

Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning

Xin Wang and Yuan-Fang Wang and William Yang Wang

University of California, Santa Barbara

{xwang, yfwang, william}@cs.ucsb.edu

Abstract

A major challenge for video captioning is to combine audio and visual cues. Existing multi-modal fusion methods have shown encouraging results in video understanding. However, the temporal structures of multiple modalities at different granularities are rarely explored, and how to selectively fuse the multi-modal representations at different levels of details remains uncharted. In this paper, we propose a novel hierarchically aligned cross-modal attention (HACA) framework to learn and selectively fuse both global and local temporal dynamics of different modalities. Furthermore, for the first time, we validate the superior performance of the deep audio features on the video captioning task. Finally, our HACA model significantly outperforms the previous best systems and achieves new state-of-the-art results on the widely used MSR-VTT dataset.

1 Introduction

Video captioning, the task of automatically generating a natural-language description of a video, is a crucial challenge in both NLP and vision communities. In addition to visual features, audio features can also play a key role in video captioning. Figure 1 shows an example where the caption system made a mistake analyzing only visual features. In this example, it could be very hard even for a human to correctly determine if the girl is singing or talking by only watching without listening. Thus to describe the video content accurately, a good understanding of the audio signature is a must.

In the multi-modal fusion domain, many approaches attempted to jointly learn temporal features from multiple modalities (Wu et al., 2014a), such as feature-level (early) fusion (Ngiam et al., 2011; Ramanishka et al., 2016), decision-level (late) fusion (He et al., 2015), model-level fusion (Wu et al., 2014b), and attention fusion (Chen



Ground Truth: A girl is singing.
A girl sings to a song.
Video Only: A woman is talking in a room.
Video + Audio: A girl is singing a song.

Figure 1: A video captioning example.

and Jin, 2016; Yang et al., 2017), etc. But these techniques do not learn the cross-modal attention and thus fail to selectively attend to a certain modality when producing the descriptions.

Another issue is that little efforts have been exerted on utilizing temporal transitions of the different modalities with varying analysis granularities. The temporal structures of a video are inherently layered since the video usually contains temporally sequential activities (*e.g.* a video where a person reads a book, then throws it on the table. Next, he pours a glass of milk and drinks it). There are strong temporal dependencies among those activities. Meanwhile, to understand each of them requires understanding many action components (*e.g.*, pouring a glass of milk is a complicated action sequence). Therefore we hypothesize that it is beneficial to learn and align both the high-level (global) and low-level (local) temporal transitions of multiple modalities.

Moreover, prior work only employed hand-crafted audio features (*e.g.* MFCC) for video captioning (Ramanishka et al., 2016; Xu et al., 2017; Hori et al., 2017). While deep audio features have shown superior performance on some audio processing tasks like audio event classification (Hershey et al., 2017), their use in video captioning needs to be validated.

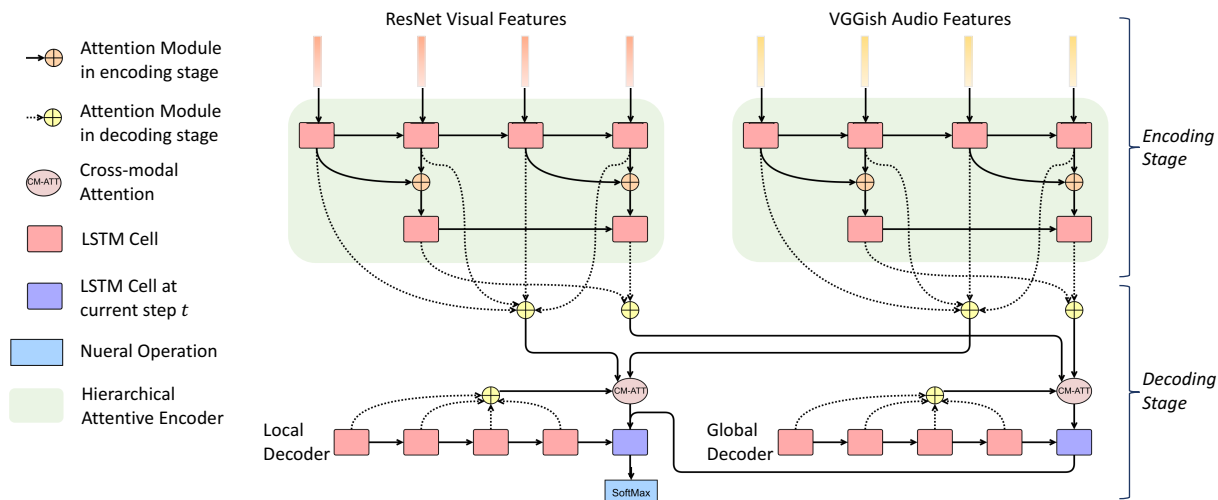


Figure 2: Overview of our HACA framework. Note that in the encoding stage, for the sake of simplicity, the step size of high-level LSTM in both hierarchical attentive encoders is 2 here, but in practice usually they are set much longer. In the decoding stage, we only show the computations of the time step t (the decoders have the same behavior at other time steps).

