Improving LSTM-based Video Description with Linguistic Knowledge Mined from Text

Subhashini Venugopalan UT Austin
Lisa Anne Hendricks UC Berkeley
Raymond Mooney UT Austin
Kate Saenko UMass Lowell

Abstract

This paper investigates how linguistic knowledge mined from large text corpora can aid the generation of natural language descriptions of videos. Specifically, we integrate both a neural language model and distributional semantics trained on large text corpora into a recent LSTM-based architecture for video description. We evaluate our approach on a collection of Youtube videos as well as two large movie description datasets showing significant improvements in grammaticality while maintaining or modestly improving descriptive quality. Further, we show that such techniques can be beneficial for describing unseen object classes with no paired training data (zeroshot captioning).

1 Introduction

The ability to automatically describe videos in natural language (NL) enables many important applications including: content-based video retrieval, video description for the visually impaired; and automated video surveillance. The past year has seen a marked increase in work on natural-language image description and a growing interest in video description. Recent works address these challenging tasks by combining the latest techniques in computer vision and natural language processing (NLP), and leveraging transformative advances in deep machine learning. In particular, recent recurrent neural network (RNN) methods (Vinyals et al., 2015; Donahue et al., 2015; Venugopalan et al., 2015b) for image and video description treat the problem as a machine translation task, translating a static image or a sequence of visual inputs (as in video) to natural language text and demonstrate very promising results.

A significant factor contributing to the success of neural network architectures for image description is the availability of large amounts of paired image-sentence corpora. In the case of videos however, there is a lack of high-quality paired video-sentence corpora. In contrast, monolingual text corpora are widely available. Despite the lack of visual grounding, plain text corpora exhibit rich linguistic structure that can aid video to text translation. Most work in statistical machine translation utilizes both a language model trained on a large corpus of monolingual data for the target language as well as a translation model trained on more limited parallel bilingual data. This paper explores methods to incorporate knowledge from language corpora to capture general linguistic regularities to aid video description.

In this work, we investigate several techniques to integrate linguistic information into an RNN-based video description system. We use RNNs based on Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) which have shown state-of-the-art performance on image captioning (Vinyals et al., 2015) and video description (Yao et al., 2015; Venugopalan et al., 2015a). Additionally, LSTMs have also shown to be very effective language models (LM) (Sundermeyer et al., 2010). Our first approach (early fusion) is to pre-train the network on text only corpora before training on parallel video-text corpora. Our next two approaches, based on Gulcehre et al. (2015), aim to integrate a trained LSTM language model with the existing video-to-text model. Further, we also explore replacing the one-hot word embedding, used in several image and video captioning works, with distributional word vectors trained on external corpora.

We present detailed comparisons between the approaches by evaluating their performance on a standard Youtube corpus and two more recent

and larger movie description datasets. The results demonstrate significant improvements in grammaticality of the descriptions (as determined by crowdsourced human evaluations) and more modest improvements in descriptive quality (as determined by both crowdsourced human judgements and standard automated comparison to human-generated descriptions).

While adding knowledge from text corpora can help correct visual interpretation and linguistic irregularities in captioning, we also demonstrate the applicability of our methods to the more compelling application of describing novel objects in images.

2 Related Work

Video Description. There is a small but growing body of work on generating NL video descriptions. Many of the earlier methods (Yu and Siskind, 2013; Rohrbach et al., 2013; Krishnamoorthy et al., 2013; Guadarrama et al., 2013; Thomason et al., 2014) followed a two-step approach where they first identify the semantic content (e.g. subject, verb, object) and then use CRFs, sentence templates, or SMT based approaches to generate a sentence from the content words (surface realization). Last year, following the Machine Translation approach by Sutskever et al. (2014), Venugopalan et al. (2015b), Yao et al. (2015) proposed CNN-RNN based methods to generate a vector representation for the video and "decode" it using an LSTM sequence model to generate a description.

Adding Linguistic Knowledge. In this work, we investigate the use of distributional semantic embeddings and LSTM-based language models trained on external text corpora to aid existing video description models. Sundermeyer et al. (2010) showed LSTMs to be very effective language models. Additionally, Gulcehre et al. (2015) developed an LSTM model for machine translation that incorporates a monolingual language model for the target language showing improved results. We utilize a similar approach to train an LSTM for translating video to text that exploits large monolingual-English corpora (Wikipedia, BNC, UkWac) to improve RNN based video description networks. Contemporaneous to us Yu et al. (2015), Pan et al. (2015) and Ballas et al. (2015) propose different neural models for video description. However their focus is primarily on improving the video representation itself. Rohrbach et al. (2015a) also incorporate scene classifiers to add to the visual inputs. In this work we choose S2VT to demonstrate our approaches, however they can also be applied to the other recent RNN-based video description networks.

