Multimodal Video Description

Vasili Ramanishka
UMass Lowell
vramanis@cs.uml.edu

Abir Das
UMass Lowell
adas@cs.uml.edu

Dong Huk Park
UC Berkeley
dong.huk.park@berkeley.edu

Subhashini Venugopalan
University of Texas Austin
vsubhashini@utexas.edu

Lisa Anne Hendricks
UC Berkeley
lisa_anne@berkeley.edu

Marcus Rohrbach
UC Berkeley
rohrbach@berkeley.edu

Kate Saenko
UMass Lowell
saenko@cs.uml.edu

ABSTRACT

Real-world web videos often contain cues to supplement visual information for generating natural language descriptions. In this paper we propose a sequence-to-sequence model which explores such auxiliary information. In particular, audio and the topic of the video are used in addition to the visual information in a multimodal framework to generate coherent descriptions of videos "in the wild". In contrast to current encoder-decoder based models which exploit visual information only during the encoding stage, our model fuses multiple sources of information judiciously, showing improvement over using the different modalities separately. We based our multimodal video description network on the state-of-the-art sequence to sequence video to text (S2VT) [26] model and extended it to take advantage of multiple modalities. Extensive experiments on the challenging MSR-VTT dataset are carried out to show the superior performance of the proposed approach on natural videos found in the web.

1. INTRODUCTION

Understanding a visual scene and expressing it in terms of natural language descriptions has drawn considerable interest from both computer vision and natural language processing communities. Early works on visual description have mostly focused on describing still images [6, 14, 15, 28]. Early efforts to generate automated video descriptions were based on a two stage pipeline which first identifies the semantic visual concepts and then stitches them in a "subject, verb, object" template [5, 8, 13, 22]. Though a template based approach separates the concept identification and description generation tasks, such templates are insufficient in modeling the richness of the language as generally found in human generated descriptions of videos or scenes.

Our model is based on the S2VT [26] model for generating natural language descriptions from videos. S2VT is an encoder-decoder based framework which maps a sequence of frames to a sequence of words. For an input sequence of video frames, the encoder first converts the video frames into a sequence of high-level feature descriptors and then encodes them into a sequence of hidden state vectors using a Long Short Term Memory (LSTM) network [11]. Once all frames are encoded, the decoder generates a sentence by first using the final encoded state as the input to the decoder and then by feeding back the generated words at each step to the decoder LSTM until the sentence is complete.

An encoder-decoder framework allows both the input and the output to be of variable length and has shown prominence in a related task - machine translation [2, 21]. However, machine translation tasks need not consider input and output from different modalities as they only deal with text. S2VT does handle two different modalities with the input being a sequence of frames and the output being sequence of words, but it is limited in the sense that both the input and the output, individually, explore information from single a modality only. Specifically, the input modality is visual while the output modality is textual.

With an eye to explore additional information that is often available with the web videos, such as audio and broad topic (category) of the videos, we have extended the S2VT framework to a multimodal video description framework denoted as the "*MMVD*". This framework supplements the visual information with audio and textual features (derived from the video category information). Such input from three different modalities (visual, audio and textual) enables the generated description to be nearly as complex and rich as human generated descriptions are. In addition, we show that employing a committee of models where each model is an expert in describing videos from a specific category is advantageous than a single model trying to describe videos from multiple categories. We have also simplified the S2VT model by using a single layer LSTM for both encoding and decoding.

The performance of the proposed approach is validated using a publicly available benchmark dataset (MSR-VTT [27]) of web videos which comes with several challenges including diverse content and diverse as well as noisy descriptions.

2. APPROACH

As our model builds on the S2VT model [26], we first describe this model briefly and then describe our approach.