In this paper, we propose a novel hierarchically aligned cross-modal attentive network (HACA) to learn and align both global and local contexts among different modalities of the video. The goal is to overcome the issues mentioned above and generate better descriptions of the input videos. Our contributions are fourfold: (1) we invent a hierarchical encoder-decoder network to adaptively learn the attentive representations of multiple modalities, including visual attention, audio attention, and decoder attention; (2) our proposed model is capable of aligning and fusing both the global and local contexts of different modalities for video understanding and sentence generation; (3) we are the first to utilize deep audio features for video captioning and empirically demonstrate its effectiveness over hand-crafted MFCC features; and (4) we achieve the new state of the art on the MSR-VTT dataset.

Among the network architectures for video captioning (Yao et al., 2015; Venugopalan et al., 2015b), sequence-to-sequence models (Venugopalan et al., 2015a) have shown promising results. Pan et al. Pan et al. (2016) introduced a hierarchical recurrent encoder to capture the temporal visual features at different levels. Yu et al. (2016) proposed a hierarchical decoder for paragraph generation, and most recently Wang et al. (2018) invented a hierarchical reinforced framework to generate the caption phrase by phrase. But none had tried to model and align the global

and local contexts of different modalities as we do. Our HACA model does only learn the representations of different modalities at different granularities, but also align and dynamically fuse them both globally and locally with hierarchically aligned cross-modal attentions.

2 Proposed Model

Our HACA model is an encoder-decoder framework comprising multiple hierarchical recurrent neural networks (see Figure 2). Specifically, in the encoding stage, the model has one hierarchical attentive encoder for each input modality, which learns and outputs both the local and global representations of the modality. (In this paper, visual and audio features are used as the input and hence there are two hierarchical attentive encoders as shown in Figure 2; it should be noted, however, that the model seamlessly extends to more than two input modalities.)

In the decoding stage, we employ two cross-modal attentive decoders: the *local decoder* and the *global decoder*. The global decoder attempts to align the global contexts of different modalities and learn the global cross-modal fusion context. Correspondingly, the local decoder learns a local cross-modal fusion context, combines it with the output from the global decoder, and predicts the next word.

2.1 Feature Extractors

To exploit visual and audio cues, we use the pre-trained convolutional neural network (CNN) models to extract deep visual features and deep audio features correspondingly. More specifically, we utilize the ResNet model for image classification (He et al., 2016) and the VGGish model for audio classification (Hershey et al., 2017).

2.2 Attention Mechanism

For a better understanding of the following sections, we first introduce the soft attention mechanism. Given a feature sequence $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ and a running recurrent neural network (RNN), the context vector \mathbf{c}_t at time step t is computed as a weighted sum over the sequence:

$$\mathbf{c}_t = \sum_{k=1}^n \alpha_{tk} \mathbf{x}_k, \quad (1)$$

These attention weights $\{\alpha_{tk}\}$ can be learned by the attention mechanism proposed in (Bahdanau et al., 2014), which gives higher weights to certain features that allow better prediction of the system’s internal state.

2.3 Hierarchical Attentive Encoder

Inspired by Pan et al. (2016), the hierarchical attentive encoder consists of two LSTMs and the input to the low-level LSTM is a sequence of temporal features $\{\mathbf{f}_i^{eL}\}$ and $i \in \{1, \dots, n\}$:

$$\mathbf{o}_i^{eL}, \mathbf{h}_i^{eL} = e_L(\mathbf{f}_i^{eL}, \mathbf{h}_{i-1}^{eL}), \quad (2)$$