Describing Novel Objects. While the idea of identifying concepts or class-labels for unseen (zero-shot) and rarely seen (few-shot) objects has been studied extensively (Rohrbach et al., 2010; Parikh and Grauman, 2011; Frome et al., 2013; Lampert et al., 2014; Akata et al., 2013) in the context of computer vision and natural language understanding, the concept of generating descriptions for such novel or rarely seen objects is quite unexplored. In recent work, Mao et al. (2015) propose a neural caption model to learn and describe objects based on few sentence examples, and Hendricks et al. (2016) propose Deep Compositional Captioner (DCC) to generate descriptions for objects not seen in paired training data, by learning to transfer knowledge from seen objects. In this work, we show that our methods can be applied to the DCC model to simplify their transfer technique and generate captions for unseen classes of objects in images.

3 LSTM-based Video Description

Here, we briefly describe the underlying video description framework, S2VT from Venugopalan et al. (2015a) used in this work. The S2VT model uses an encoder-decoder approach (Sutskever et al., 2014; Cho et al., 2014) which reads a sequence of inputs $\vec{x} = (x_1, ..., x_T)$ (video frame features), generates a fixed dimensional vector representation of the sequence, and then decodes this latent representation into a sequence of output words $\vec{y} = (y_1, ..., y_N)$. In S2VT, the encoder and decoder are both modeled using LSTM recurrent neural networks.

3.1 Long Short Term Memory

LSTMs incorporate explicitly controllable memory units that allow it to learn long-range temporal dependencies, which are very difficult to learn using traditional recurrent networks. We refer interested readers to the original works for detailed explanation, here we only present the functions in the recurrence (Eqn 1) that allow the LSTM to encode a sequence of inputs $\vec{x}$ to a vector and decode from it to generate the output sequence $\vec{y}$.

Let x_t, c_t, and h_t denote the input, cell memory, and hidden control states at each time step. Given a sequence of inputs $(x_1, \dots, x_T)$ the LSTM computes the cell memory sequences $(c_1, \dots, c_T)$ and hidden control sequences $(h_1, \dots, h_T)$ as follows:

$$\begin{aligned} i_t &= \text{sigm}(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \text{sigm}(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \text{sigm}(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad (1)$$

where $\odot$ represents the element-wise product, and the weight matrices denoted by W_{ij} and biases b_j are the trained parameters. When LSTMs are stacked, the $\vec{h}$-sequence of one layer is provided as input ($\vec{x}$) to the following layer. During decoding, the model essentially defines a probability over the output sequence $\vec{y}$ by decomposing the joint probability into ordered conditionals:

$$p(\vec{y}|x_1, \dots, x_T) = \prod_{t=1}^{N} p(y_t|h_T, y_1, \dots, y_{t-1})$$

This is done by applying a softmax function on the decoder LSTM's h-sequence. Hence, for a word in the vocabulary ($w \in V$),

$$p(y_t = w|h_T, \vec{y}_{<t}) = \text{softmax}(W_v h_T + b_v) \quad (2)$$

Thus the overall objective of the network is to maximize log-likelihood of the output word sequence.

$$\log p(\vec{y}|\vec{x}) = \sum_{t=1}^{N} \log p(y_t|h_T, \vec{y}_{<t}) \quad (3)$$

3.2 S2VT Model

The S2VT network (Fig. 1) employs a stack of two LSTM layers. The input $\vec{x}$ to the first LSTM layer is a sequence of frame features obtained from the penultimate layer (fc_7) of a Convolutional Neural Network (CNN) after the ReLu operation. The first layer of LSTMs process these frame-feature inputs to encode the video sequence. At each time step, the hidden control state h_t is provided as input to a second LSTM layer. After viewing all the frames, the second LSTM layer learns to decode from this state to output a sequence of words. We can interpret this architecture as one LSTM layer modeling the visual features, and a second LSTM layer modeling language conditioned on the visual representation. Our work presents modifications to this architecture to incorporate linguistic knowledge at different stages of the training and generation process. Although our methods are evaluated primarily on the S2VT architecture, they are sufficiently general and could be incorporated into other CNN-RNN based captioning models.

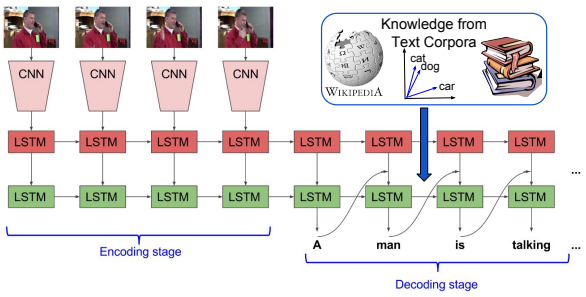


Figure 1: The S2VT architecture that encodes a sequence of frames and decodes them to a sentence. We propose to add knowledge from text corpora to enhance the quality of video description.

