Noname manuscript No.
(will be inserted by the editor)

MSVD-Turkish: A Comprehensive Multimodal Video Dataset for Integrated Vision and Language Research in Turkish

Begum Citamak · Ozan Caglayan · Menekse Kuyu · Erkut Erdem* · Aykut Erdem · Pranava Madhyastha · Lucia Specia

Received: date / Accepted: date

Abstract Automatic generation of video descriptions in natural language, also called *video captioning*, aims to understand the visual content of the video and produce a natural language sentence depicting the objects and actions in the scene. This challenging integrated vision and language problem, however, has been predominantly addressed for English. The lack of data and the linguistic properties of other languages limit the success of existing approaches for such languages. In this paper we target Turkish, a morphologically rich and agglutinative language that has very different properties compared to English. To do so, we create the first large-scale video captioning dataset for this language by carefully translating the English descriptions of the videos in the MSVD (Microsoft Research Video Description Corpus) dataset into Turkish. In addition to enabling research in video captioning in Turkish, the parallel English-Turkish descriptions also enable the study of the role of video context in (multimodal) machine translation. In our experiments, we build models for both video captioning and multimodal machine translation and investigate the effect of different word segmentation approaches and different neural architectures to better address the properties of Turkish. We hope that the MSVD-Turkish dataset and the results reported in this work will lead to bet-

* Corresponding author.

Begum Citamak · Menekse Kuyu · Erkut Erdem
Department of Computer Engineering, Hacettepe University, Ankara/Turkey
E-mail: {n16221821, n17132209, aykut, erkut} @ cs.hacettepe.edu.tr

Aykut Erdem
Department of Computer Engineering, Koç University, Istanbul/Turkey
E-mail: aerdem @ ku.edu.tr

Ozan Caglayan · Pranava Madhyastha · Lucia Specia
Department of Computing, Imperial College London, UK
E-mail: {o.caglayan, pranava, l.specia} @ imperial.ac.uk

ter video captioning and multimodal machine translation models for Turkish and other morphology rich and agglutinative languages.

Keywords Video description dataset · Turkish · Video captioning · Video understanding · Neural machine translation · Multimodal machine translation

1 Introduction

Recent developments in computer vision (CV) and natural language processing (NLP) have led to a surge of new problems which lie at the intersection of these two fields, creating a new area of research in general called "integrated vision and language" (iVL). Video captioning is one of the important problems in iVL research, which has gained significant attention in both the CV and NLP communities. It aims at understanding the visual content of a given video clip and contextually generating a natural language description of this clip.

Although a considerable amount of literature has revolved around this challenging task in recent years, all existing work is monolingual that it has mainly focused on the English language. Hence, whether or not the state-of-the-art video captioning methods can be effectively adapted to languages other than English, especially for low-resource languages, remains an open problem. Moreover, linguistic differences between English and other languages, particularly those that are morphologically richer than English, introduce new challenges that need to be addressed. Before these questions can be answered, however, we require video datasets containing descriptions from languages other than English to further enable iVL research.

As a first step towards this direction, in this paper, we extend the MSVD (Microsoft Research Video Description Corpus) (Chen and Dolan, 2011) dataset and introduce a new multilingual dataset that we call MSVD-Turkish which contains approximately 2k video clips and a total of 80k Turkish video descriptions. In particular, we collect these Turkish descriptions by manually translating the original English video descriptions from MSVD into Turkish. Compared to the original English descriptions, Turkish descriptions have a larger vocabulary size and more importantly reflect the highly inflected and highly agglutinative nature of Turkish.

We demonstrate the multilingual, multimodal capabilities of the proposed MSVD-Turkish dataset, by exploring two distinct iVL tasks shown in Figure 1, namely video captioning and multimodal machine translation (MT), but with a special focus on Turkish. As far as we are aware of, this work is the first to investigate generating Turkish descriptions depicting visual content of videos. To this end, we analyse different segmentation strategies for Turkish. Additionally, we explore multimodal MT as the second task on MSVD-Turkish where we examine the use of supplementary visual cues within videos to potentially improve the translation quality. Our primary contributions in this paper are:

– To foster research in multilingual, multimodal language generation, we collect a new large-scale dataset called MSVD-Turkish by translating the En-

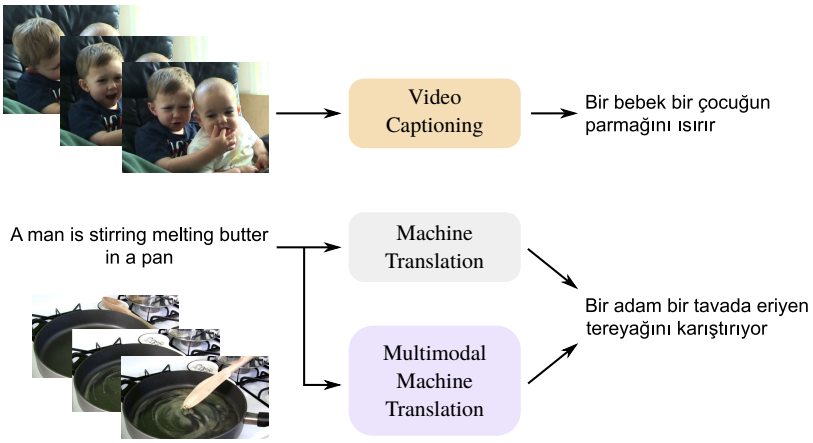

Fig. 1 The depiction of the video captioning and MT tasks on the MSVD-Turkish Dataset.

glish descriptions of the videos from the well-known MSVD dataset into Turkish.
– We investigate the performance of several (multimodal) MT and video captioning models on the proposed MSVD-Turkish dataset.
– To address the rich and agglutinative morphology of Turkish, we explore different word segmentation approaches in Turkish.

The paper is organised as follows: In Section 2, we briefly review the state-of-the-art in multimodal MT and video captioning. In Section 3, we introduce the MSVD-Turkish dataset, examine our data collection strategy and provide some statistics regarding the dataset. We introduce the details regarding the visual and textual representations considered in our machine translation and video captioning models in Section 4, and describe the models themselves in Section 5. In Section 6, we present our experimental results and discuss our findings and finally, we provide a summary of our work and discuss possible future research directions in Section 7.

2 Related Work

In the following, we review the efforts towards two related tasks within the integrated vision and language research, namely multimodal MT and video captioning.

2.1 Multimodal Machine Translation

The predominant approaches in state-of-the-art in MT use neural models (NMT) which consist of an encoder to map a given sentence into a latent

representation, and a decoder to map this representation into a translation in the target language (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). NMT models are trained with maximum likelihood estimation (MLE), i.e. the training objective maximises the likelihood of source-target training pairs.