where e_L is the low-level encoder LSTM, whose output and hidden state at step i are \mathbf{o}_i^{eL} and \mathbf{h}_i^{eL} respectively. As shown in Figure 2, different from a stacked two-layer LSTM, the high-level LSTM here operates at a lower temporal resolution and runs one step every s time steps. Thus it learns the temporal transitions of the segmented feature chunks of size s . Furthermore, an attention mechanism is employed between the connection of these two LSTMs. It learns the context vector of the low-level LSTM’s outputs of the current feature chunk, which is then taken as the input to the high-level LSTM at step j . In formula,

$$\mathbf{f}_j^{eH} = \sum_{k=s(j-1)+1}^{sj} \alpha_{jk} \mathbf{o}_k^{eL}, \quad (3)$$

$$\mathbf{o}_j^{eH}, \mathbf{h}_j^{eH} = e_H(\mathbf{f}_j^{eH}, \mathbf{h}_{j-1}^{eH}), \quad (4)$$

where e_H denotes the high-level LSTM whose output and hidden state at j are \mathbf{o}_j^{eH} and \mathbf{h}_j^{eH} .

Since we are utilizing both the visual and audio features, there are two hierarchical attentive encoders (v for visual features and a for audio features). Hence four sets of representations are learned in the encoding stage: high-level and low-level visual feature sequences ($\{\mathbf{o}_j^{vH}\}$ and $\{\mathbf{o}_i^{vL}\}$), and high-level and low-level audio feature sequences ($\{\mathbf{o}_j^{aH}\}$ and $\{\mathbf{o}_i^{aL}\}$).

2.4 Globally and Locally Aligned Cross-modal Attentive Decoder

In the decoding stage, the representations of different modalities at the same granularity are aligned separately with individual attentive decoders. That is, one decoder is employed to align those high-level features and learn a high-level (global) cross-modal embedding. Since the high-level features are the temporal transitions of larger chunks and focus on long-range contexts, we call the corresponding decoder as *global decoder* (d_G). Similarly, the companion *local decoder* (d_L) is used to align the low-level (local) features that attend to fine-grained and local dynamics.

At each time step t , the attentive decoders learn the corresponding visual and audio contexts using the attention mechanism (see Figure 2). In addition, our attentive decoders also uncover the attention over their own previous hidden states and learn aligned decoder contexts \mathbf{c}_t^{dL} and \mathbf{c}_t^{dG} :

$$\mathbf{c}_t^{dL} = \sum_{k=1}^{t-1} \alpha_{tk}^{dL} \mathbf{h}_k^{dL}, \quad \mathbf{c}_t^{dG} = \sum_{k=1}^{t-1} \alpha_{tk}^{dG} \mathbf{h}_k^{dG}. \quad (5)$$

Paulus et al. (2017) also show that decoder attention can mitigate the phrase repetition issue.

Each decoder is equipped with a *cross-modal attention*, which learns the attention over contexts of different modalities. The cross-modal attention module selectively attends to different modalities and outputs a fusion context \mathbf{c}_t^f :

$$\mathbf{c}_t^f = \tanh(\beta_{tv} \mathbf{W}_v \mathbf{c}_t^v + \beta_{ta} \mathbf{W}_a \mathbf{c}_t^a + \beta_{td} \mathbf{W}_d \mathbf{c}_t^d + b), \quad (6)$$

where \mathbf{c}_t^v , \mathbf{c}_t^a , and \mathbf{c}_t^d are visual, audio and decoder contexts at step t respectively; \mathbf{W}_v , \mathbf{W}_a and \mathbf{W}_d are learnable matrices; β_{tv} , β_{ta} and β_{td} can be learned in a similar manner of the attention mechanism in Section 2.2.

The global decoder d_G directly takes as the input the concatenation of the global fusion context \mathbf{c}_t^{fG} and the word embedding of the generated

word w_{t-1} at previous time step:

$$\mathbf{o}_t^{d_G}, \mathbf{h}_t^{d_G} = d_G([\mathbf{c}_t^{f_G}, \text{emb}(w_{t-1})], \mathbf{h}_{t-1}^{d_G}). \quad (7)$$

The global decoder’s output $\mathbf{o}_t^{d_G}$ is a latent embedding which represents the aligned global temporal transitions of multiple modalities. Differently, the local decoder d_L receives the latent embedding $\mathbf{o}_t^{d_G}$, mixes it with the local fusion context $\mathbf{c}_t^{f_L}$, and then learns a uniform representation $\mathbf{o}_t^{d_L}$ to predict the next word. In formula,

$$\mathbf{o}_t^{d_L}, \mathbf{h}_t^{d_L} = d_L([\mathbf{c}_t^{f_L}, \text{emb}(w_{t-1}), \mathbf{o}_t^{d_G}], \mathbf{h}_{t-1}^{d_L}). \quad (8)$$