4 Approach

Existing video and image captioning models are trained solely on text from the caption datasets and tend to exhibit some linguistic irregularities associated with a restricted language model and a small vocabulary. Here, we investigate several techniques to integrate prior linguistic knowledge into a CNN/LSTM-based network for video to text (S2VT) and evaluate their effectiveness at improving the overall description. We also show that such techniques are particularly useful for the task of generating descriptions of novel/previously-unseen objects.

Our first approach (*early fusion*) is to pre-train portions of the network modeling language on large corpora of raw NL text and then continue "fine-tuning" the parameters on the paired video-text corpus. Our next two approaches, *late fusion* and *deep fusion* (Figure 2), based on Gulcehre et al. (2015), aim to integrate a separate LSTM-based language model (trained on external text corpora) to improve the predictions of the video description model. Finally, we employ pretrained distributional word embeddings to replace the one-hot vector encoding of words commonly used in several image and video captioning works.

Early Fusion. In the early fusion approach, we consider just the language LSTM (second) layer of the S2VT model and pre-train a language model using web-scale text corpora. An LSTM model learns to estimate the probability of an output sequence given an input sequence. To learn a language model, we train the LSTM layer to predict the next word given the previous words. Fol-

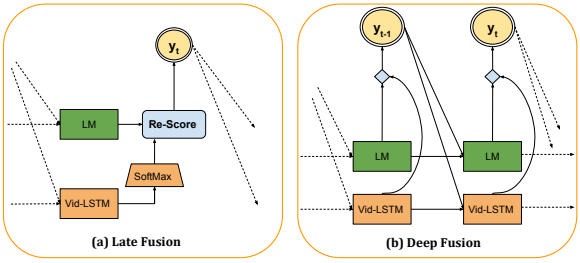


Figure 2: Illustration of our late and deep fusion approaches to integrating an independently trained LM to aid video captioning. The deep fusion model learns jointly from the hidden representations of the LM and S2VT video-to-text model (Vid-LSTM), whereas the late fusion re-scores the softmax output of the video-to-text model.

lowing the S2VT architecture, we embed one-hot encoded words in lower dimensional vectors. This lexical representation is then provided as input to the LSTM layer. The network is trained on external text and the parameters are learned through backpropagation using stochastic gradient descent.[1] The weights from this network are then used to *initialize* the embedding and weights of the LSTM layers of S2VT, which is then trained on video-text data. This trained LM is also used as the LSTM LM in the late and deep fusion models.

Late Fusion. Our late fusion approach is similar to how neural machine translation models incorporate a trained language model during decoding. At each step of sentence generation, the video caption model proposes a distribution over the vocabulary words. We then use the language model to re-score the final output by considering the weighted average of the sum of scores proposed by the LM as well as the S2VT video-description model (VM). More specifically, if y_t denotes the output at time step t, and if p_{VM} and p_{LM} denote the proposal distributions of the video captioning model, and the language models respectively, then for all words $y' \in V$ in the vocabulary we can recompute the score of each new word, $p(y_t = y')$ as:

$$\alpha \cdot p_{VM}(y_t = y') + (1 - \alpha) \cdot p_{LM}(y_t = y') \quad (4)$$

Hyper-parameter α is tuned on the validation set.

Deep Fusion. In the deep fusion approach, we integrate the LM a step deeper in the generation process by concatenating the hidden state of the language model LSTM (h_t^{LM}) with the hidden state of the S2VT video description model (h_t^{VM}) and use the combined latent vector to predict the output word. This is similar to the technique proposed by Gulcehre et al. (2015) for incorporating language models trained on monolingual corpora for machine translation. However, our approach differs in two key ways: (1) we only concatenate the hidden states of the S2VT LSTM and language LSTM and do not use any additional context information, (2) we fix the weights of the LSTM language model but train the full video captioning model with the additional LM model input. In this case, the probability of the predicted word at time step t is:

$$p(y_t|\vec{y}_{<t}, \vec{x}) \propto \exp(\mathrm{Wf}(h_t^{VM}, h_t^{LM}) + b) \quad (5)$$

where $\vec{x}$ is the visual feature input, W is the weight matrix, and b the biases. We avoid tuning the LSTM LM to prevent overwriting already learned weights of a strong language model. But we train the full video caption model to incorporate the LM outputs while training on the caption domain.