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2984066

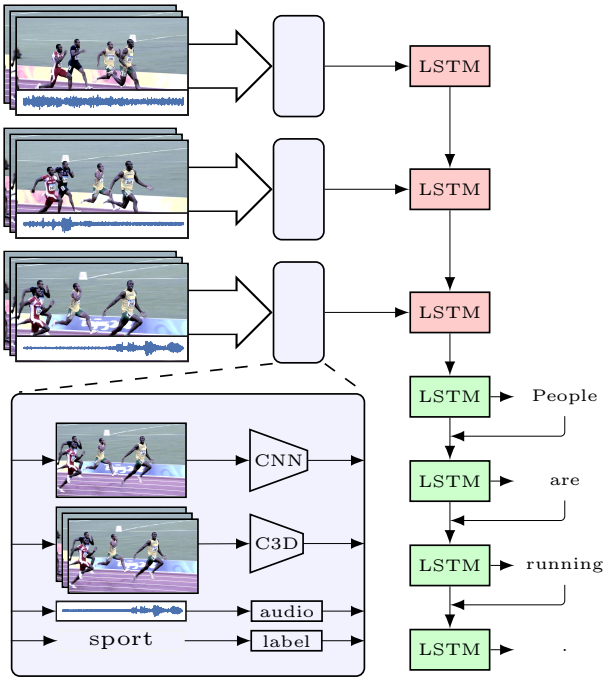

Figure 1: Proposed MMVD approach exploits different modalities of the video, which are first encoded using an LSTM, and then decoded to predict a sentence description. To truly understand the video we propose to rely on static frame features, 3D temporal features across frames, audio, and the video domain or category.

2.1 S2VT video description framework

The S2VT model reads a sequence of input frames ($x_1, x_2, ..., x_n$), encodes each frame to a fixed dimensional vector representation, and then decodes this vector to a sequence of output words ($y_1, y_2, ..., y_m$). S2VT uses LSTM [11] units for modeling long-range temporal dependencies in sequences. LSTMs encode the input sequence in their hidden representations (h_n) and then decode the output sequence from this representation. In the encoding phase, an LSTM computes a sequence of hidden states ($h_1, h_2, ..., h_n$) from the input sequence ($x_1, x_2, ..., x_n$). During decoding, the model defines the joint probability over the output sequence ($y_1, y_2, ..., y_m$) as a product of the conditionals as $\prod_{t=1}^{m} p(y_t|h_{n+t-1}, y_{t-1})$, where the conditional probability of each output word y_t is obtained by a *softmax* over all the words in the vocabulary. The model is trained end-to-end using standard backpropagation techniques where the following log-likelihood of the predicted words is maximized for the model parameters θ, where θ denotes all trainable weights and biases.

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{m} \log p(y_t|h_{n+t-1}, y_{t-1}; \theta) \quad (1)$$

After encoding the frames, the <BOS> (beginning-of-sentence) tag is fed to the LSTM while an <EOS> (end-of-sentence) tag terminates each sentence. During decoding at training time, the input to the LSTM is the embedded representations of the ground truth words while at test time, the word with the maximum probability after the softmax is input to the LSTM until <EOS> token is emitted.

2.2 MMVD model

Our approach MMVD (Multimodal Video Description) extends S2VT [26] to exploit additional multimodal features and is implemented in TensorFlow [1]. Additionally, instead of using a stack of two LSTM layers like S2VT, we use a single layer of LSTMs which not only simplifies the model but also reduces the memory requirement considerably, thus allowing us to incorporate many additional features. A schematic of the MMVD model is shown in Fig. 1.

2.2.1 *Multimodal input features*

Specifically, MMVD incorporates the following features from different modalities to generate the descriptor for the sequence of input frames from each video.

Object recognition features: The deep CNN models applied to image classification and detection tasks (ImageNet) provide a strong visual representation of objects and scenes depicted in the video frames. We used the state-of-the-art ResNet [9] (winner of ILSVRC 2015) features which uses special skip connections and features a heavy use of batch normalization. Transferring knowledge from convnets pre-trained on 1.2M+ images with 1000 category labels helps create open ended descriptions of videos in the wild with large vocabularies. We use 2048 dimensional features from the global average pooling (pool5) layer after the last convoluational layer of ResNet.

Action recognition features: Though ResNet CNN features efficiently capture different visual concepts in static frames, they lack dynamic features that can capture movement/motion in videos. A simple yet effective approach to learn temporal dynamics in videos was proposed by extending a 2-D convnet to a deep 3-D convolutional network (C3D) [23]. Unlike S2VT where optical flow features are used to model motion patterns of activity, the use of C3D features in MMVD allows it to learn relevant motion information from videos in an end-to-end fashion. In our experiments, C3D features from the fc-6 layer of the model pre-trained on Sport1M [12] were used.