The success of such approaches has led to a rising interest in more sophisticated NMT architectures that can handle multiple input/output modalities, a framework often referred to as "multimodal machine translation" (MMT). MMT seeks to improve translation quality by taking into account visual (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) or speech modality (Sulubacak et al., 2019). The prominent *end-to-end* approaches to MMT with visual information can be divided into two main categories:

1. **Multimodal attention** extends the classical attention mechanism (Bahdanau et al., 2015) applied on top of the textual representations, with a spatial one (Xu et al., 2015) applied to convolutional feature maps. Specifically, Caglayan et al. (2016a) explore a *shared attention* across the modalities while Calixto et al. (2016) experiment with *dedicated attention*. Caglayan et al. (2016b) later propose several variants where the level of parameter sharing across modality attentions is configurable. In all these models, the outputs of attention mechanisms are simply fused together via addition or concatenation. Libovický and Helcl (2017) replace this step with another attention layer which could in theory, selectively integrate information coming from different modalities. Huang et al. (2016) do not implement a fully multimodal attention in the decoder but enrich the source word embedding sequence with visual feature vectors, in the hope that the decoder attention will learn to pay attention to visual feature vectors when needed.
2. **Simple conditioning** makes use of non-spatial features such as the fully-connected (`FC`) layer features for VGG-style (Simonyan and Zisserman, 2014) and global average pooled features for ResNet-style (He et al., 2016) networks. Specifically, a *single* feature vector per image is used to condition arbitrary layers in the network, with the objective of learning *grounded* textual representations. The visual conditioning is often performed through (i) initializing the hidden state of the recurrent encoders and/or decoders with the visual features (Calixto and Liu, 2017), (ii) multiplicative interactions of source and/or target embeddings with the visual features (Caglayan et al., 2017a), and (iii) the use of auxiliary training objectives such as the *Imagination* architecture (Elliott and Kádár, 2017) which tries to reconstruct the visual features from the textual encoder states.

2.2 Video Captioning

Video captioning aims at generating a single sentence description from a short video clip summarising the actors and actions depicted in the clip. It involves unique challenges over image captioning, since it additionally requires analyzing the temporal evolvement of concepts and their relationships. The methods

proposed for this task can be, in general, categorised into three classes (for a more thorough review, please refer to Aafaq et al. (2019)):

1. **Classical video captioning** approaches include the early works that integrate the traditional computer vision and NLP techniques (Hakeem et al., 2004; Barbu et al., 2012; Hanckmann et al., 2012; Kojima et al., 2012; Guadarrama et al., 2013; Krishnamoorthy et al., 2013; Thomason et al., 2014). These methods commonly generate a description of a clip in two phases. In the first phase, they generally detect the most important objects, recognise their actions and extract object-object interactions along with the scene information. The second phase employs these extracted visual entities and some rule-based, pre-defined sentence templates to construct video descriptions. While this strategy generates grammatically sound sentences, the sentences lack naturalness and more importantly become too constrained for open-domain videos.
2. **Statistical video captioning** methods such as Rohrbach et al. (2013) have been proposed to fill in this gap by additionally taking into account some statistical cues while generating a natural language description of a given input video. Accordingly, this provides more accurate and natural depictions compared to the classical approaches.
3. **Deep video captioning** approaches which are specifically motivated from the recent NMT models. They all consider two sequential stages which are realised with an encoder-decoder architecture. The basic difference between these deep learning models and the first two groups of work lies in how they represent the visual content. While the earlier approaches employ recognition and detection methods to extract a set of word tokens, deep models represent the video in terms of a vector representation, either with a fixed or dynamic embedding. Deep learning-based video captioning models can be categorised into further groups by their encoder-decoder structure and by their learning methodology. For instance, the most common model architecture (Donahue et al., 2015; Venugopalan et al., 2015; Yao et al., 2015) employs convolutional neural networks to extract visual content in the encoding stage and a recurrent neural network in the decoding phase to perform the video-driven sentence generation. Some other work (Srivastava et al., 2015; Yu et al., 2016) extends this structure by considering recurrent neural networks in both encoding and decoding stages. The final group of studies includes reinforcement learning based video captioning models (Chen et al., 2018; Wang et al., 2018).

3 MSVD-Turkish Dataset

3.1 Overview

Existing datasets for video captioning typically contain short video clips (a few seconds in duration) and descriptions depicting the content of videos in a natural language. The early examples such as MPII Cooking (Rohrbach et al., 2012), YouCook (Das et al., 2013), TACoS (Regneri et al., 2013), TACoS Multi-Level (Senina et al., 2014), and YouCook II (Zhou et al., 2017) include videos about everyday actions, which were usually collected from video-sharing sites such as YouTube by querying keywords related to cooking. In contrast, other datasets such as MSVD (Chen and Dolan, 2011), M-VAD (Torabi et al., 2015), MPII-Movie Description (MPII-MD) (Rohrbach et al., 2015), TGIF (Li et al., 2016), MSR-VTT (Xu et al., 2016), VTW (Zeng et al., 2016), Charades (Sigurdsson et al., 2016), LSMDC (Rohrbach et al., 2017), ActyNet-Cap (Krishna et al., 2017), ANet-Entities (Zhou et al., 2019), and VideoStory (Gella et al., 2018) are open-domain datasets. M-VAD, MPII-MD, and LSMDC datasets differ from the others in that they contain movie-clip descriptions constructed by professionals for descriptive video service purposes. In recent years, with the increase in the use of social media platforms, social media has become a major source of data, and ANet-Entities (Zhou et al., 2019) and VideoStory (Gella et al., 2018) datasets include videos shared in these mediums. Table 1 summarises the characteristics of these datasets in detail. In particular, for each dataset, we report the domain of the videos, the number of action classes seen in the videos along with some statistics such as the number of videos/clips and their average length. We also provide the number of sentences/words and the size of the vocabulary.

It is important to mention that all these datasets are monolingual and contain only English descriptions.[1] The only exception is the recently proposed VaTeX dataset (Wang et al., 2019), which has both Chinese and English descriptions for each video clip. Even for image captioning which has been studied more extensively than video captioning, multilingual datasets are scarce. There exist only a few datasets such as (i) the TasvirEt dataset (Unal et al., 2016) which extends the original Flickr8k dataset with two crowdsourced Turkish descriptions per image, and (ii) the STAIR dataset (Yoshikawa et al., 2017) which provides five crowd-sourced Japanese descriptions for 164,062 MSCOCO (Lin et al., 2014) images. Similarly, the image-based MMT task requires a multimodal dataset with images and their (translated) descriptions in at least two languages. The well-known Multi30k dataset (Elliott et al., 2016) fulfilled this requirement by augmenting the popular image captioning dataset Flickr30k (Plummer et al., 2015), with German, French and Czech descriptions that are direct translations of the original Flickr30k English descriptions.

[1] Note that the original MSVD dataset also contains annotations obtained for many different languages. The number of these multilingual descriptions is, however, very low compared to the number of original English descriptions. Moreover, these descriptions were not shared with the community.