2.5 Cross-Entropy Loss Function

The probability distribution of the next word is

$$p(w_t|w_{1:t-1}) = \text{softmax}(\mathbf{W}_p[\mathbf{o}_t^{d_L}]), \quad (9)$$

where \mathbf{W}_p is the projection matrix. $w_{1:t-1}$ is the generated word sequence before step t . θ be the model parameters and $w_{1:T}^*$ be the ground-truth word sequence, then the cross entropy loss

$$\mathcal{L}(\theta) = - \sum_{t=1}^T \log p(w_t^*|w_{1:t-1}^*, \theta). \quad (10)$$

3 Experimental Setup

Dataset and Preprocessing We evaluate our model on the MSR-VTT dataset (Xu et al., 2016), which contains 10,000 videos clips (6,513 for training, 497 for validation, and the remaining 2,990 for testing). Each video contains 20 human annotated reference captions collected by Amazon Mechanical Turk. To extract the visual features, the pretrained ResNet model (He et al., 2016) is used on the video frames which are sampled at $3fps$. For the audio features, we process the raw WAV files using the pretrained VGGish model as suggested in Hershey et al. (2017)¹.

Evaluation Metrics We adopt four diverse automatic evaluation metrics: BLEU, METEOR, ROUGE-L, and CIDEr-D, which are computed using the standard evaluation code from MS-COCO server (Chen et al., 2015).

¹<https://github.com/tensorflow/models/tree/master/research/audioset>

Models	BLEU-4	METEOR	ROUGE-L	CIDEr
Top-3 Results from MSR-VTT Challenge 2017				
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
State Of The Arts				
CIDEnt-RL	40.5	28.4	61.4	51.7
Dense-Cap	41.4	28.3	61.1	48.9
HRL	41.3	28.7	61.7	48.0
Our Models				
ATT(v)	39.6	27.4	59.7	45.8
CM-ATT(va)	41.7	28.6	61.2	48.2
CM-ATT(vad)	41.9	29.1	61.5	48.0
HACA(w/o align)	42.8	29.0	61.8	48.9
HACA	43.4	29.5	61.8	49.7

Table 1: Results on the MSR-VTT dataset.

Training Details All the hyperparameters are tuned on the validation set. The maximum number of frames is 50, and the maximum number of audio segments is 20. For the visual hierarchical attentive encoders (HAE), the low-level encoder is a bidirectional LSTM with hidden dim 512 (128 for the audio HAE), and the high-level encoder is an LSTM with hidden dim 256 (64 for the audio HAE), whose chunk size s is 10 (4 for the audio HAE). The global decoder is an LSTM with hidden dim 256 and the local decoder is an LSTM with hidden dim 1024. The maximum step size of the decoders is 16. We use word embedding of size 512. Moreover, we adopt Dropout (Srivastava et al., 2014) with a value 0.5 for regularization. The gradients are clipped into the range $[-10, 10]$. We initialize all the parameters with a uniform distribution in the range $[-0.08, 0.08]$. Adadelta optimizer (Zeiler, 2012) is used with batch size 64. The learning rate is initially set as 1 and then reduced by a factor 0.5 when the current CIDEr score does not surpass the previous best for 4 epochs. The maximum number of epochs is set as 50, and the training data is shuffled at each epoch. Schedule sampling (Bengio et al., 2015) is employed to train the models. Beam search of size 5 is used during the test time inference.

4 Results

4.1 Comparison with State Of The Arts

In Table 1, we first list the top-3 results from the MSR-VTT Challenge 2017: v2t_navigator (Jin et al., 2016), Aalto (Shetty and Laaksonen, 2016), and VideoLAB (Ramanishka et al., 2016). Then we compare with the state-of-the-art methods on the MSR-VTT dataset: CIDEnt-RL (Pasunuru and

Features	BLEU-4	METEOR	ROUGE-L	CIDEr
video only	39.6	27.4	59.7	45.8
video + MFCC	40.3	28.5	60.8	47.5
video + VGGish	41.7	28.6	61.2	48.2

Table 2: Performance of the cross-modal attention model with various audio features.

Bansal, 2017), Dense-Cap (Shen et al., 2017), and HRL (Wang et al., 2018). Our HACA model significantly outperforms all the previous methods and achieved the new state of the art on BLEU-4, METEOR, and ROUGE-L scores. Especially, we improve the BLEU-4 score from 41.4 to 43.1. The CIDEr score is the second best and only lower than that of CIDEr-RL which directly optimizes the CIDEr score during training with reinforcement learning. Note that all the results of our HACA method reported here are obtained by supervised learning only.

