Audio-Visual Interpretable and Controllable Video Captioning

Yapeng Tian[1], Chenxiao Guan[1], Justin Goodman[2], Marc Moore[3], and Chenliang Xu[1]

[1]University of Rochester [2]University of Maryland [3]Mississippi State University

1. Introduction

Video captioning [8] aims to automatically generate a natural language sentence to describe the dynamic, potentially complex multimodal scene inside of a video. Most of the previous works [8, 10] focus on exploring better vision-and-language modelings and put less emphasis on the multimodal aspect of video captioning, where audio often reveals important in-scene and out-of-scene information and contributes to the language generation in its unique ways. For example, it adds the difficulty to describe the singing event by watching the audio-mute video in Fig. 1. Although this is not new to the multimedia community, many works over there aim to optimize the video captioning metrics with an uninterpretable fusion strategy. The basic questions as **to what extent different modalities (auditory and visual) contribute to a particular sentence**, and furthermore, **to individual words in a sentence** remain underexplored. It is our belief that unfolding these questions is valuable to making fundamental progress on video captioning.

At first glance, it is seemingly impossible to answer the above questions. The first challenge is that there is no annotation denoting the individual contributions of the auditory or visual modality made to texts in any of the existing video captioning datasets—such a process is difficult to quantify without breakthroughs in Neurophysiology and Psychophysics. Instead, we study these questions from a computational perspective, where we mine signals from audio and video and compete their associations to text.

The second challenge lies on the computational framework. Recurrent neural networks (RNNs) are widely used as decoders for video captioning. Despite the success in modeling sequential dependencies, RNN decoder-based architectures have inherent limitations to perform modality-interpretable video captioning. When generating a word, besides using current provided/attended features and previous words, these models always exploit hidden states of the RNN decoder. The latter contains memorized information from different modalities, which makes the models impossible to disentangle the contributions from individual modalities for predicting the words.

In this paper, we aim to disentangle the interplay of the two modalities and make the first attempt to interpretable audio-visual video captioning. Concretely, we propose a novel multimodal convolutional neural network-based audio-visual video captioning framework without a RNN decoder to ease the design of interpretable structure, and introduce a modality-aware feature aggregation module with defined activation energy to distinguish which modality is more informative for generating words. In addition, the interpretability endows our framework ability on audio-visual controllable sentence generation. In practice, we introduce an audio-visual controller to manipulate the parameters in the modality-aware feature aggregation module allowing the proposed model generate diverse modality-aware captions.

Humans: (1) people singing and dancing.
(2) a group of people singing and dancing.

Interpretability: a group of people are singing and dancing.

Controllability: (a) a man and a woman are singing a song.
(b) a group of people are singing and dancing.
(c) a group of women are dancing.

Figure 1. Audio-visual video captioning with interpretability on word generation and controllability on sentence prediction. The automatically detected audio/visual activated words are highlighted with **red/blue**. We see that visual modality is dominated when generating *people* and *dancing*, and audio content is more informative for predicting *singing*. The trained single model can generate different sentences by setting an audio-visual controller as different values.

2. Overview of Proposed Approach

Given an input visual and audio clip pair $\{V, A\}$, our captioning network aims to generate a natural language sentence $S = (s_1, s_2, \ldots, s_{T_s})$ containing T_s words. Unlike previous RNN-based encoder-decoder video captioning networks, we propose a 2D MMCNN-based audio-visual video captioning framework illustrated in Fig. 2, which is capable of learning decoupled audio-text and visual-text deep feature hierarchies and is more convenient for achieving modality interpretability. We utilize pre-trained CNN models to extract visual features $v \in \mathcal{R}^{T_v \times D_v}$ and audio features $a \in \mathcal{R}^{T_a \times D_a}$ from the input visual clip V and audio clip A. Here, we sample T_v video frames from the given visual clip V and T_a seconds from the given audio clip A. Visual feature dimension for each frame is D_v and audio

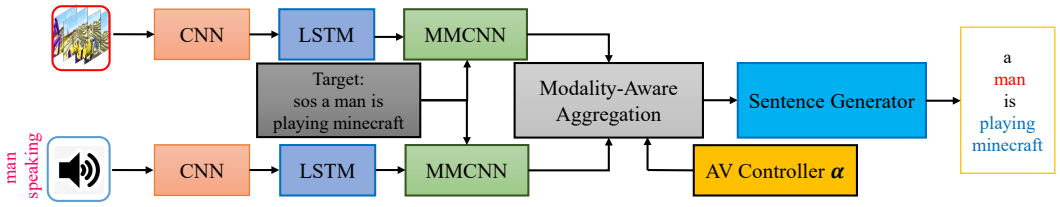

Figure 2. The proposed audio-visual interpretable and controllable video captioning framework. During testing, words in the sentence will be predicted one-by-one. The input video frames only contain content of the video game, but there is man speaking sound in the audio channel. The word *man* will be inferred from activated the auditory modality, and the words *playing* and *minecraft* are mainly from visual modality. We make modality selection decision based on values of audio activation energy and visual activation energy. There is an audio-visual controller α in the modality-aware aggregation module, which balances the importance between audio and visual modalities during sentence prediction.