Table 1 Statistics of Video Captioning Datasets

Dataset	Domain	Classes	Videos	Avg len	Clips	Sents	Words	Vocab
MPII Cooking	cooking	65	44	600 sec	-	5,609	-	-
YouCook	cooking	6	88	-	-	2,688	42,457	2,711
TACoS	cooking	26	127	360 sec	7,206	18,227	146,771	28,292
TACos-MLevel	cooking	67	185	360 sec	14,105	52,593	2,000	-
MPII-MD	movie	-	94	3.9 sec	68,337	68,375	653,467	24,549
M-VAD	movie	-	92	6.2 sec	48,986	55,904	519,933	17,609
MSR-VTT	open	257	7,180	20 sec	10k	200k	1,856,523	29,316
Charades	open	157	9,848	30 sec	9,848	27,847	-	-
VTW	open	-	18,100	90 sec	18,100	44,613	-	-
YouCook II	cooking	89	2,000	316 sec	15.4k	15.4k	-	2,600
ActyNet Cap	open	200	20k	180 sec	100k	100k	1,348,000	-
ANet-Entities	social media	-	14,281	180 sec	52k	-	-	-
VideoStory	social media	-	20k	-	123k	123k	-	-
VaTeX-English	open	600	41.3k	10 sec	41.3k	826k	12,580,000	11k
VaTeX-Chinese	open	600	41.3k	10 sec	41.3k	826k	11,523,000	14k
MSVD(-English)	open	218	1970	10 sec	1,970	80,827	567,874	12,592
MSVD-Turkish	open	218	1970	10 sec	1,970	80,676	432,250	18,312

Multi30k is so far the only dataset which provides actual translations aligned to images, rather than independent descriptions as in TasvirEt and STAIR datasets.

3.2 Data Collection

In this study, we aim to contribute to this new area of research, *multilingual video captioning*, by collecting a large video dataset consisting of videos and their English and Turkish descriptions. We believe that selecting Turkish will fill an important gap for the analysis of morphologically-rich and low-resource languages in the video-captioning literature. We name our dataset as MSVD-Turkish, after the MSVD dataset. Since MSVD-Turkish has parallel Turkish-English sentences, it can be used not only for video captioning task but also for multimodal MT.[2]

In the data collection phase, we first translated the English captions into Turkish with the free Google Translate API. Using automatic systems for data annotation has become a common practice in the community to obtain large-scale training data for deep learning approaches. Needless to say, this strategy comes with its own drawbacks. For instance, in our case, the major risk of using automatic translation systems for constructing Turkish descriptions is that the generated translations could be of low-quality. We observed that in some of the translations, suffixes were incorrectly attached or they were completely

[2] We make our dataset publicly available at `https://hucvl.github.io/MSVD-Turkish/`.

MSVD Descriptions	MSVD-Turkish Descriptions
a woman is cutting a vegetable	bir kadın bir sebze kesiyor
a man cuts a piece of zucchini in half and slices it thinly	bir adam bir parça kabağı ikiye keser ve ince dilimler
a person is slicing cucumbers	bir kişi salatalık dilimliyor
someone sliced a cucumber or a zucchini into small pieces	birisi küçük parçalar halinde bir salatalık ya da bir kabak dilimlenmiş
a person is slicing a cucumber into pieces	bir kişi bir salatalığı parçalara ayırıyor
a chef is slicing a vegetable	bir şef bir sebze dilimliyor
a person is slicing a cucumber	bir kişi bir salatalık dilimliyor
a woman is slicing vegetables	bir adam sebze dilimliyor
a woman is slicing a cucumber	bir kadın bir salatalık dilimliyor
a person is slicing cucumber with a knife	bir kişi bir bıçakla salatalık dilimliyor

Fig. 2 A sample clip and its English captions with the corresponding Turkish translations from the MSVD-Turkish dataset.

missing. Moreover, there were some translation issues regarding ambiguous words.

To ensure the adequacy of the computer-generated descriptions in our MSVD-Turkish dataset, we gathered two bilingual M.Sc. students who speak Turkish and English and have some experience on image captioning. They controlled the automatically translated sentences for obvious errors and made the necessary corrections. During this step, we noticed that some of the English descriptions from the original MSVD dataset were very noisy, with some not even in English. We left those descriptions out and did not translate them into Turkish. In Figure 2, we depict an example video clip with the original English descriptions from MSVD and their translations into Turkish as provided in the resulting MSVD-Turkish dataset.

Moreover, after the data collection step, we performed a user study to further assess the quality of the Turkish descriptions. In particular, we randomly selected 100 video clips from our dataset, and for each selected video clip, we sampled a single Turkish description from a pool. We then asked each user to evaluate a given Turkish description from three different aspects: translation quality (*how good is the translation from English to Turkish?*), relevance (*how relevant is the Turkish description to the given image?*), and fluency (*how grammatically correct is the Turkish description?*). For simplicity, the users rated each aspect using a 5-point Likert scale ranging from 'excellent' (score 5) to 'poor' (score 1).

Table 2 presents the results of our user study in which 15 different users participated. These results are encouraging in that the average translation quality and grammatical correctness of Turkish descriptions seem to be fairly good and the relevance between the videos and their descriptions is highly satisfactory. To compute the statistical measure of agreement we further obtain the Krippendorff's alpha coefficient (Krippendorff, 1970) and observe that the inter-annotator agreement for the translation quality study is $\alpha = 0.389$.

Table 2 Human evaluation of automatically translated Turkish descriptions. For each aspect, mean scores are given along with the standard deviations (σ).

Translation Quality	Relevance	Fluency
4.46 (σ = .83)	4.59 (σ = .82)	4.57 (σ = .76)

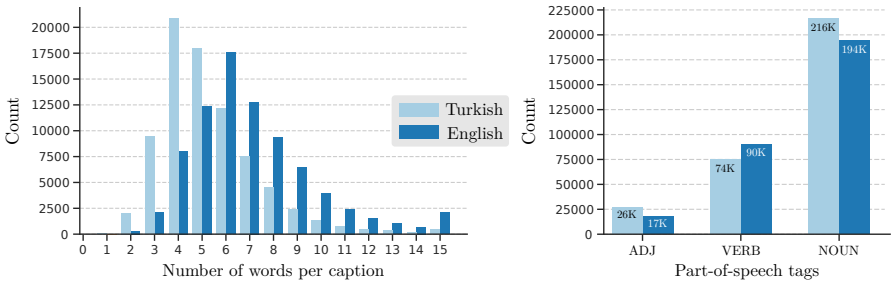

Fig. 3 Description length (left) and part-of-speech tag distribution (right) across the MSVD and MSVD-Turkish datasets.

3.3 Data Analysis

The original MSVD dataset contains a total of 1,970 video clips collected from YouTube and on average, each clip is 10 seconds long (Table 1). The training, validation and test splits of MSVD contain 1200, 100 and 670 videos, respectively. The number of descriptions per video is variable and has a mean of 40.9 (σ = 10.3). The number of descriptions in MSVD-Turkish is slightly lower due to some of the original noisy captions being excluded for translation. Therefore, we also provide a fully parallel version with 80,673 aligned descriptions in total. In the left panel of Figure 3, we compare the distribution of number of words between English and Turkish, and observe that the distribution is *right-skewed* for MSVD-Turkish. This is expected to some extent due to the rich and agglutinative morphology of Turkish i.e. individual words can represent multi-word expressions in English. This is also reflected in Table 1 where MSVD-Turkish is shown to have a substantially larger vocabulary than MSVD (18,312 compared to 12,592). Finally, the average caption lengths for English and Turkish are 7.1 and 5.4 words, respectively.

We further perform part-of-speech tagging to extract the distribution of adjectives, verbs and nouns across the English and Turkish descriptions. For this purpose, we utilise the `Stanza` toolkit (Qi et al., 2020) for English and the `Zemberek` toolkit (Akin and Akin, 2007) for Turkish. We lowercase all descriptions and remove punctuation for consistency. The right panel of Figure 3 shows the final counts for the selected part-of-speech tags where we see that MSVD-Turkish descriptions have more adjectives and nouns than the original English descriptions.