Distributional Word Representations. The S2VT network, like most image and video captioning models, represents words using a 1-of-N (one hot) encoding. During training, it learns to embed "one-hot" words into a lower 500 dimensional space by applying a linear transformation, learning its parameters via back-propagation. However, the embedding is learned only from the text in the limited and possibly noisy text in the caption data. There are many approaches (Mikolov et al., 2013; Pennington et al., 2014) that use large text corpora to learn vector-space representations of words that capture fine-grained semantic and syntactic regularities. It is only natural to take advantage of these to aid video description. Specifically, we used 300-dimensional GloVe vectors (Pennington et al., 2014) pre-trained on 6B tokens from Gigaword and Wikipedia 2014. While S2VT mapped the one-hot vectors to 500d, we modify this to initialize an embedding based on the GloVe vectors. In addition to using the distributional vectors for the input, we also explore variations where the model predicts both the one-hot word (trained on the $\mathrm{softmax}$ loss, Eqn 2), as well as predicting the distributional vector from the LSTM hidden state using Euclidean loss as the objective. Here the output is computed as $y_t = (W_g h_t + b_g)$, and the loss is given by:

$$\mathbb{L}(y_t, w_{glove}) = \|(W_g h_t + b_g) - w_{glove}\|^2 \quad (6)$$

where h_t is the output of the LSTM and w_{glove} is the GloVe embedding of the word. The network

[1] The LM was trained to achieve a perplexity of 120

then essentially becomes a multi-task model with two loss functions. However, we use this loss only to influence the weights learned by the network, the predicted word embedding is not used.

Ensembling. The overall loss function (Equation 3) of the video captioning network is non-convex, and difficult to optimize. In practice, using an ensemble of networks trained slightly differently can improve performance (Hansen and Salamon, 1990). Hence we also present results over an ensemble of the models by averaging their predictions.

5 Datasets

We first describe the external text corpora and video description datasets used in our experiments, and then in the next section we present the evaluation results.

External Text Corpus (WebCorpus). We extract sentences from Gigaword, the British National Corpus (BNC), UkWaC, and Wikipedia. Stanford CoreNLP 3.4.2 (Manning et al., 2014) was used to extract tokenizations. This dataset was used to train the LSTM language model. The model's vocabulary consisted of the 80,000 most frequent tokens from the combined corpus. For experiments with the distributional embeddings, we refined this vocabulary further to a set of 72,700 words that also had pre-trained GloVe embeddings.

In-Domain Corpus. For experiments on the Youtube video description dataset, we also compare our approaches by training on sentences from "in-domain" corpora of visual descriptions. Specifically, we choose training sentences from the image caption corpus MSCOCO (Lin et al., 2014).

5.1 Video Description datasets

We report results on three video annotation datasets: a collection of Youtube videos and two movie description datasets.

Youtube video dataset The Microsoft Video Description corpus (Chen and Dolan, 2011) consists of 1,970 short, single-activity, 10-25s length clips from Youtube. Each clip has about 40 English sentences describing the main event in the video. Following the experimental settings used in prior work (Venugopalan et al., 2015b) we use 1,200 videos for training, 100 for validation and 670 for test.

MPII Movie Description Dataset (MPII-MD) MPII-MD (Rohrbach et al., 2015b) contains around 68,000 video clips extracted from 94 Hollywood movies. A single sentence description accompanies each clip, sourced from movie scripts and audio description (AD) data. The AD or Descriptive Video Service (DVS) is an additional audio track for the visually impaired describing the visual elements that occur on screen.

Montreal Video Annotation Dataset (M-VAD) The M-VAD movie description corpus (Torabi et al., 2015) is similar to MPII-MD in that it contains AD data from 92 movies. The dataset consists of about 49,000 short video clips with mostly single-sentence descriptions.

6 Experiments

Automatic Evaluation Metrics. We evaluate performance using machine translation (MT) metrics METEOR (Denkowski and Lavie, 2014) and BLEU (Papineni et al., 2002) to compare the machine-generated descriptions to human ones. The metrics score the generated description based on alignment and similarity to the set of candidate reference sentences. For the movie corpora that have just a single description we use only the METEOR metric which is more robust when the number of references are small (Vedantam et al., 2015). We use the code accompanying the Microsoft COCO Evaluation Server.

Human Evaluation. We also collected human judgements of the test sentences from some of the best performing models using Amazon Mechanical Turk. We evaluated sentences from a random subset of 200 video clips from each dataset. For the Youtube dataset, we obtained judgements for relevance and grammar, and for the movie corpora we only evaluated for grammar (for video copyright reasons). For relevance, we asked workers to watch a video, and rate sentences generated by the different models on a Likert scale of 1 to 5 based on how relevant the sentences were to the video, where 5 meant the sentence was "strongly relevant" and 1 meant the sentence was "irrelevant". For grammar, we do not provide the video, but asked workers to rate the description on a 5 point scale based on sentence formation, vocabulary, and grammar.