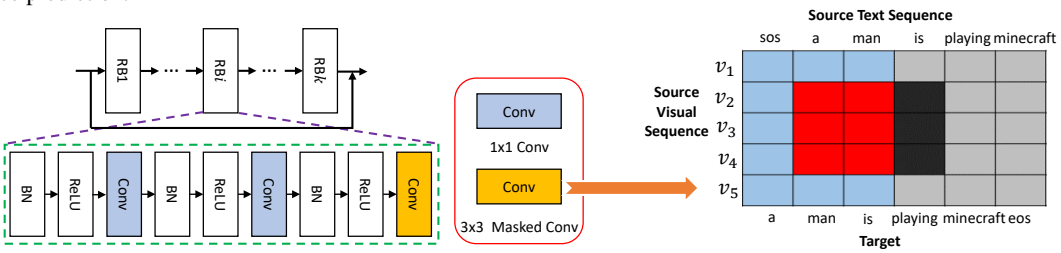

Figure 3. Residual Network with k residual units in the proposed MMCNN and an example for illustrating the masked convolution operation in residual units. On the right, there is a 3×3 masked convolution kernel colorized by red and black boxes for processing a visual-text joint embedding. The masked convolution will only use features from red boxes and cannot access information in black boxes. When predicting the word *is*, the network will only use source visual information and words *sos*, *a*, and *man* (previous words).

feature dimension for each second audio segment is D_a. To explore temporal structures of audio and visual modalities individually, we use two separated LSTMs: one takes visual feature v as input and the other takes audio feature a as input. They model temporal dependencies independently for the two modalities and implicitly align them with the textual sequence. Concerning the interpretability on word generation, we build two separated MMCNNs: one for each modality. Taking the aggregated hidden states from the audio LSTM and the sentence S as inputs, our audio-text MMCNN will predict a joint deep audio-text embedding F^a. Similarly, we can obtain a joint deep visual-text embedding F^v from the visual-text MMCNN. The modality-aware aggregation module takes these embeddings as inputs along with an audio-visual controller and generates a feature for the final sentence generation. The sentence generator predicts words parallel during training and one-by-one during inference.

2.1. Multimodal Convolutional Neural Network

Taking the visual-text MMCNN as an example, we introduce the detail. The visual-text MMCNN mainly contains two parts: visual-text tensor construction and joint deep visual-text feature extraction.

Tensor Construction: For a target sentence S, we first extract word embedding $e_t \in \mathcal{R}^{D_s}$ for each word s_t in S and then combine all words into a matrix $e \in \mathcal{R}^{T_s \times D_s}$. Given the aggregated visual hidden states $h^v \in \mathcal{R}^{T_v \times D_v}$ for a video clip V and word embedding e for the sentence, we construct a 3D tensor $I^v \in \mathcal{R}^{T_s \times T_v \times D_{vs}}$, where $D_{vs} = D_v + D_s$ and $I^v_{ij} = [e_i \; h^v_j]$. Note that, for designing an autoregressive language model, the first word in the sentence will be set to $< sos >$. This tensor is then input to the joint deep feature learning module.

Joint Deep Feature Learning: To learn joint deep representations for visual and textual modalities, we feed the tensor I^v into a deep residual 2D CNN network f_v. The joint visual-text embedding $F^v \in \mathcal{R}^{T_s \times T_v \times D_{vs}}$ can be obtain: $F^v = f_v(I^v)$. Following the deign of residual blocks in the ResNet and considering computation efficiency, we utilize the residual block layout as illustrated in Fig. 3.

Similarly, we can build a audio-text MMCNN to predict the joint deep embedding $F^a \in \mathcal{R}^{T_s \times T_a \times D_{as}}$.

2.2. Modality-Aware Aggregation

The modality-aware aggregation module will adaptively select features over different time steps and across different modalities for captioning generation.

Given $F^a \in \mathcal{R}^{T_s \times T_a \times D_{as}}$ and $F^v \in \mathcal{R}^{T_s \times T_v \times D_{vs}}$, we first use two fully connected layers to align the two tensors with a same feature dimension D_c, and then construct a new tensor $F^c \in \mathcal{R}^{T_s \times T_c \times D_c}$ by concatenating the two tensors along the audio-visual channel, where $T_c = T_v + T_a$.