Video category information: In our ablation studies, we show that the category or domain of a video (*e.g.*, sports, cooking, music *etc.*) carries a lot of information for generating a proper description of the video. As shown in section 3.3, the background knowledge about the videos in form of their category information helps to discern between videos which have only fine-grained visual differences. We use the category labels supplied with the MSR-VTT video clips as an additional feature descriptor and show that the video domain information can be an useful resource for natural language video description.

Audio features: Although audio tracks associated with the web videos can be useful for generating descriptions of such videos, this information has, traditionally, not been used to effect. Towards this objective, we used the popular audio feature - Mel Frequency Cepstral Coefficients which have been used widely in various audio processing tasks such as automatic speech recognition, music transcription, and environment classification [4, 10, 17]. Feature extraction was performed with *pyAudioAnalysis* [7] for evenly sampled 1 second audio segments aligned with corresponding video frames sampled for object recognition features. Actual audio descriptors were represented by average value and standard deviation of 34 audio features (including 13 MFCCs).

We embed the features from various modalities to a lower dimensional space, where the parameters of the embedding are learned jointly with the description task. Then we concatenate the different embedded features for every time step to create a single feature vector as input to the LSTM.

We also trained models specifically for each video category. For this, a base model is trained on the whole training data by using the combined features as described above. Then the different "*expert*" models are obtained by finetuning the base model with paired video-caption data from specific categories only. During inference every "*expert*" is used to produce descriptions for the corresponding category (the category is provided with the video at test time).

2.2.2 *Model details*

Our model starts by sampling 26 uniformly spaced frames from each video clip. The ResNet and C3D features features are extracted from these 26 frames as the anchor frames. Similarly, audio features are also extracted from one second clips starting with these anchor frames. For category label features, we used one-hot representations of the possible 20 category labels. Though S2VT allows variable length input, to reduce computation we experimented with a fixed number of frames sampled uniformly. The performance difference turned out to be negligible with significantly lower computation overhead.

The increase in the number of input modalities comes with the cost of increased memory and computational requirements as the total dimensionality of the input to the system increases with the number of different features. As a result, we reduce the dimensionality of each feature by passing them through embedding layers (trainable fully connected layers with no activation function). The embedded features are concatenated and passed in as input to the LSTM. The encoder and decoder LSTMs share the same weights. No preprocessing of the text data was done. We used a vocabulary of size 23,667 (of which 13,627 tokens have 2 or more occurrences). The LSTM hidden state size was 512. To avoid overfitting, we apply dropout with a rate of 0.5, which has been shown to be beneficial in video description [20]. The model was optimized by the Adam optimizer with initial learning rate of 5×10^{-4} in batches of size 100.

Some of the hyperparameters are motivated by the properties of the MSR-VTT dataset with a goal to simplify the final model. For example, the strategy to sample 26 uniformly spaced frames was chosen as this number made sure that events in videos from the dataset are not underrepresented by the sampling strategy.

3. EXPERIMENTS

In this section we first describe the dataset and discuss its challenges. Then we detail the evaluation protocol and analyze the experimental evaluation.

3.1 Dataset

In this work, we train and evaluate our models on the MSR-VTT dataset [27]. MSR-VTT is a large-scale dataset collected for the task of describing videos with natural language. The dataset provides 41.2 hours of web videos as 10,000 clips covering diverse visual contents in 20 broad categories or domains. Each clip is annotated with 20 natural language descriptions produced by AMT workers. In addition to "in the wild" nature of the videos, an interesting characteristic of this dataset is the presence of audio. The audio information is complimentary to the visual features and our method takes advantage of it as the human annotators also based their descriptions not only on the video but also on the accompanying sound.

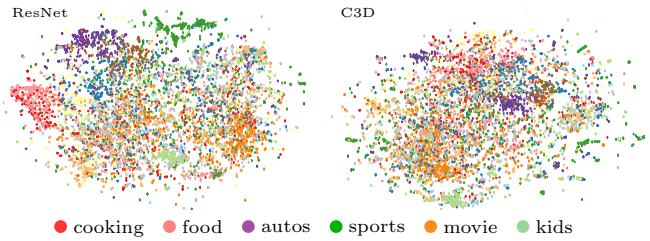

Figure 2: Feature embedding visualizations of visual descriptors on MSR-VTT using t-SNE. It should be noted that semantically close categories like 'food/drinks' and 'cooking' form a common cluster while generic category like 'movie' has large diversity. Each clip is visualized as a point and clips belonging to same category have same color. While for semantically different categories the visual features may be efficient they may not be efficient in generating specific descriptions for visually similar categories.