6.1 Youtube Video Dataset Results

We compare the performance of our proposed techniques to the base S2VT model in Table 1. The fusion techniques show modest improvements in METEOR scores, with Deep Fusion performing best on both METEOR and BLEU. Incorporating Glove embeddings in place of one-hot encoding shows a more substantial increase in METEOR, compared to any of the fusion techniques. The single best model incorporated Glove embeddings on the input as well as deep fusion with a language model trained on external corpora. Our final model that performed best on this dataset is an ensemble of the S2VT model with Glove, and the Glove+Deep Fusion models trained on both the Web Corpus and In Domain COCO sentences. We note here that the state-of-the-art on this dataset is achieved by HRNE (Pan et al., 2015) (METEOR 33.1) and by h-RNN-Vgg (Yu et al., 2015) (BLEU 44.3) both of which propose superior visual processing pipelines using attention to encode the video.

Embedding Influence. In our experiments, we compared multiple methods of incorporating word embeddings: **(1) GloVe for the input.** Using GloVe embeddings as the input to the LSTM performed best on METEOR, and all the scores reported in Tables 1 and 3 correspond to this setting. **(2) Fine-tuning Embedding.** We also experimented with initializing the input embedding with pre-trained vectors and subsequently fine-tuning them via backpropagation. This actually reduced performance on our validation set by 0.4 METEOR. **(3) Input and Predict Embedding.** As described in Section 4, we trained a model to take word embeddings as input the and predict both the one-hot outputs and the word embedding incorporating two loss functions. This model did not perform better than using GloVe vectors as input.

Human Evaluation. We obtained human judgements on relevance and grammar for the baseline S2VT model and the best performing models; the mean ratings are shown in Table 2, including evaluation of the human-generated "ground truth" descriptions. Human relevance ratings appear to correlate well with the METEOR scores, confirming that our methods give a modest improvement in descriptive quality. However, our methods for incorporating linguistic knowledge significantly improve the grammaticality of the results, making them more comprehensible to human users. Significance was determined using the Wilcoxson Sign Rank Test, and the improvements were significant with $p < 0.02$ on relevance and $p < 0.001$ on grammar.

Model	METEOR	B-4
S2VT	29.2	37.0
Early Fusion	29.6	37.6
Late Fusion	29.4	37.2
Deep Fusion	29.6	39.3
Glove	30.0	37.0
Glove + Deep Fusion		
- Web Corpus	30.3	38.1
- In-Domain	30.3	38.8
Ensemble	**31.4**	**42.1**
h-RNN-Vgg	31.1	44.3
HRNE	32.1	43.6
HRNE (+attn)	33.1	43.8

Table 1: Youtube dataset: Video description evaluation using MT metrics, B-4 denotes BLEU@4, and METEOR in %. Our model's best results are in bold.

Model	Relevance	Grammar
S2VT	2.06	3.76
Glove+Deep (Web Corpus)	2.12	4.05*
Glove+Deep (In Domain)	2.21*	4.17*
Ensemble	**2.24***	**4.20***
GroundTruth	4.52	4.47

Table 2: Youtube dataset: Mean scores on Human Evaluation. Sentences were rated on a scale of 1 to 5 for relevance and grammatical correctness. Higher values are better. Our model's best results are in bold, * indicates significant improvement.

6.2 Movie Description Results

Results on the movie corpora are presented in Table 3. Both MPII-MD and M-VAD videos have only a single ground truth description for each video, which makes both learning and evaluation very challenging. METEOR scores are also significantly lower since generated sentences are compared to a single reference translation. We compare all results against a re-implementation of the baseline S2VT model using the same vocabulary and architecture. We observe that the ability of external linguistic knowledge to improve METEOR scores on these challenging datasets is small but consistent. Again, human evaluations of grammaticality show that incorporating linguistic knowledge significantly improves the quality of sentence generation. Examples in Fig. 4 demonstrate the superior sentence quality produced by our models, especially in terms of grammaticality. For example, the description in Fig. 4 (top) changes

Dataset	Metric	S2VT*	Early Fusion	Late Fusion	Deep Fusion	Glove	Glove+Deep Fusion
MPII-MD	METEOR	6.5	6.7	6.5	6.8	6.7	6.8
	Grammar	2.6	-	-	-	3.9	**4.1**
M-VAD	METEOR	6.6	6.8	6.7	6.8	6.7	6.8
	Grammar	2.2	-	-	-	3.1	**3.3**

Table 3: Movie Description results comparing our approaches. S2VT* is a reimplementation using the same underlying architecture and vocabulary as the fusion models. Human judgements were obtained to evaluate grammatical correctness of the sentences on a scale of 1-5 (higher being better). The most significant improvement is indicated in bold.

from "Someone looks at the car, someone looks at the car" (S2VT) to "Someone is driving down the street, looking around the road" (Glove+Deep).