Let $F^c_i \in \mathcal{R}^{T_c \times D_c}$ be the i-th row of F^c. We will use F^c_i to generate a feature vector $x_i \in \mathcal{R}^{D_c}$, and then predict the $(i+1)$-th word S_{i+1}. The naive and simple way to generate x_i from F^c_i is by max-pooling or mean-pooling. Since mean-pooling will regard the T_c feature vectors are

equally important and max-pooling will highly mix information from different feature vectors, the two methods are modality-ambiguous. Motivated by a self-attention mechanism in [4], we introduce a modality-aware aggregation module to compute x_i from F_i^c:

$$x_i = \sum_{j=1}^{T_c} w_j F_{ij}^c \ , \tag{1}$$

where $F_{ij}^c \in \mathcal{R}^{D_c}$ and $w_j \in [0, 1]$. We define respective activation energies for audio and video, and measure the dominant one that generates a noun or a verb word. The visual activation energy is defined as: $e_i^v = \sum_{j=1}^{T_v} w_j^2$. Similarly, the audio activation energy can be computed as: $e_i^a = \sum_{j=T_v+1}^{T_c} w_j^2$. When the generated word is a noun or a verb, if $e_i^v > e_i^a$, visual content is more important for generating the word; if $e_i^v < e_i^a$, the word is more related to the auditory modality. In this way, our model will have interpretable ability for modality selection during word generation. The weights in Eq. 1 can be computed by:

$$w_1, ..., w_{T_c} = softmax(u) \ , \tag{2}$$

$$u = fc_3(\delta(fc_2(fc_1(F_i)))) \ , \tag{3}$$

where the first Fully-Connected (FC) layer fc_1 aggregates features at each position j of F_i, $fc_1(F_i) \in \mathcal{R}^{T_c}$, the second and third FC layers have $\lfloor T_c/2 \rfloor$ and T_c output neurons, respectively. Here, $u \in \mathcal{R}^{T_c}$, and δ is the ReLU function.

2.3. Audio-Visual Controllable Captioning

The weights $\{w_1, ..., w_{T_c}\}$ indicate the importance of corresponding features for word generation, and two sets of weights: $\{w_1, ..., w_{T_v}\}$ and $\{w_{T_v+1}, ..., w_{T_c}\}$ are associated with visual and audio features, respectively. We introduce a controller $\alpha \in [0, 1]$ to generate two parameters, α_a and α_v, to manipulate the audio and visual weights for audio-visual controllable video captioning:

$$\alpha_v = \begin{cases} \frac{\alpha}{1-\alpha}, & \text{if } \alpha < 0.5 \ , \\ 1, & \text{otherwise} \ , \end{cases} \tag{4}$$

$$\alpha_a = \begin{cases} 1, & \text{if } \alpha < 0.5 \ , \\ \frac{1-\alpha}{\alpha}, & \text{otherwise} \ . \end{cases} \tag{5}$$

With the defined α_a and α_v, we revisit Eq. 1 to compute the feature x_i for audio-visual controllable word generation:

$$x_i = \sum_{j=1}^{T_v} \alpha_v w_j F_{ij}^c + \sum_{j=T_v+1}^{T_c} \alpha_a w_j F_{ij}^c \ . \tag{6}$$

Clearly, the Eq. 1 is a special case of the Eq. 6 ($\alpha = 0.5$). Setting the controller to different values during inference, it will assign different importance to both audio and visual modalities for audio-visual video captioning to generate diverse descriptions for a single video. However, when we directly take a trained model with the modality-aware aggregation module defined in Eq. 1, it fails to generate meaningful and logical sentences for certain α values like $\alpha = 1$. When $\alpha = 1$, $\alpha_a = 0$ and $\alpha_v = 1$, but both α_v and α_a are equal to 1 for the Eq. 1 during training. Therefore, the audio-text MMCNN branch may also make contributions to language modeling and we can not ensure that the visual-text MMCNN learns a individual good language model, which will lead to inaccurate sentence generation.

To overcome the above issue, we introduce a random controller α to train the network for keeping training and testing be more consistent. During training, α will be uniformly sampled from $[0, 1]$ for each batch. In this way, the network can randomly sample different α to penalize audio or visual modality, which makes the model be able to explore the associations between words and individual modalities; adaptively learn to be aware of visual-related, audio-related, or both audio- and visual-related words for sentence generation and discover corresponding events from audio or visual modalities. With the competing (α_a or α_v may be close to 0), both audio-text and visual-text MMCNNs will learn good language models.


Humans: (1) a man is playing with garrys mod.
(2) two men demonstrate a video game.
Our: a man is playing a video game. (man is audible)

Humans: (1) girls dance and sing in a gym.
(2) a group of young girls sing and dance.
Our: a group of girls are singing.