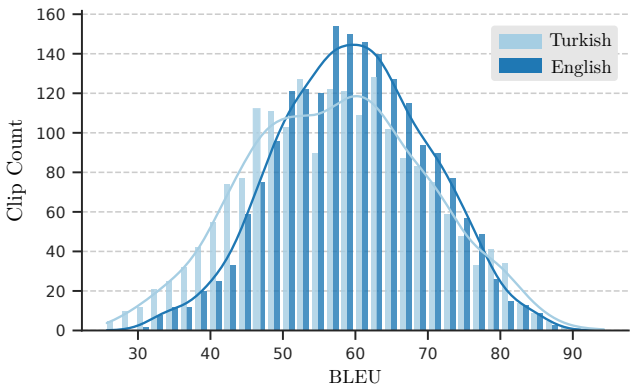

Fig. 4 The distribution of clip-level paraphrase BLEU scores: the mean BLEU across all video clips is 59.8 and 57.5 for English and Turkish, respectively. A lower score denotes a higher lexical diversity.

3.3.1 Paraphrasing Quality

The primary motivation behind the creation of the MSVD dataset was the collection of a highly parallel corpus for paraphrase evaluation (Chen and Dolan, 2011). In this experiment, we assess the impact of our data collection procedure on paraphrasing quality. Specifically, we use the `sacreBLEU` utility (Post, 2018) to measure the sentence-level BLEU (Papineni et al., 2002) scores across each clip's description set. More formally, we compute the average BLEU score B_i within the description set $\{C_i\}$ of the video clip i as follows:

$$B_i = \frac{1}{|C_i|} \sum_{j}^{|C_i|} \text{BLEU}(\text{candidate} = C_{ij}, \text{references} = \{C_{ik} | k \neq j\}))$$

Since BLEU measures the n-gram overlap between a candidate sentence and the set of corresponding references, a *lower* BLEU suggests higher lexical variability for a given video's description set. Figure 4 plots the histogram of clip-level BLEU scores across MSVD and MSVD-Turkish. The distribution shows that the English descriptions for a given video are more lexically similar to each other than the Turkish counterparts. We especially observe that a substantial amount of clip descriptions in Turkish obtain a BLEU score of less than 50 whereas for English, the distribution is centered around 60. In other words, the paraphrasing quality of the original English descriptions does not seem to decrease when they are automatically translated to Turkish.

4 Modality Representations

Both the video captioning and MT tasks can be cast as sequence to sequence (S2S) problems, and they involve processing of visual and/or textual infor-

mation. In the following, we briefly explain our choices for video and textual representations.

4.1 Visual representations

We represent each video clip with a fixed-length vector by using an ImageNet pre-trained VGG16 network (Simonyan and Zisserman, 2014). In particular, we sample 30 equidistant frames from each video and feed them to VGG16 to extract 4096-dimensional `FC7` feature vectors. The final video representation that will be used by the video captioning models is this temporal feature matrix of size 30×4096.

4.2 Textual representations

In addition to the default setting of using words for both languages, we conduct several experiments using two widely-known unsupervised segmentation techniques. The motivation behind these experiments is to understand what kind of token representation is ideal for video captioning and MT into Turkish, an agglutinative language with rich morphology.

The first segmentation approach makes use of the so-called "byte-pair encoding (BPE)" algorithm, which proceeds by deconstructing sentences to characters and then works out in a bottom-up manner to recursively merge frequent byte pairs altogether (Sennrich et al., 2016). The final "open vocabulary" is biased towards keeping frequent words intact while splitting out rare words into frequent subwords. The main hyperparameter of the BPE algorithm determines how many merge operations will be done during the learning step, which approximately reflects the final vocabulary size. In contrast to the BPE algorithm which is deterministic, `SentencePiece` (SPM) employs a unigram language modeling approach to maximize the likelihood of a given corpus with the probability of each sentence defined as the sum of its candidate segmentations (Kudo and Richardson, 2018). Overall, this yields a probabilistic mixture model from which it is possible to sample arbitrarily many multi-level segmentations (characters, subwords and words) for a given sentence. SPM can also be applied to non-tokenized sentences, removing the necessity of using language-dependent tokenization and detokenization pipelines.

5 Tasks and Models

In this section, we present two tasks exploring the proposed bilingual MSVD dataset where the common objective is to generate natural language sentences in Turkish. We first start by describing the sequence-to-sequence framework within the context of monolingual and multimodal MT. We then present the

Task	Architecture	Input
Machine Translation (MT)	Recurrent NMT (Bahdanau et al., 2015)	English
	Transformer NMT (Vaswani et al., 2017)	
Multimodal MT	Recurrent NMT (Bahdanau et al., 2015)	
	+ Multimodal Att. (Caglayan et al., 2016a)	English + CNN Features
	+ Simple Conditioning (Calixto and Liu, 2017)	
Video Captioning	LSTM-based (Venugopalan et al., 2015)	CNN Features
	Transformer-based (Vaswani et al., 2017)	

Table 3 The summary of the tasks & models explored in our study.

video-captioning approaches that we follow, which can be considered as extensions to the S2S framework. We note that some design choices and hyperparameters are different between the neural architectures and the video captioning and MMT tasks because they were empirically selected for each architecture and task.

Our primary objectives here are to investigate (i) the performances of several different (multimodal) MT and video captioning models on MSVD-Turkish dataset, and (ii) to explore the usefulness of various word segmentation techniques to deal with the rich and agglutinative morphology of Turkish, in the context of MSVD-Turkish dataset. The summary of the explored tasks and their configurations are given in Table 3.

5.1 Neural Machine Translation

In what follows, we introduce our recurrent and transformer-based NMT models and their multimodal counterparts. To represent the words in the sentences, we experiment with the segmentation approaches previously mentioned in Section 4. To achieve multimodality, we provide frame-level image features extracted using a pre-trained CNN model (Section 4) as a secondary input modality along the English sentences (Figure 5). Model-specific details will be given in the respective subsections.

5.1.1 Recurrent NMT

We follow the attentive encoder-decoder approach (Bahdanau et al., 2015) for the recurrent models. The *attention mechanism* is crucial to obtain state-of-the-art results in NMT. This mechanism avoids encoding the whole source sentence into a single vector as in Sutskever et al. (2014) by looking to the latent encodings of the source sentence *at each timestep* of the decoder. In other words, the decoder is conditioned on a different representation of the source sentence (namely the context c_t) when generating target words, rather than reusing the same fixed-size encoding vector. The model thus estimates the probability of a target token y_t by conditioning on the previous target token y_{t-1} and the dynamic context c_t *i.e.* $P(y_t|y_{t-1}, c_t)$.

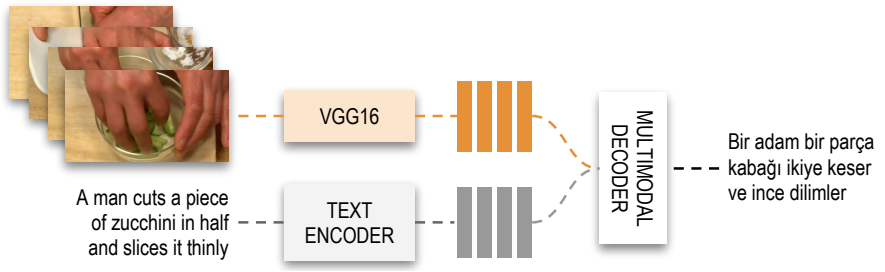

Fig. 5 Multimodal machine translation decoder on MSVD-Turkish.