3.2 Challenges of the data

During our experiments we observed several additional challenges in the dataset. These include noisy text data, high variation in video and sentence length, unavailable audio streams *etc.* as described next.

- Though the total vocabulary size of all the sentences in the training split is 23,667, a total of 10,040 words appear only once. In addition, a comparison with the 400K length vocabulary (based on Wikipedia 2014 + Gigaword 5) used for training a Glove [19] embedding, revealed that 836 words out of 23,667 words are out of vocabulary yet they appear in the dataset multiple times (at least twice). These primarily contain misspelled words like 'backetball' or 'peson'. Such noisy captions add to the challenge of learning a good language model for this dataset.
- As described in section 2.2.1 and as will be shown later, audio features are useful in generating descriptions for "in the wild" videos. However, around 12.5 % of the videos in this dataset do not contain audio and the absence of this useful information makes the task more challenging.
- We also noticed that 95% of the sentences are shorter than 17 tokens (or words) and 95% of video clips in dataset are shorter that 26 seconds. Such clip duration is another motivation to choose 26 anchor frames for each video as this makes sure that the majority of the videos are not under-represented as a result of sampling.
- We also studied the semantic separability of the video clips using the visual descriptors (ResNet and C3D). Fig. 2 visualizes the feature embedding on training and validation clips. We extracted both features from the 26 frames for all training and validation clips. From each clip, the features are mean-pooled across the frames and are projected on 2-dimensional space using t-SNE [24]. The generally intermingled clusters indicate the diversity between the clips along with strong intra category variation. Fig. 2 shows that these visual descriptors may not be a good candidate for describing semantically similar categories of videos (*e.g.*, 'food' and 'cooking') or a generic category like 'movie'. In such cases, audio or topic information of the video become handy to get human like descriptions.

3.3 Results and discussion

We experimented with different modalities individually and with several combinations of them. Table 1 provides the performance of the model on the MSR-VTT validation data

Figure 3: Examples of sentences produced by the proposed ResNet, ResNet+categories, ResNet+all modalities, and the Committee of experts models. Each video clip is represented by four frames with category label shown in top-left.

after training on the training split. Quantitave evaluation of the model is done using 4 different metrics - BLEU@4 [18], METEOR [3], CIDEr [25] and ROUGE-L [16]. We adopted an early stopping criterion and stopped the training when the model starts to overfit (roughly 10 epochs for all the different trials). The best model was chosen by first normalizing all 4 scores on validation data and then selecting the model which gives the highest sum of scores.

The first two rows in the table show that the model performance is pretty low when only category or audio information is used. This is reasonable as the description is generated without using any visual information. The next two rows show the performance of the model with visual feature descriptors. These two features give a significant boost in terms of all 4 metrics. However, the slightly better performance of ResNet descriptors compared to the C3D can possibly be attributed to greater depth and use of skip connections in state-of-the-art ResNet. Moreover, the data on which the two feature descriptors were pretrained are significantly different. The ImageNet data (in case of ResNet) comes with images from more than 1000 diverse categories while the Sport1M data (in case of C3D) contains sports related videos only, resulting in ResNet serving as a better descriptor for "in the wild" videos seen in MSR-VTT.

The next 3 rows show the performance when the respective features are added in order. Thus '+C3D' implies the use of both ResNet and C3D features. Similarly, '+categories' means the use of three features (ResNet, C3D and category information) simultaneously and so on. We see that the use of C3D features and category information does not change the performance much in terms of BLEU@4, METEOR or ROUGE-L. However, the CIDEr score gets continually improved with the addition of these features. While METEOR compares exact token matches, stemmed tokens, paraphrase matches and WordNet synonyms based semantic matches, BLEU@4 tries to give more weight to human-like grammatically correct sentences. On the other hand, CIDEr measures similarity of a generated sentence to reference sentences by counting TF-IDF weighted common n-grams. Consequently, this metric rewards sentences for generating n-grams of uncommon words and is a good fit for measuring the quality of the generated sentences where a diverse set of reference sentences for the videos are available as is the case with MSR-VTT. The performance is boosted in terms of all the metrics by the incorporation of audio features signifying the importance of these features in generating natural language descriptions for web videos.