4.2 Result Analysis

We also evaluate several baselines to validate the effectiveness of the components in our HACA framework (see Our Models in Table 1). $ATT(v)$ is a generic attention-based encoder-decoder model that specifically attends to the visual features only. $CM-ATT$ is a cross-modal attentive model, which contains one individual encoder for each input modality and employs a cross-modal attention module to fuse the contexts of different modalities. $CM-ATT(va)$ denotes the CM-ATT model consisting of visual attention and audio attention, while $CM-ATT(vad)$ has an additional decoder attention.

As presented in Table 1, our $ATT(v)$ model achieves comparable results with the top-ranked results from MSR-VTT challenge. Comparing between $ATT(v)$ and $CM-ATT(va)$, we observe a substantial improvement by exploiting the deep audio features and adding cross-modal attention. The results of $CM-ATT(vad)$ further demonstrates that decoder attention was beneficial for video captioning. Note that to test the strength of the aligned attentive decoders, we provide the results of $HACA(w/o align)$ model, which shares almost same architecture with the HACA model, except that it only has one decoder to receive both the global and local contexts. Apparently, our HACA model obtains superior performance, which therefore proves the effectiveness of the context alignment mechanism.

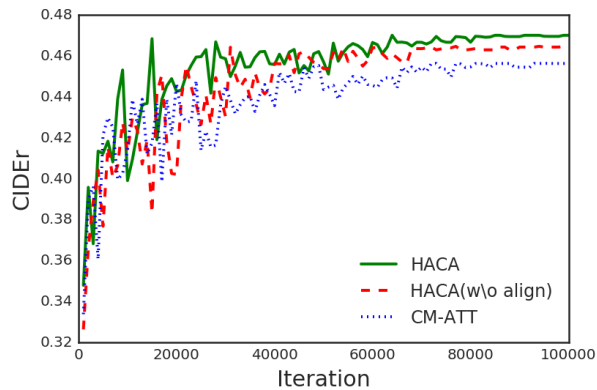


Figure 3: Learning curves of the CIDEr scores on the validation set. Note that greedy decoding is used during training, while beam search is employed at test time, thus the testing scores are higher than the validation scores here.

4.3 Effect of Deep Audio Features

In order to validate the superiority of the deep audio features in video captioning, we illustrate the performance of different audio features applied in the CM-ATT model in Table 2. Evidently, the deep VGGish audio features work better than the hand-crafted MFCC audio features for the video captioning task. Besides, it also shows the importance of understanding and describing a video with the help of audio features.

4.4 Learning Curves

For a more intuitive view of the model capacity, we plot the learning curves of the CIDEr scores on the validation set in Figure 3. Three models are presented: HACA, $HACA(w/o align)$, and CM-ATT. They are trained on same input modalities and all paired with visual, audio and decoder attentions. We can observe that the HACA model performs consistently better than others and has the largest model capacity.

5 Conclusion

We introduce a generic architecture for video captioning which learns the aligned cross-modal attention globally and locally. It can be plugged into the existing reinforcement learning methods for video captioning to further boost the performance. Moreover, in addition to the deep visual and audio features, features from other modalities can also be incorporated into the HACA framework, such as optical flow and C3D features.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*. pages 1171–1179.
- Shizhe Chen and Qin Jin. 2016. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, pages 571–575.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 770–778.
- Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 73–80.
- Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, pages 131–135.
- Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *The IEEE International Conference on Computer Vision (ICCV)*.
- Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. 2016. Describing videos using multi-modal fusion. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, pages 1087–1091.
- Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multi-modal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. pages 689–696.
- Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 1029–1038.
- Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 979–985.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
- Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, pages 1092–1096.
- Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly supervised dense video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rakshith Shetty and Jorma Laaksonen. 2016. Frame- and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, pages 1073–1076.
- Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015a. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. pages 4534–4542.
- 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 *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Xin Wang, Wenhua Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. 2014a. Survey on audiovisual emotion recognition:

databases, features, and data fusion strategies. *AP-SIPA transactions on signal and information processing* 3.

Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. 2014b. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pages 167–176.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 5288–5296.

Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. 2017. Learning multimodal attention lstm networks for video captioning. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, pages 537–545.

Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. 2017. Deep multimodal representation learning from temporal data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. pages 4507–4515.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 4584–4593.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.