Our model is composed of two bidirectional GRU (Cho et al., 2014) layers in the encoder. The Turkish decoder follows the *Conditional GRU* (CGRU) design (Sennrich et al., 2017) where the attention mechanism operates between two GRU layers. The input and output embeddings of the decoder are *tied* for parameter efficiency (Inan et al., 2016; Press and Wolf, 2017). The hidden dimensions of encoders, decoders and the attention mechanism are globally set to 320 while the source and target embeddings are 200-dimensional. Dropout (Srivastava et al., 2014) is employed at three places, namely, on top of the source embeddings ($p = 0.4$), encoder outputs ($p = 0.5$) and before the softmax layer ($p = 0.5$). We train the models using the open-source sequence-to-sequence toolkit `nmtpytorch` (Caglayan et al., 2017b). We evaluate model performance after each epoch by computing the BLEU score for the validation set translations. The learning rate is *halved* if the performance does not improve for two consecutive epochs. After five consecutive epochs with no improvement, the training is *stopped*. At test time, the translations are decoded using the best model checkpoint, with beam size set to 12.

5.1.2 Recurrent MMT

For multimodal MT, we experiment with two different visual integration approaches: (i) a dedicated multimodal attention mechanism (Caglayan et al., 2016a) and (ii) simple conditioning (Calixto and Liu, 2017) by initialising the encoders and the decoders of the MMT by the max-pooled visual feature vector.

For attentive MMT, we first project the visual features to the size of bidirectional textual encoder states. This way, the multimodal decoder receives two sets of compatible semantic representations for the input modalities. At each decoding timestep t, the model now computes an auxiliary visual context vector c'_t alongside the textual context c_t (§ 5.1.1). The distribution of target language tokens is now estimated by $P(y_t|y_{t-1}, f(c_t, c'_t))$ where the function $f()$ implements the concatenation operator (Caglayan, 2019).

Fig. 6 Architecture of the LSTM-based video captioning model.

5.1.3 Transformer-based NMT

Transformer-based NMT models (Vaswani et al., 2017) are feed-forward architectures which extend the idea of attention and avoid the need for recurrent layers. This has mainly two advantages: (i) it accommodates for more layers (depth) as the gradients will flow more easily than in recurrent NMT models, and (ii) the removal of the sequential dependence between hidden states allows for parallelised training. The expressiveness of recurrence is replaced with self-attention layers which takes into account all hidden representations at a given depth. Transformer NMT models are currently the state of the art, especially in large-scale NMT setups.

Since the dataset is relatively small, we use the base Transformer model with 6 encoders and 6 decoder layers, each having 4 attention heads. The model and feed-forward dimensions are set to 256 and 1024, respectively. Label smoothing with $\epsilon = 0.1$ is applied to the cross-entropy loss. The dropout rate is fixed as 0.3. We use the open-source `fairseq` toolkit (Ott et al., 2019) for training the models. The models are trained for a maximum of 50 epochs, and the test set translations are generated with beam size 12, using the checkpoint that achieved the lowest validation loss during training.

5.2 Video Captioning

Similar to NMT models, deep approaches to video captioning also employ encoder-decoder architectures. While the encoder module takes the individual frame representations as input and encodes them into a feature representation, the decoder generates a natural language description of the video by considering the encoded visual information. In this study, we test two popular types of network architectures, namely a recurrent one which uses the LSTM (Hochreiter and Schmidhuber, 1997) variant and a Transformer (Vaswani et al., 2017). As we mentioned in Section 4, we utilise the ImageNet pre-trained VGG16 CNN to encode the video frames, and as for the textual representation, we investigate different word segmentation strategies using SPM and BPE algorithms.

5.2.1 Recurrent Video Captioning

For our recurrent video captioning model, we adapt the architecture proposed by Venugopalan et al. (2015) in which the encoder and the decoder are implemented with two separate LSTM networks (Figure 6). The encoder computes a sequence of hidden states by sequentially processing the frame-level visual features, extracted from the uniformly sampled video frames. The decoder module then takes the *final hidden state* of the encoder, and outputs a sequence of tokens as the predicted video caption. There is no attention mechanism involved in this model. Both the encoder and decoder LSTM networks have 500 hidden units.

We use Adam (Kingma and Ba, 2014) as the optimiser and set the initial learning rate and batch size to 0.0004 and 32, respectively. We choose the models by using early stopping. In particular, we take into account the validation loss values to decide on the checkpoint that will be used to generate Turkish descriptions at inference time.

5.2.2 Transformer-based Video Captioning

Our Transformer-based video captioning model is built upon the base Transformer model (Vaswani et al., 2017). In the encoder, we first consider a linear transformation layer to project the extracted visual features to 512. We then treat these transformed features as our visual tokens, and consider positional encodings to preserve temporal information of the frames. The decoder module is responsible for generating a description conditioned on the input video frames encoded by the encoder. Figure 7 shows an illustration of our Transformer-based video captioning model.

We train the models using the `tensor2tensor` toolkit (Vaswani et al., 2018). We use the base Transformer model containing 3 encoder and 3 decoder layers, each with 8 attention heads, since the dataset contains few video samples. During training, we employ the cross entropy loss with label smoothing ($\epsilon = 0.1$) and a batch size of 1024. The dropout rate is fixed as 0.1. The model parameters are optimised using Adam by setting the initial learning rate to 0.0005. We employ approximate BLEU score for early stopping, and at test time, descriptions are obtained by using beam search with a beam size of 4.

6 Experimental Results

In this section we present quantitative and qualitative results for the MT and video captioning experiments. We begin by using the MT experiments as a guide to find the optimum segmentation strategy for Turkish, which is subsequently adopted for the multimodal MT and the video captioning experiments.

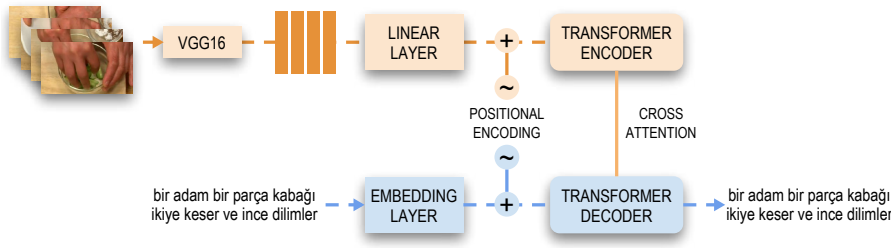

Fig. 7 Illustrative architecture of the Transformer-based video captioning model.

Table 4 BLEU comparison of English and Turkish segmentation choices with Recurrent NMT. A word→word baseline is also provided for comparison purposes. The segmentation models are learned on the training set of the MSVD dataset.