Table 1: Performance comparison for various feature descriptors on the validation set. +C3D, +categories, +audio were added in this order to ResNet model. Thus +audio line shows results for all modalities. "Committee" refers to the "committee of experts".

Descriptors	BLEU@4	METEOR	CIDEr	ROUGE-L
categories	0.298	0.228	0.236	0.548
audio	0.301	0.222	0.184	0.544
C3D	0.374	0.264	0.389	0.594
ResNet	0.389	0.269	0.400	0.605
+C3D	0.385	0.267	0.411	0.601
+categories	0.381	0.270	0.418	0.597
+audio	0.395	0.277	0.442	0.610
committee	0.407	0.286	0.465	0.610

The last row shows the performance of the "*committee of experts*" where "+audio" was taken as the base model on which the experts were trained. It gives a significant improvement in terms of almost all metrics (except ROUGE-L for which it stays the same). A possible reason for this boost is the fact that the experts learn to produce better sentences specific to different categories when they are fine tuned with paired video-caption data specific to a video category.

Fig. 3 shows descriptions generated by MMVD on sample test videos from MSR-VTT using 4 different combinations of modalities. It can be seen that the use of all modalities helps to produce more accurate sentences, *e.g.*, it can identify that the left video is about a cartoon character. "Committee of experts" can even tell that the video is a result of a person playing a video game. Similarly, for the video on the right, "committee" can correctly identify that it is a "how to" video.

We used the following ablation experiment to estimate the human level performance on this dataset. We removed one random ground truth sentence from every annotated video and treated these sentences as a human generated description for each video. The remaining 19 sentences for each video were treated as the ground truth descriptions with which the BLEU@4 and METEOR score were calculated. This naive approach gave us a METEOR score of 0.3 and a BLEU@4 score of 0.36 as an estimate of the human performance on MSR-VTT dataset. From this study and from the results shown in Table 1, it can be seen that our approach outperforms humans in terms of BLUE@4 score and gets close to human level performance in terms of METEOR. Additional results, analysis and code are available in https://github.com/VisionLearningGroup/MMVD

4. CONCLUSION

This paper proposes a MultiModal approach to Video Description (MMVD). Our MMVD model exploits information from several diverse modalities to generate natural language description of web videos using a recurrent encoder-decoder framework. In particular, we have shown that the use of audio information and the topic of a video can supplement the visual features significantly for this task. Despite its conceptual simplicity, our model achieves state-of-the-art performance on the difficult MSR-VTT dataset which comes with challenges both in audio-visual and language domain. The future directions of our research will be to explore different fusion strategies including speech recognition on audio tracks for better video description.

Acknowledgments This work was supported by DARPA under AFRL grant FA8750-13-2-0026 and a Google Faculty Research Award.

5. REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*, 2015.

[3] S. Banerjee and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Association for Computational Linguistics Workshop*, 2005.

[4] F. Beritelli and R. Grasso. A Pattern Recognition System for Environmental Sound Classification based on MFCCs and Neural Networks. In *IEEE International Conference on Signal Processing and Communication Systems*, pages 1–4, 2008.

[5] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. In *European Conference on Computer Vision*, 2010.

[7] T. Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS ONE*, 10(12):1–17, 12 2015.

[8] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition. In *IEEE International Conference on Computer Vision*, 2013.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[11] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[13] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. In *AAAI Conference on Artificial Intelligence*, 2013.

[14] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and Generating Simple Image Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[15] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing Simple Image Descriptions using Web-scale n-grams. In *Conference on Computational Natural Language Learning*, 2011.

[16] C.-Y. Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Association for Computational Linguistics Workshop*, volume 8, 2004.

[17] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *International Symposium on Music Information Retrieval*, 2000.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Association for Computational Linguistics*, pages 311–318, 2002.

[19] J. Pennington, R. Socher, and C. D. Manning. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

[20] A. Rohrbach, M. Rohrbach, and B. Schiele. The Long-Short Story of Movie Description. In *German Conference on Pattern Recognition*, 2015.

[21] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 2014.

[22] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild. In *International Conference on Computational Linguistics*, 2014.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*, 2015.

[24] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[25] R. Vedantam, L. C. Zitnick, and D. Parikh. Cider: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[26] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to Sequence - Video to Text. In *IEEE International Conference on Computer Vision*, 2015.

[27] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[28] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image Parsing to Text Description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.