very different types of descriptions (long/short, simple/retrieved from the training data, etc.) as discussed above and that we only have a single reference to compute these metrics. While we believe that these metrics can still provide reasonable scores for similar models, comparing very diverse methods and results, requires human evaluation. However, also for human evaluation, further studies are needed in the future, to determine what are the best evaluation protocols.

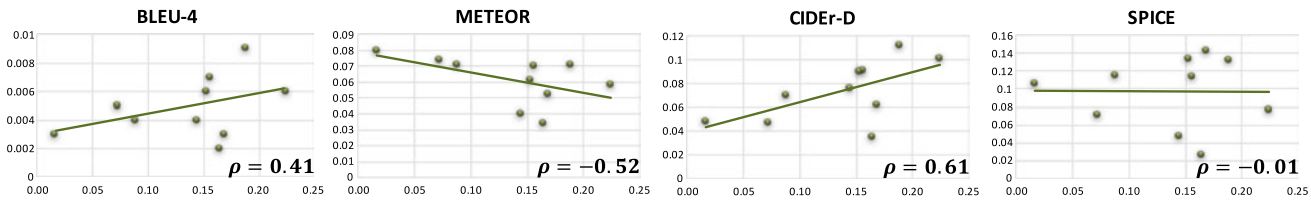


Fig. 10 LSDMC 16: We plot the correlation between human evaluation score (*x axis*) and 4 automatic measures (*y axis*)

	Approach	Sentence
	Visual-Labels (ours)	Someone lies on the bed.
	S2VT	Someone lies asleep on his bed.
	Frame-Video-Concept Fusion	Someone lies on the bed.
	Temporal Attention	Someone lies in bed.
	Reference	Someone lies on her side facing her new friend.
	Visual-Labels (ours)	Someone sits down.
	S2VT	Someone sits on the couch and looks at the tv.
	Frame-Video-Concept Fusion	Someone sits at the table.
	Temporal Attention	Someone looks at someone.
	Reference	Someone takes a seat and someone moves to the stove.
	Visual-Labels (ours)	Someone walks to the front of the house.
	S2VT	Someone looks at the house.
	Frame-Video-Concept Fusion	Someone walks up to the house.
	Temporal Attention	Someone looks at someone.
	Reference	Someone sets down his young daughter then moves to a small wooden table.
	Visual-Labels (ours)	Someone turns to someone.
	S2VT	Someone looks at someone.
	Frame-Video-Concept Fusion	Someone turns to someone.
	Temporal Attention	Someone stands alone.
	Reference	Someone dashes for the staircase.
	Visual-Labels (ours)	Someone takes a deep breath and takes a deep breath.
	S2VT	Someone looks at someone and looks at him.
	Frame-Video-Concept Fusion	Someone looks up at the ceiling.
	Temporal Attention	Someone stares at someone.
	Reference	Someone digs out her phone again, eyes the display, and answers the call.

Fig. 11 Qualitative comparison of our approach Visual-Labels, S2VT (Venugopalan et al. 2015b), Frame-Video-Concept Fusion (Shetty and Laaksonen 2015) and Temporal Attention (Yao et al. 2015) on the blind test set of the LSMDC. Discussion see Sect. 6.3

6.3 LSMDC Qualitative Results

Figure 11 shows qualitative results from the competing approaches submitted to LSMDC 15. The first two examples are success cases, where most of the approaches are able to describe the video correctly. The third example is an interesting case where visually relevant descriptions, provided by most approaches, do not match the reference description, which focuses on an action happening in the background of the scene ("Someone sets down his young daughter then moves to a small wooden table."). The last two rows contain partial and complete failures. In one all approaches fail to recognize the person running away, only capturing the "turning" action which indeed happened before running. In the other one, all approaches fail to recognize that the woman interacts with the small object (phone).