7 Compositional captioning

In addition to improving standard video description tasks, incorporating linguistic knowledge can potentially benefit other scenarios in which limited paired training data is available. In this section, we demonstrate how incorporating linguistic knowledge simplifies training when learning to describe novel objects unseen in paired training data.

Task. We consider the problem of generating a sentence description for objects that have not been seen in the image caption training data. This problem is interesting because current visual recognition systems are capable of recognizing several thousand object categories, but state-of-the-art captioning models lack the ability to describe these objects without explicit training on image-sentence pairs containing the objects.

Recent work has explored caption generation for novel objects with limited paired training data (Mao et al. (2015)) or no paired training data (Hendricks et al. (2016)). We focus on the latter scenario and demonstrate that adding linguistic knowledge, specifically word embeddings trained on large text corpora, can simplify training the Deep Compositional Captioner (DCC) model proposed in Hendricks et al. (2016).

7.1 DCC Caption Model

We briefly highlight the main aspects of the DCC model shown in Figure 3. DCC consists of three main modules, 1) a pre-trained visual classifier (yellow), 2) an LSTM language model (blue), and 3) a multi-modal unit (orange) that learns to combine predictions from the visual classifier and the LM to generate a caption. DCC is advantageous because the visual classifier and LM can be trained independently on object recognition and monolingual text data. To generate sentences conditioned on image content, the visual classifer and LM model are combined by training a multi-modal unit on paired image-caption data. To generate sentences which incorporate objects unseen in paired training data but seen in unpaired image and text data, learned parameters in the multi-modal unit (W_L and W_I in Fig 3) are transferred from seen object classes to novel object classes. Though the transfer mechanism produces good sentences about a variety of objects, DCC cannot be trained end-to-end and, as discussed in Hendricks et al. (2016), sentence quality is dependent on determining good weights to transfer to unseen categories.

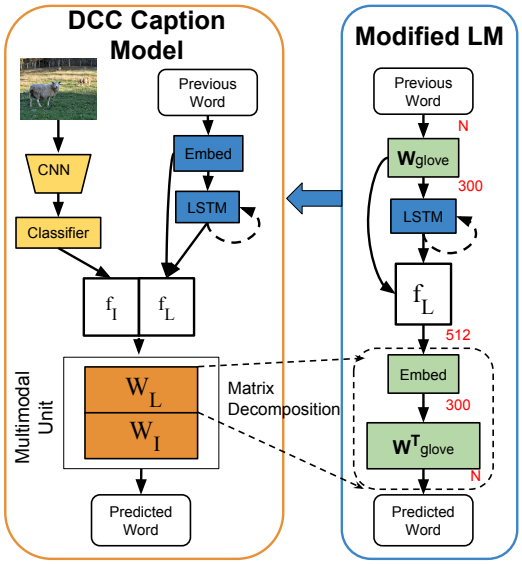


Figure 3: Left: DCC image captioning model. Right: Our modifications to the LM to incorporate distributional word embeddings in the input and output. This allows the model to generate captions of unseen classes without any explicit transfer from the language unit.

7.2 Proposed Modified LM

We propose a modified language model (Fig 3, right) based on models outlined in Section 4 that can generate captions for novel object classes, without employing the "transfer" mechanism for

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg.
DCC	2.4	22.5	27.9	25.0	**65.1**	**54.8**	0	76.7	**34.3**
Ours (no LM transfer)	0	**39.3**	27.0	23.0	4.7	0	0	76.6	21.3

Table 4: Compositional Captioning: F1 scores of our embedding approach on the compositional captioning task on held-out objects not seen during training. Our approach uses distributional vector representation in the language and is able to caption objects without employing the Language Model "Transfer" mechanism used by DCC. We note that F1 scores for the DCC model without Language transfer is 0.

language weights. Given a new object class, the transfer mechanism essentially identifies a seen object class that is "similar" to the novel class and then transfers (copies) its parameters. To circumvent the transfer of multi-modal language weights (W_L), we use an embedding space that naturally encapsulates this concept, such as pre-trained word2vec or GloVe vectors, where semantically similar objects have similar weights. The crucial change we make to the model in order to achieve this is to linearly decompose the weight matrix W_L into two matrices such that, $W_L = W_{glove}^T W_{embed}$. We fix the GloVe weights throughout training to ensure that its properties are retained in the final model. Tuning weights, even when the input and output weights are shared, changes parameters significantly and does not generate sentences about novel classes.