Figure 4. Audio-visual video captioning results with modality selection visualizations. Here, audio activated words and visual activated words are highlighted with **red** and **blue** texts, respectively.

3. Example Results

In this work, we train and evaluate the proposed audio-visual video captioning model on the MSR-VTT [10]. MSR-VTT is a large-scale video description dataset, which has 10,000 video clips over 20 video categories with diverse video content and descriptions, as well as multimodal audio and video streams. We use four commonly used automatic evaluation metrics: BLEU [6], METEOR [2], ROUGE-L [5], and CIDEr [7] to measure similarity between ground truth and automatic video description results.

Figure 4 illustrates the audio-visual interpretability on

Table 1. Performances of the proposed model and other state-of-the-art methods on MSR-VTT test dataset [10].

Models	BLEU-4	METEOR	ROUGE-L	CIDEr
PickNet [3]	38.9	27.2	59.5	42.1
HRL [9]	41.3	28.7	61.7	48.0
GRU-EVE [1]	38.3	28.4	60.7	48.1
Ours	42.7	28.5	61.5	47.2

modality selection of the proposed MMCNN-based audio-visual video captioning framework. For the first example, sound source (*man*) is not visible, but our network predicts the word *man* by activating auditory modality. For the second example, the model finds *girls* from visual information and predict the *singing* based on auditory signal. From these results, we observe that the auditory signal tends to be activated when predicting words related to audio events; and visual information will dominate word generation when describing visual events. To demonstrate the capability of the proposed framework on controllable audio-visual video captioning, some results are illustrated in Fig. 5. We can find that the proposed framework with a single trained model by setting different audio-visual controller values can generate diverse sentences for each video. Fig. 5(I) shows that the singing audio event dominates the sentence generation when $\alpha <= 0.5$; when the audio modality is penalized $\alpha >= 0.6$, the model infers that this is a scene from a movie by partially considering the background music and leveraging whole visual information; when $\alpha = 1$, only visual modality is available, the model predicts there is a soldiers' talking event. For Fig. 5(II), when $\alpha <= 0.2$, the model tries to only describe sound in the video but fails to generate accurate contents; when α becomes larger, it generates audio-visual comprehensive descriptions by predicting *woman* and *talking* from sound and *dish* and *cooking* from visual domain; when audio modality is further penalized ($\alpha >= 0.7$), only the visual event is described and the model is blind on finding who is cooking, because the *woman* is not visible in the video.

Table 1 shows the performances of the proposed method and other SOTA methods on the MSR-VTT test dataset. We can see that the proposed approach equipped with interpretability and controllability can still achieve comparable performance with the current SOTA models.

Acknowledgement: J. Goodman and M. Moore did the work during their REU program with the University of Rochester. This work was supported in part by NSF IIS 1741472, IIS 1813709, and CHE 1764415. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

Humans: (1) music video with slow motion.
(2) clipping from the movie.
Comments: we can hear a man's singing sound from the audio channel and it is a movie clip.
Ours: (a) a man is singing. ($\alpha = 0, ..., 0.5$)
(b) a scene from a movie. ($\alpha = 0.6, \ldots, 0.9$)
(c) soldiers are talking to each other. ($\alpha = 1$)

(I)

Humans: (1) in a kitchen there is women preparing some fried dishes.
(2) there is a woman is talking about a new food.
Ours: (a) a woman is talking about a baby stroller. ($\alpha = 0$)
(b) a woman is talking about a woman's hair. ($\alpha = 0.1, 0.2$)
(c) a woman is talking about a dish. ($\alpha = 0.3$)
(d) a woman is talking about cooking. ($\alpha = 0.4$)
(e) a woman is cooking. ($\alpha = 0.5, 0.6$)
(f) a person is cooking. (α =0.7, 0.8)
(g) someone is cooking food. (α =0.9)
(h) in a kitchen someone is cooking. (α =1)

(II)

Figure 5. Audio-visual controllable video captioning results. During testing, we use a single trained model with setting the audio-visual controller α as 0, 0.1, ..., 1 to generate different captions. As expected, when $\alpha < 0.5$, the model tends to produce more audio-related sentences; when $\alpha > 0.5$, it tends to generate more visual-related sentences; when $\alpha = 0.5$, it may be bias to one modality or make comprehensive audio-visual descriptions.

References

[1] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, 2019. 4

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop*, 2005. 3

[3] Y. Chen, S. Wang, W. Zhang, and Q. Huang. Less is more: Picking informative frames for video captioning. In *ECCV*, 2018. 4

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[5] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 3

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 3

[7] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 3

[8] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 1

[9] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, 2018. 4

[10] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 3, 4