Figure 12 compares all LSMDC 15 approaches with the LSMDC 16 winner, Tel Aviv University, on a sequence of 5 consecutive clips. We can make the following observations from these examples. Although, Tel Aviv University is a retrieval-based approach, it does very well in many cases, providing an added benefit of fluent and grammatically correct descriptions. One side-effect of retrieval is that when it fails, it produces a completely irrelevant description, e.g. the second example. Tel Aviv University and Visual-Labels are able to capture important details, such as sipping a drink, which the other methods fail to recognize. Descriptions generated by Visual-Labels and S2VT tend to be longer and noisier than the ones by Frame-Video-Concept Fusion and Temporal Attention, while Temporal Attention tends to produce generally applicable sentences, e.g. "Someone looks at someone".

	Approach	Sentence
	Visual-Labels (ours)	Someone takes a seat on the table and takes a seat on his desk.
	S2VT	Someone looks at someone and smiles.
	Frame-Video-Concept Fusion	Someone looks at someone.
	Temporal Attention	Someone gets up.
	Tel Aviv University	Farther along, the mustached stranger sits on a bench.
	Reference	Later, someone sits with someone and someone.
	Visual-Labels (ours)	Someone gets out of the car and walks off.
	S2VT	Someone walks up to the front of the house.
	Frame-Video-Concept Fusion	Someone walks up to the front door.
	Temporal Attention	Someone gets out of the car.
	Tel Aviv University	He sees a seated man on the TV gesturing.
	Reference	Now someone steps out of the carriage with his new employers.
	Visual-Labels (ours)	Someone walks up to the street, and someone is walking to the other side of.
	S2VT	Someone walks over to the table and looks at the other side of the house.
	Frame-Video-Concept Fusion	Someone walks away.
	Temporal Attention	Someone gets out of the car.
	Tel Aviv University	Later smiling, the two walk hand in hand down a busy sidewalk noticing every hat-wearing man they pass.
	Reference	The trio starts across a bustling courtyard.
	Visual-Labels (ours)	Someone sips his drink.
	S2VT	Someone sits at the table and looks at someone.
	Frame-Video-Concept Fusion	Someone sits up.
	Temporal Attention	Someone looks at someone.
	Tel Aviv University	Someone sits at a table sipping a drink.
	Reference	As the men drink red wine, someone and someone watch someone take a sip.
	Visual-Labels (ours)	Someone takes a bite.
	S2VT	Someone sits at the table.
	Frame-Video-Concept Fusion	Someone looks at someone.
	Temporal Attention	Someone looks at someone.
	Tel Aviv University	Later at the dinner table.
	Reference	Someone tops off someone's glass.

Fig. 12 Qualitative comparison of our approach Visual-Labels, S2VT (Venugopalan et al. 2015b), Frame-Video-Concept Fusion (Shetty and Laaksonen 2015), Temporal Attention (Yao et al. 2015), and Tel Aviv University on 5 consecutive clips from the blind test set of the LSMDC. Discussion see Sect. 6.3

7 Conclusion

In this work we present the Large Scale Movie Description Challenge (LSMDC), a novel dataset of movies with aligned descriptions sourced from movie scripts and ADs (audio descriptions for the blind, also referred to as DVS). Altogether the dataset is based on 200 movies and has 128,118 sentences with aligned clips. We compare AD with previously used script data and find that AD tends to be more correct and relevant to the movie than script sentences.

Our approach, *Visual-Labels*, to automatic movie description trains visual classifiers and uses their scores as input to an LSTM. To handle the weak sentence annotations we rely on three ingredients. (1) We distinguish three semantic groups of labels (verbs, objects, and places). (2) We train them separately, removing the noisy negatives. (3) We select only the most reliable classifiers. For sentence generation we show the benefits of exploring different LSTM architectures and learning configurations.