7.3 Evaluation and Results

To evaluate the effect of incorporating linguistic knowledge, we follow the protocol outlined in Hendricks et al. (2016) to (pre-)train our visual classifiers and language model. We train the full captioning model on a subset of the MSCOCO dataset with 8 objects removed (refer to Table 4) and evaluate our model by generating sentences for images which include at least one of the 8 held-out "unseen" object classes. We compute the F1 score to evaluate the model's ability to correctly incorporate novel objects into captions. We consider generated sentences "positive" if they contain at least one mention of a held out word, and ground truth sentences "positive" if a word is mentioned in any ground truth annotation that describes an image. Our results (Table 4) show that our model performs comparably to the DCC model on half of the classes without explicit language transfer. We note that without language transfer, the F1 scores for the original DCC model is 0 for all objects. We believe incorporating linguistic knowledge through word embeddings is a promising approach to train models in which there is limited paired training examples in an end-to-end fashion.

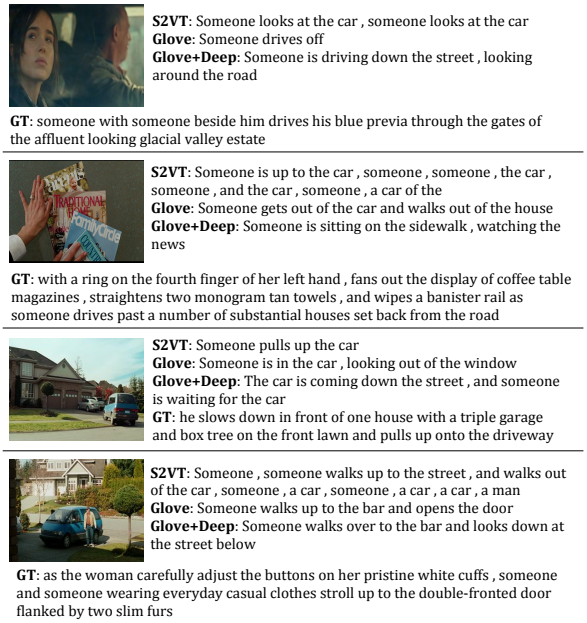


Figure 4: MPII-MD Example: Representative frames from contiguous clips from the movie "Juno". S2VT is the baseline model, Glove indicates the model trained with input Glove vectors, and Glove+Deep uses input Glove vectors with the Deep Fusion approach. GT indicates groundtruth sentence.

8 Conclusion

This paper investigates multiple techniques to incorporate linguistic knowledge from text corpora to aid in description generation when limited training data is available, such as in video description or in zero-shot captioning. For video description, we empirically evaluate our approaches on Youtube clips and two large and challenging movie description datasets, MPII-MD and MVAD. Our results show significant improvements on human evaluations of grammar while also maintaining or improving the overall descriptive quality of sentences on all three datasets. Further, we show that such techniques can benefit models generating captions for novel object classes. While methods presented in this work are applied to specific video or image captioning networks, the methods are sufficiently generic to be applied to other video and image description models.