	Word→Word	**3K→3K**	**3K→6K**	**6K→3K**	**6K→6K**	**Avg**
Words	**36.2** ±0.3	–	–	–	–	–
BPE	–	35.4 ±0.4	34.5 ±0.7	36.1 ±0.3	35.1 ±0.6	35.3
SPM	–	35.9 ±0.4	35.9 ±0.3	36.0 ±0.2	35.3 ±0.2	35.8

6.1 Segmentation & Machine Translation

We first focus our attention on the word segmentation problem. For that, we use of the parallel captions from the MSVD-Turkish dataset and learn separate English and Turkish segmentation models using the BPE and the SPM approaches. We explore four settings where the size of the final vocabularies would be approximately 3K and 6K tokens. We train three recurrent NMT systems per each segmentation setup and report the mean and the standard deviation of the test set BLEU scores. We also train a non-segmented word→word baseline for comparison. Table 4 presents the results for these experiments. We observe that in almost all cases SPM outperforms BPE, with the average gap being 0.5 BLEU in favor of SPM. Surprisingly, the best average performance of 36.2 is obtained when words are used without any segmentation at all.

6.1.1 Pre-trained segmentation models

These results do not favor any of the proposed segmentation approaches. We posit that this may be because of limitations of the segmentation models themselves. Therefore, we direct our attention to pre-training Turkish BPE and SPM models on a large external corpus instead of the small set of MSVD-Turkish captions. This way, we expect to learn slightly more linguistically sound segmentation models. For this purpose, we used a snapshot (2019-07-

Table 5 BLEU comparison of Turkish Wikipedia pre-trained segmentation models for recurrent and Transformer NMT models. The English vocabulary is fixed to word units and contains 9,321 words in total. The corresponding Turkish vocabulary sizes are given in parentheses.

	→Word (13400)	**→BPE30K (8400)**	**→SPM30K (7900)**
Transformer NMT	35.8 ±0.2	36.5 ±0.2	36.7 ±0.1
Recurrent NMT	36.2 ±0.3	36.8 ±0.3	**37.0** ±0.6

Table 6 Multimodal machine translation results on MSVD-Turkish. The Turkish vocabulary uses `SPM30K` learned on Wikipedia and English vocabulary consists of words.

Model	**# Params**	**BLEU**
Recurrent NMT	8.3M	**37.0** ±0.6
+ Multimodal Attention	11.3M	36.5 ±0.1
+ Enc-Dec Initialisation	9.4M	36.7 ±0.1

20) of the official Turkish Wikipedia dump, pre-processed it[3] and ended up with 955K Turkish sentences.

To focus on the effect of segmentation in Turkish, we fix the segmentation of the English vocabulary to word units and explore only target language segmentation strategies. Table 5 presents the results for both recurrent NMT and Transformer-based NMT. We see that the `SPM30K` model trained on Wikipedia performs consistently better than the others for both types of architectures. This result clearly shows the benefit of segmentation by using a large external corpus for the morphologically-rich Turkish language. We also note that for the MSVD-Turkish dataset, the recurrent CGRU architecture slightly outperforms the Transformer models in every setting.

6.1.2 Multimodal Machine Translation

We now fix the choice of segmentation to `words` for English and to `SPM30K` for Turkish, and proceed with the multimodal MT results. Here, we limit the experiments to recurrent MMT models since the monomodal results did not reveal any advantages for Transformer NMT in terms of performance (Table 5).

Table 6 shows the results for the multimodal MT experiments. We observe that none of the multimodal architectures can surpass[4] the strong recurrent baseline on average. This could be because of the multimodal fusion strategy employed here, which perhaps may be improved by a more sophisticated multimodal design.

[3] Pre-processing consists of lowercasing, length filtering with minimum token count set to 5, punctuation removal and deduplication.

[4] It should be noted we reused the hyper-parameters from the NMT experiments and did not conduct a hyper-parameter search for our MMT models.

Table 7 Quantitative comparison of the LSTM and Transformer-based Turkish video captioning models in terms of BLEU, METEOR, ROUGE-L and CIDEr metrics: word, BPE, SPM-based scores and the number of trainable parameters are reported, with **bold**-face denoting the best performance for each architecture.

	Vocab	BLEU	METEOR	ROUGE-L	CIDEr	# Params
LSTM	Word	**23.2** ±1.6	23.4 ±0.5	**55.3** ±1.3	**25.4** ±1.7	25.2M
	BPE30K	22.7 ±2.6	**24.7** ±1.4	53.8 ±1.8	**25.4** ±1.8	17.1M
	SPM30K	22.1 ±1.0	23.9 ±0.3	54.8 ±0.4	24.6 ±1.2	16.3M
Transformer	Word	**24.1** ±0.2	26.1 ±0.1	58.5 ±0.2	38.3 ±0.2	24.1M
	BPE30K	23.8 ±0.1	26.7 ±0.0	58.4 ±0.2	**40.0** ±0.6	21.1M
	SPM30K	23.9 ±0.0	**26.9** ±0.0	**59.1** ±0.1	38.2 ±0.1	20.7M

6.2 Video Captioning

In our quantitative analysis, we employ four commonly used evaluation metrics in captioning, namely BLEU, METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). The scores are computed with `coco-eval` toolkit (Lin et al., 2014). In our experiments, we train each model five times with different random seeds and report the average performances over all the runs and the corresponding standard deviations.

As stated earlier, we experiment with both LSTM and Transformer-based architectures and analyze three different segmentation strategies: word-, BPE- and SPM-level segmentations. As can be seen from the results presented in Table 7, Transformer-based models generate more accurate descriptions than the LSTM-based models. Moreover, switching from words to subword units extracted using SPM or BPE improves the performance in general except for the BLEU metric. Additionally, SPM- and BPE-based models have a smaller number of trainable parameters than word-based models. We note that the results in Table 7 are not directly comparable to the ones in Table 6 since for the MT experiments we consider a single reference, whereas for the video captioning experiments we consider all available references. This follows from previous work in these areas, where for MT every source-reference segment pair of is treated as an additional instance.

In Figure 8, we show some qualitative results of the proposed LSTM and Transformer-based Turkish captioning models and monomodal and multimodal recurrent MT models. In the top row, we show some sample results from our models where they give satisfactory translations and generations. In contrast, we provide some corner cases in the bottom row, where the proposed models produce semantically and/or grammatically incorrect outputs. This demonstrates that there are still some open challenges and room for further research.

Fig. 8 Video captioning and machine translation results on MSVD test set: TF (Transformer) and LSTM refer to the video captioning outputs, MMT system is the *enc-dec initialisation* variant from Table 6. The ground-truth captions, English *gloss* translations and incorrect generations are given in blue, gray and red, respectively. The examples (A) to (C) and (D) to (F) depict good and bad outputs, respectively.

7 Conclusion

In this paper we introduced and described a new large-scale video description dataset called MSVD-Turkish, which was constructed by carefully translating original English descriptions of MSVD dataset (Chen and Dolan, 2011) to Turkish. Our dataset will allow research on novel video captioning models for Turkish, a highly inflected and agglutinative language, as well as on

multilingual video captioning approaches, including those based on translation. Additionally, as our the Turkish descriptions are direct translations of English descriptions, the dataset can be used for research in novel approaches to multimodal MT.

We also provided baselines using popular neural models based on recurrent neural networks and Transformer architectures. For these neural architectures, we analysed the use of word segmentation approaches such as BPE and SPM and demonstrated how they help both in description generation as well as MT. We hope that our dataset will serve as a good resource for future efforts on multilingual, multimodal language generation. As an avenue of future work, it would be interesting to study the intrinsic annotation biases or linguistic differences between English and Turkish descriptions in the MSVD and MSVD-Turkish datasets.