To evaluate different approaches for movie description, we organized a challenge at ICCV 2015 (LSMDC 15) where we evaluated submissions using automatic and human evaluation criteria. We found that the approaches S2VT and our Visual-Labels generate longer and more diverse descriptions than the other submissions but are also more susceptible to content or grammatical errors. This consequently leads to worse human rankings with respect to correctness and grammar. In contrast, Frame-Video-Concept Fusion wins the challenge by predicting medium length sentences with intermediate diversity, which gets rated best in human evaluation for correctness, grammar, and relevance. When ranking sentences with respect to the criteria "helpful for the blind", our Visual-Labels is well received by human judges, likely because it includes important aspects provided by the strong visual labels. Overall all approaches have problems with the challenging long-tail distributions of our data. Additional training data cannot fully ameliorate this problem because a new movie might always contain novel parts. We expect new techniques, including relying on different modalities, see e.g. Hendricks et al. (2016), to overcome this challenge.

The second edition of our challenge (LSMDC 16) was held at ECCV 2016. This time we introduced a new human evaluation protocol to allow comparison of a large number of approaches. We found that the best approach in the new evaluation with the "helpful for the blind" criteria is a retrieval-based approach from Tel Aviv University. Likely,

human judges prefer the rich while also grammatically correct descriptions provided by this method. In the future work the movie description approaches should aim to achieve rich yet correct and fluent descriptions. Our evaluation server will continue to be available for automatic evaluation.

Our dataset has already been used beyond description, e.g. for learning video-sentence embeddings or for movie question answering. Beyond our current challenge on single sentences, the dataset opens new possibilities to understand stories and plots across multiple sentences in an open domain scenario on a large scale.

Acknowledgements Open access funding provided by Max Planck Society. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *European conference on computer vision (ECCV)*.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Ballas, N., Yao, L., Pal, C., & Courville, A. (2016). Delving deeper into convolutional networks for learning video representations. In *International conference on learning representations (ICLR)*.

Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S. et al. (2012). Video in sentences out. In *Proceedings of the conference on Uncertainty in artificial intelligence (UAI)*.

Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2013). Finding actors and actions in movies. In *International conference on computer vision (ICCV)*.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In *European conference on computer vision (ECCV)*.

Bruni, M., Uricchio, T., Seidenari, L., & Del Bimbo, A. (2016). Do textual descriptions help action recognition? In *Proceedings of the ACM on multimedia conference (MM)*, pp. 645–649.

Chen, D. & Dolan, W. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Chen, X, & Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *Conference on computer vision and pattern recognition (CVPR)*.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325.

Cour, T., Jordan, C., Miltsakaki, E., & Taskar, B. (2008). Movie/script: Alignment and parsing of video and text transcription. In *European conference on computer vision (ECCV)*.

Cour, T., Sapp, B., Jordan, C., & Taskar, B. (2009). Learning from ambiguously labeled images. In *Conference on computer vision and pattern recognition (CVPR)*.

Das, D., Martins, A. F. T., & Smith, N. A. (2012). An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Das, P., Xu, C., Doell, R., & Corso, J. (2013). Thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Conference on computer vision and pattern recognition (CVPR)*.

de Melo, G., & Tandon, N. (2016). Seeing is believing: The quest for multimodal knowledge. *SIGWEB Newsletter*, (Spring). doi:10.1145/2903513.2903517.

Del Corro, L., & Gemulla, R. (2013). Clausie: Clause-based open information extraction. In *Proceedings of the international world wide web conference (WWW)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Conference on computer vision and pattern recognition (CVPR)*.

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X. et al. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. et al. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Conference on computer vision and pattern recognition (CVPR)*.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., & Ponce, J. (2009). Automatic annotation of human actions in video. In *International conference on computer vision (ICCV)*.

Elliott, D., & Keller, F. (2013). Image description using visual dependency representations. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pp. 1292–1302.

Everingham, M., Sivic, J., & Zisserman, A. (2006). "hello! my name is... buffy"—Automatic naming of characters in tv video. In *Proceedings of the british machine vision conference (BMVC)*.