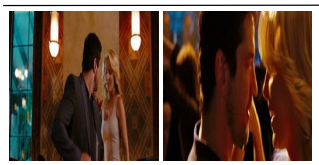
Correct		**S2VT**: The sun is is over the water of the water. **Glove**: The unk mountains of the river, which is filled with a large sea. **Glove+Deep**: The hogwarts express chugs through the barren moorland. **GT**: Steam billows from the funnel as the hogwarts express travels through the rain beside the edge of a vast lake.
Related (but doesn't match GroundTruth)		**S2VT**: Someone takes a back, someone and someone to the door. **Glove**: Someone sits on the couch and watches her phone. **Glove+Deep**:Someone sits on the couch, watching her, her feet on her lap. **GT**: Someone drops the flowers and kisses someone.
		**S2VT**: Someone walks up. **Glove**: Someone looks at someone , then turns to someone. **Glove+Deep**: Someone looks at someone , who is still standing in the doorway , watching the tv. **GT**: Someone thrusts a wet umbrella at someone.
		**S2VT**: Someone walks to the door, someone walks up to the door. **Glove**: Someone walks into the kitchen and sits down. **Glove+Deep**: Someone walks over to the window and looks out. **GT**: Someone is still eating and watching television.
		**S2VT**: Someone , the man and the man, the water, the water of the ground. **Glove**: Someone is sitting on the ground, his head bowed. **Glove+Deep**: Someone is walking along the sidewalk, a tall camel, a man in a ferret, a bloodhound drooling. **GT**: A magnificent creature stands in front of them.
		**S2VT**: Someone takes a head. The man on a door. **Glove**: Someone unk her gaze. Someone and someone dance. **Glove+Deep**: Someone and someone watch the dance floor. Someone and someone dance. **GT**: He leads her to the dance floor and flings off his jacket. He raises her arms above her head.
		**S2VT**: Someone and someone to the door, the man of the man of the floor. **Glove**: Someone pulls out a car. Someone glances at the wheel, then turns to the side of the road. **Glove+Deep**: Someone pulls out a pair of doors and slides out of the car. He pulls out a pistol. **GT**: Drawing his gun, someone returns fire. Someone cowers . The pick-up swerves onto the one-way street and jams itself alongside the delta, mangling the convertibles headlight and someone. The vehicles separate. Someone bashes the pick-up.
Incorrect		**S2VT**: Someone, someone walks into the window. **Glove**: Someone is in the back of the car. **Glove+Deep**: Someone grabs the phone and punches it at someone. **GT**: Someone grabs the tablecloth.
		**S2VT**: Someone takes a hand and someone and someone to the door. **Glove**: Someone turns to the door and finds someone who sits on a bench. The penguins steps closer. **Glove+Deep**: Someone and someone step into the elevator. Someone glances at the security guard. Someone stands and walks off. **GT**: A customer opens the door for the arriving painters. As they wheel in the hand truck, at street level, the two guards watch the upper lobby from the top of the stairs. Two painters fasten the doors shut.

Figure 5: Representative frames from clips in the movie description corpora. S2VT is the baseline model, Glove indicates the model trained with input Glove vectors, and Glove+Deep uses input Glove vectors with the Deep Fusion approach. GT indicates groundtruth sentence.

Acknowledgements

This work was supported by NSF awards IIS-1427425 and IIS-1212798, and ONR ATL Grant N00014-11-1-010, and DARPA under AFRL grant FA8750-13-2-0026. Raymond Mooney and Kate Saenko also acknowledge support from a Google grant. Lisa Anne Hendricks is supported by the National Defense Science and Engineering Graduate (NDSEG) Fellowship.

References

[Akata et al.2013] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826.

[Ballas et al.2015] Nicolas Ballas, Li Yao, Chris Pal, and Aaron C. Courville. 2015. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432.

[Chen and Dolan2011] David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

[Cho et al.2014] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv:1409.1259*.

[Denkowski and Lavie2014] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *EACL*.

[Donahue et al.2015] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

[Frome et al.2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129.

[Guadarrama et al.2013] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *ICCV*.

[Gulcehre et al.2015] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.C. Lin, F. Bougares, H. Schwenk, and Y. Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

[Hansen and Salamon1990] L. K. Hansen and P. Salamon. 1990. Neural network ensembles. *IEEE TPAMI*, 12(10):993–1001, Oct.

[Hendricks et al.2016] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8).

[Krishnamoorthy et al.2013] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J. Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, July.

[Lampert et al.2014] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465.

[Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

[Manning et al.2014] C. Manning, M Surdeanu, J Bauer, J Finkel, S J Bethard, and Dx McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

[Mao et al.2015] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Pan et al.2015] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2015. Hierarchical recurrent neural encoder for video representation with application to captioning. *CoRR*, abs/1511.03476.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

[Parikh and Grauman2011] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

[Rohrbach et al.2010] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. 2010. What helps where–and why? semantic relatedness for knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 910–917. IEEE.

[Rohrbach et al.2013] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *ICCV*.

[Rohrbach et al.2015a] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. 2015a. The long-short story of movie description. *GCPR*.

[Rohrbach et al.2015b] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015b. A dataset for movie description. In *CVPR*.

[Sundermeyer et al.2010] M. Sundermeyer, R. Schluter, and H. Ney. 2010. Lstm neural networks for language modeling. In *INTERSPEECH*.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

[Thomason et al.2014] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J. Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*.

[Torabi et al.2015] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070v1*.

[Vedantam et al.2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *CVPR*.

[Venugopalan et al.2015a] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. 2015a. Sequence to sequence - video to text. *ICCV*.

[Venugopalan et al.2015b] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015b. Translating videos to natural language using deep recurrent neural networks. In *NAACL*.

[Vinyals et al.2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *CVPR*.

[Yao et al.2015] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. *arXiv:1502.08029v4*.

[Yu and Siskind2013] Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from videos described with sentences. In *ACL*.

[Yu et al.2015] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2015. Video paragraph captioning using hierarchical recurrent neural networks. *CoRR*, abs/1510.07712.