Acknowledgements This work was supported in part by GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem, BAGEP 2021 Award of the Science Academy to A. Erdem, and the MMVC project funded by TUBITAK and the British Council via the Newton Fund Institutional Links grant programme (grant ID 219E054 and 352343575). Lucia Specia, Pranava Madhyastha and Ozan Caglayan also received support from MultiMT (H2020 ERC Starting Grant No. 678017).

References

Aafaq N, Mian A, Liu W, Gilani SZ, Shah M (2019) Video description: A survey of methods, datasets and evaluation metrics and description. ACM Comput Surv 52(6)

Akin AA, Akin MD (2007) Zemberek, an open source NLP framework for Turkic languages

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proc. International Conference on Learning Representations (ICLR), San Diego, California, USA, URL http://arxiv.org/pdf/1409.0473

Barbu A, Bridge A, Burchill Z, Coroian D, Dickinson S, Fidler S, Michaux A, Mussman S, Narayanaswamy S, Salvi D, Schmidt L, Shangguan J, Siskind JM, Waggoner J, Wang S, Wei J, Yin Y, Zhang Z (2012) Video in sentences out. In: Proc. 28th Conference on Uncertainty in Artificial Intelligence (UAI2012), Catalina Island, California, USA, URL http://arxiv.org/abs/1204.2742, 1204.2742

Barrault L, Bougares F, Specia L, Lala C, Elliott D, Frank S (2018) Findings of the third shared task on multimodal machine translation. In: Proc. Third Conference on Machine Translation, Volume 2: Shared Task Papers, Association for Computational Linguistics, Brussels, Belgium, pp 308–327, URL http://www.aclweb.org/anthology/W18-6402

Caglayan O (2019) Multimodal Machine Translation. Theses, Université du Maine, URL https://tel.archives-ouvertes.fr/tel-02309868

Caglayan O, Aransa W, Wang Y, Masana M, García-Martínez M, Bougares F, Barrault L, van de Weijer J (2016a) Does multimodality help human and machine for translation and image captioning? In: Proc. First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany, pp 627–633, URL http://www.aclweb.org/anthology/W/W16/W16-2358

Caglayan O, Barrault L, Bougares F (2016b) Multimodal attention for neural machine translation. Computing Research Repository arXiv:1609.03976, URL http://arxiv.org/pdf/1609.03976

Caglayan O, Aransa W, Bardet A, García-Martínez M, Bougares F, Barrault L, Masana M, Herranz L, van de Weijer J (2017a) LIUM-CVC submissions for WMT17 multimodal translation task. In: Proc. Second Conference on Machine Translation, Volume 2: Shared

Task Papers, Association for Computational Linguistics, Copenhagen, Denmark, pp 432–439, URL http://www.aclweb.org/anthology/W17-4746

Caglayan O, García-Martínez M, Bardet A, Aransa W, Bougares F, Barrault L (2017b) NMTPY: A flexible toolkit for advanced neural machine translation systems. Prague Bull Math Linguistics 109:15–28, DOI 10.1515/pralin-2017-0035, URL https://ufal.mff.cuni.cz/pbml/109/art-caglayan-et-al.pdf

Calixto I, Liu Q (2017) Incorporating global visual features into attention-based neural machine translation. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Copenhagen, Denmark, pp 992–1003, URL https://www.aclweb.org/anthology/D17-1105

Calixto I, Elliott D, Frank S (2016) DCU-UvA multimodal MT system report. In: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany, pp 634–638, URL http://www.aclweb.org/anthology/W/W16/W16-2359

Chen D, Dolan W (2011) Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 190–200, URL https://www.aclweb.org/anthology/P11-1020

Chen Y, Wang S, Zhang W, Huang Q (2018) Less is more: Picking informative frames for video captioning. In: Proc. European Conference on Computer Vision (ECCV), Munich, Germany, pp 367–384

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp 1724–1734, URL http://www.aclweb.org/anthology/D14-1179

Das P, Xu C, Doell R, Corso J (2013) A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, USA, pp 2634–2641

Denkowski M, Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. In: Proc. 9th Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, pp 376–380, DOI 10.3115/v1/W14-3348, URL http://aclweb.org/anthology/W14-3348

Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, USA, URL http://arxiv.org/abs/1411.4389

Elliott D, Kádár À (2017) Imagination improves multimodal translation. In: Proc. 8th International Joint Conference on Natural Language Processing (IJCNLP) (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, pp 130–141, URL http://aclweb.org/anthology/I17-1014

Elliott D, Frank S, Sima'an K, Specia L (2016) Multi30k: Multilingual english-german image descriptions. In: Proc. 5th Workshop on Vision and Language, Association for Computational Linguistics, Berlin, Germany, pp 70–74, URL http://anthology.aclweb.org/W16-3210

Elliott D, Frank S, Barrault L, Bougares F, Specia L (2017) Findings of the second shared task on multimodal machine translation and multilingual image description. In: Proc. 2nd Conference on Machine Translation, Volume 2: Shared Task Papers, Association for Computational Linguistics, Copenhagen, Denmark, pp 215–233, URL http://www.aclweb.org/anthology/W17-4718

Gella S, Lewis M, Rohrbach M (2018) A dataset for telling the stories of social media videos. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Brussels, Belgium, pp 968–974, DOI 10.18653/v1/D18-1117, URL https://www.aclweb.org/anthology/D18-1117

Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K (2013) Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proc. IEEE International Conference

on Computer Vision (ICCV), Sydney, Australia, pp 2712–2719, URL http://www.cs.utexas.edu/users/ai-labpub-view.php?PubID=127409

Hakeem A, Sheikh Y, Shah M (2004) $case^e$: A hierarchical event representation for the analysis of videos. In: Proc. Association for the Advancement of Artificial Intelligence (AAAI), San Jose, California, USA, pp 263–268

Hanckmann P, Schutte K, Burghouts GJ (2012) Automated textual descriptions for a wide range of video events with 48 human actions. In: Proc. European Conference on Computer Vision (ECCV), Firenze, Italy

He K, Xiangyu Z, Shaoqing R, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, pp 770–778, DOI 10.1109/CVPR.2016.90

Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. Neural computation 9(8):1735–1780

Huang PY, Liu F, Shiang SR, Oh J, Dyer C (2016) Attention-based multimodal neural machine translation. In: Proc. First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany, pp 639–645, URL http://www.aclweb.org/anthology/W/W16/W16-2360

Inan H, Khosravi K, Socher R (2016) Tying word vectors and word classifiers: A loss framework for language modeling. CoRR arXiv:1611.01462, URL http://arxiv.org/pdf/1611.01462

Kingma D, Ba J (2014) Adam: A method for stochastic optimization. CoRR arXiv:1412.6980, URL http://arxiv.org/pdf/1412.6980

Kojima A, Tamura T, Fukunaga K (2012) 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

Krippendorff K (1970) Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement 30(1):61–70, DOI 10.1177/001316447003000105

Krishna R, Hata K, Ren F, Li F, Niebles JC (2017) Dense-captioning events in videos. CoRR abs/1705.00754, URL http://arxiv.org/abs/1705.00754, 1705.00754

Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarrama S (2013) Generating natural-language video descriptions using text-mined knowledge. In: Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, USA, pp 10–19, URL https://www.aclweb.org/anthology/W13-1302

Kudo T, Richardson J (2018) Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, pp 66–71, URL http://www.aclweb.org/anthology/D18-2012

Li Y, Song Y, Cao L, Tetreault JR, Goldberg L, Jaimes A, Luo J (2016) TGIF: A new dataset and benchmark on animated GIF description. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, URL http://arxiv.org/abs/1604.02748, 1604.02748

Libovický J, Helcl J (2017) Attention strategies for multi-source sequence-to-sequence learning. In: Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, pp 196–202, DOI 10.18653/v1/P17-2031, URL http://aclweb.org/anthology/P17-2031

Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, pp 74–81

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common objects in context. In: Proc. European Conference on Computer Vision (ECCV), Springer, Zurich, Switzerland, pp 740–755

Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M (2019) fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In: Proc. Conference of the North

American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Minneapolis, Minnesota, USA

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: A method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, DOI 10.3115/1073083.1073135, URL http://dx.doi.org/10.3115/1073083.1073135

Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp 2641–2649, DOI 10.1109/ICCV.2015.303

Post M (2018) A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, pp 186–191, DOI 10.18653/v1/W18-6319, URL https://www.aclweb.org/anthology/W18-6319

Press O, Wolf L (2017) Using the output embedding to improve language models. In: Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, URL http://arxiv.org/abs/1608.05859, 1608.05859

Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf

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 1:25–36, URL https://www.aclweb.org/anthology/Q13-1003

Rohrbach A, Rohrbach M, Tandon N, Schiele B (2015) A dataset for movie description. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, USA, pp 3202–3212, DOI 10.1109/CVPR.2015.7298940

Rohrbach A, Torabi A, Rohrbach M, Tandon N, Pal C, Larochelle H, Courville A, Schiele B (2017) Movie description. Int J Comput Vision 123(1):94–120, DOI 10.1007/s11263-016-0987-1, URL https://doi.org/10.1007/s11263-016-0987-1

Rohrbach M, Amin S, Andriluka M, Schiele B (2012) A database for fine grained activity detection of cooking activities. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Providence, Rhode Island, USA, pp 1194–1201

Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B (2013) Translating video content to natural language descriptions. In: Proc. IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Sydney, Australia, pp 433–440

Senina A, Rohrbach M, Qiu W, Friedrich A, Amin S, Andriluka M, Pinkal M, Schiele B (2014) Coherent multi-sentence video description with variable level of detail. CoRR abs/1403.6173, URL http://arxiv.org/abs/1403.6173, 1403.6173

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, pp 1715–1725, URL http://www.aclweb.org/anthology/P16-1162

Sennrich R, Firat O, Cho K, Birch-Mayne A, Haddow B, Hitschler J, Junczys-Dowmunt M, Läubli S, Miceli Barone A, Mokry J, Nadejde M (2017) Nematus: a toolkit for neural machine translation. In: Proceedings of the EACL 2017 Software Demonstrations, Association for Computational Linguistics (ACL), Valencia, Spain, pp 65–68

Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A (2016) Hollywood in homes: Crowdsourcing data collection for activity understanding. CoRR abs/1604.01753, URL http://arxiv.org/abs/1604.01753, 1604.01753

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. URL http://arxiv.org/abs/1409.1556, cite arxiv:1409.1556

Specia L, Frank S, Sima'an K, Elliott D (2016) A shared task on multimodal machine translation and crosslingual image description. In: Proc. of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany, pp 543–553, URL http://www.aclweb.org/anthology/W/W16/W16-2346

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1):1929–1958, URL http://dl.acm.org/citation.cfm?id=2627435.2670313

Srivastava N, Mansimov E, Salakhutdinov R (2015) Unsupervised learning of video representations using lstms. CoRR abs/1502.04681, URL http://arxiv.org/abs/1502.04681, 1502.04681

Sulubacak U, Caglayan O, Grönroos SA, Rouhe A, Elliott D, Specia L, Tiedemann J (2019) Multimodal Machine Translation through Visuals and Speech. arXiv preprint arXiv:191112798

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proc. 27th International Conference on Neural Information Processing Systems (NeurIPS), MIT Press, Montreal, Canada, pp 3104–3112, URL http://dl.acm.org/citation.cfm?id=2969033.2969173

Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R (2014) Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proc. International Conference on Computational Linguistics (COLING): Technical Papers, Dublin, Ireland, pp 1218–1227, URL https://www.aclweb.org/anthology/C14-1115

Torabi A, Pal CJ, Larochelle H, Courville AC (2015) Using descriptive video services to create a large data source for video annotation research. CoRR abs/1503.01070, URL http://arxiv.org/abs/1503.01070, 1503.01070

Unal ME, Citamak B, Yagcioglu S, Erdem A, Erdem E, Cinbis NI, Cakici R (2016) TasvirEt: A benchmark dataset for automatic Turkish description generation from images. In: Proc. 24th Signal Processing and Communication Application Conference (SIU), pp 1977–1980, DOI 10.1109/SIU.2016.7496155

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp 5998–6008, URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez AN, Gouws S, Jones L, Kaiser L, Kalchbrenner N, Parmar N, Sepassi R, Shazeer N, Uszkoreit J (2018) Tensor2tensor for neural machine translation. CoRR abs/1803.07416, URL http://arxiv.org/abs/1803.07416

Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: Consensus-based image description evaluation. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, USA, pp 4566–4575

Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K (2015) Sequence to Sequence-Video to Text. In: Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile

Wang X, Chen W, Wu J, Wang Y, Wang WY (2018) Video captioning via hierarchical reinforcement learning. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, United States, URL http://arxiv.org/abs/1711.11135, 1711.11135

Wang X, Wu J, Chen J, Li L, Wang YF, Wang WY (2019) Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proc. IEEE International Conference on Computer Vision (ICCV), Seoul, Korea

Xu J, Mei T, Yao T, Rui Y (2016) MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Proc. 32nd International Conference on Machine Learning (ICML), JMLR Workshop and Conference Proceedings, Lille, France, pp 2048–2057, URL http://jmlr.org/proceedings/papers/v37/xuc15.pdf

Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. In: Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile

Yoshikawa Y, Shigeto Y, Takeuchi A (2017) STAIR captions: Constructing a large-scale japanese image caption dataset. In: Proc. 55th Annual Meeting of the Association

for Computational Linguistics, Vancouver, Canada, URL http://arxiv.org/abs/1705.00823, 1705.00823

Yu H, Wang J, Huang Z, Yang Y, Xu W (2016) Video paragraph captioning using hierarchical recurrent neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, URL http://arxiv.org/abs/1510.07712, 1510.07712

Zeng K, Chen T, Niebles JC, Sun M (2016) Title generation for user generated videos. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, URL http://arxiv.org/abs/1608.07068, 1608.07068

Zhou L, Xu C, Corso JJ (2017) Procnets: Learning to segment procedures in untrimmed and unconstrained videos. CoRR abs/1703.09788, URL http://arxiv.org/abs/1703.09788, 1703.09788

Zhou L, Kalantidis Y, Chen X, Corso JJ, Rohrbach M (2019) Grounded video description. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, California, USA, URL http://arxiv.org/abs/1812.06587, 1812.06587