Fang, H., Gupta, S., Iandola, F. N., Srivastava, R., Deng, L., Dollár, P. et al. (2015). From captions to visual concepts and back. In *Conference on computer vision and pattern recognition (CVPR)*.

Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. et al. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision (ECCV)*.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: The MIT Press.

Gagnon, L., Chapdelaine, C., Byrns, D., Foucher, S., Heritier, M., & Gupta, V. (2010). A computer-vision-assisted system for videodescription scripting. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR workshops)*.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T. et al. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *International conference on computer vision (ICCV)*.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In *Conference on computer vision and pattern recognition (CVPR)*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoffman, J., Guadarrama, S., Tzeng, E., Donahue, J., Girshick, R., Darrell, T., & Saenko, K. (2014). LSDA: Large scale detection through adaptation. In *Conference on neural information processing systems (NIPS)*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Conference on computer vision and pattern recognition (CVPR)*.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006). Extending verbnet with novel verb classes. In *Proceedings of the international conference on language resources and evaluation (LREC)*.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal neural language models. In *International conference on machine learning (ICML)*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics (TACL)*, *14*, 595–603.

Klein, B., Lev, G., Sadeh, G., & Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. In *Conference on computer vision and pattern recognition (CVPR)*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N. et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision (IJCV)*, *50*(2), 171–184.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Conference on neural information processing systems (NIPS)*.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *Conference on computer vision and pattern recognition (CVPR)*.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Kuznetsova, P., Ordonez, V., Berg, T. L., Hill, UNC Chapel, & Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. In *Proceedings of the Transactions of the association for computational linguistics (TACL)*.

Lakritz, J. & Salway, A. (2006). The semi-automatic generation of audio description from screenplays. Technical report, Department of Computing Technical Report, University of Surrey.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on computer vision and pattern recognition (CVPR)*.

Lev, G., Sadeh, G., Klein, B., & Wolf, L. (2015). RNN fisher vectors for action recognition and image annotation. In *European conference on computer vision (ECCV)*.

Li, G., Ma, S., & Han, Y. (2015). Summarization-based video caption via deep neural networks. In *Proceedings of the 23rd annual ACM conference on multimedia conference*.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale N-grams. In *Proceedings of the fifteenth conference on computational natural language learning (CoNLL)*. Association for Computational Linguistics.

Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., & Luo, J. (2016). TGIF: A new dataset and benchmark on animated GIF description. In *Conference on computer vision and pattern recognition (CVPR)*.

Liang, C., Xu, C., Cheng, J., & Lu, H. (2011). Tvparser: An automatic tv video parsing method. In *Conference on computer vision and pattern recognition (CVPR)*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pp. 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D. et al. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2015). Deep captioning with multimodal recurrent neural networks (m-rnn). In *International conference on learning representations (ICLR)*.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Conference on computer vision and pattern recognition (CVPR)*.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X. et al. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the conference of the European chapter of the association for computational linguistics (EACL)*.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Conference on neural information processing systems (NIPS)*.

Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Shaw, B., Smeaton, A. F., & Quéenot, G. (2012). Trecvid 2012—An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA.

Pan, P., Xu, Z., Yang, Y., Wu, F., & Zhuang, Y. (2016a). Hierarchical recurrent neural encoder for video representation with application to captioning. In *Conference on computer vision and pattern recognition (CVPR)*.

Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016b). Jointly modeling embedding and translation to bridge video and language. In *Conference on computer vision and pattern recognition (CVPR)*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.

Ramanathan, V., Joulin, A., Liang, P., & Fei-Fei, L. (2014). Linking people in videos with "their" names using coreference resolution. In *European conference on computer vision (ECCV)*.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, *1*, 25–36.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German conference on pattern recognition (GCPR)*.

Rohrbach, A., Rohrbach, M., & Schiele, B. (2015a). The long-short story of movie description. arXiv:1506.01698.

Rohrbach, A., Rohrbach, M., & Schiele, B. (2015b). The long-short story of movie description. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015c). A dataset for movie description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B. (2013). Translating video content to natural language descriptions. In *International conference on computer vision (ICCV)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Salway, A. (2007). A corpus-based analysis of audio description. In *Media for all: Subtitling for the deaf, audio description and sign language* (pp. 151–174).

Salway, A., Lehane, B., & O'Connor, N. E. (2007). Associating characters with events in films. In *Proceedings of the ACM international conference on image and video retrieval (CIVR)*.

Schuler, K. K., Korhonen, A., & Brown, S. W. (2009). Verbnet overview, extensions, mappings and applications. In *Proceedings of the conference of the North American chapter of the association for computational linguistics (NAACL)*.

Shetty, R., & Laaksonen, J. (2015). Video captioning with recurrent networks based on frame-and video-level features and visual content classification. arXiv:1512.02949.

Shetty, R., & Laaksonen, J. (2016). Frame-and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the ACM on multimedia conference (MM)*, pp. 1073–1076.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR)*.

Sivic, J., Everingham, M., & Zisserman, A. (2009). "who are you?"-learning person specific classifiers from video. In *Conference on computer vision and pattern recognition (CVPR)*.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics (TACL)*, *2*, 207–218.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Conference on computer vision and pattern recognition (CVPR)*.

Tandon, N., de Melo, G., De, A., & Weikum, G. (2015). Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings on CIKM*.

Tapaswi, M., Baeuml, M., & Stiefelhagen, R. (2012). "knock! knock! who is it?" probabilistic person identification in tv-series. In *Conference on computer vision and pattern recognition (CVPR)*.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Conference on computer vision and pattern recognition (CVPR)*.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., & Mooney, R. J. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the international conference on computational linguistics (COLING)*.

Torabi, A., Pal, C., Larochelle, H., & Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. arXiv:1503.01070v1.

Torabi, A., Tandon, N., & Sigal, L. (2016). Learning language-visual embedding for movie understanding with natural-language. arXiv:1609.08124.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology*. Association for Computational Linguistics.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Conference on computer vision and pattern recognition (CVPR)*.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015a) Sequence to sequence—video to text. arXiv:1505.00487v2.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015b) Sequence to sequence—video to text. In *International Conference on Computer Vision (ICCV)*.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2015c). Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the conference of the North American chapter of the association for computational linguistics (NAACL)*.

Venugopalan, S., Hendricks, L. A., Mooney, R., & Saenko, K. (2016). Improving LSTM-based video description with linguistic knowledge mined from text. arXiv:1604.01729.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, H. & Schmid, C. (2013). Action recognition with improved trajectories. In *International conference on computer vision (ICCV)*.

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)*, *103*(1), 60–79.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on computer vision and pattern recognition (CVPR)*.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Conference on computer vision and pattern recognition (CVPR)*.

Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*.

Xu, R., Xiong, C., Chen, W., & Corso, J. J. (2015b). Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Conference on artificial intelligence (AAAI)*.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *International conference on computer vision (ICCV)*.

Yao, L., Ballas, N., Cho, K., Smith, J. R., & Bengio, Y. (2016). Empirical performance upper bounds for image and video captioning. In *International conference on learning representations (ICLR)*.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, *2*, 67–78.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016a). Video paragraph captioning using hierarchical recurrent neural networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Yu, Y., Ko, H., Choi, J., & Kim, G. (2016b). Video captioning and retrieval models with semantic attention. arXiv preprint arXiv:1610.02947.

Zeng, K.-H., Chen, T.-H., Niebles, J. C., & Sun, M. (2016). Title generation for user generated videos. In *European conference on computer vision*.

Zhong, Z., & Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Conference on neural information processing systems (NIPS)*.

Zhu, L., Xu, Z., Yang, Y., & Hauptmann, A. G. (2015a). Uncovering temporal context for video question and answering. arXiv:1511.04670.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015b). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International conference on computer vision (ICCV)*.