

## Integrating Both Visual and Audio Cues for Enhanced Video Caption

Wangli Hao,<sup>1,4</sup> Zhaoxiang Zhang,<sup>1,2,3,4,\*</sup> He Guan<sup>1,4</sup>

<sup>1</sup>Research Center for Brain-inspired Intelligence, CASIA

<sup>2</sup>National Laboratory of Pattern Recognition, CASIA

<sup>3</sup>CAS Center for Excellence in Brain Science and Intelligence Technology, CAS

<sup>4</sup>University of Chinese Academy of Sciences

{haowangli2015,zhaoxiang.zhang,guanhe2015}@ia.ac.cn

### Abstract

Video caption refers to generating a descriptive sentence for a specific short video clip automatically, which has achieved remarkable success recently. However, most of the existing methods focus more on visual information while ignoring the synchronized audio cues. We propose three multimodal deep fusion strategies to maximize the benefits of visual-audio resonance information. The first one explores the impact on cross-modalities feature fusion from low to high order. The second establishes the visual-audio short-term dependency by sharing weights of corresponding front-end networks. The third extends the temporal dependency to long-term through sharing multimodal memory across visual and audio modalities. Extensive experiments have validated the effectiveness of our three cross-modalities fusion strategies on two benchmark datasets, including Microsoft Research Video to Text (MSRVTT) and Microsoft Video Description (MSVD). It is worth mentioning that sharing weight can coordinate visual-audio feature fusion effectively and achieve the state-of-art performance on both BELU and METEOR metrics. Furthermore, we first propose a dynamic multimodal feature fusion framework to deal with the part modalities missing case. Experimental results demonstrate that even in the audio absence mode, we can still obtain comparable results with the aid of the additional audio modality inference module.

### Introduction

Automatically describing video with natural sentences has potential applications in many fields, such as human-robot interaction, video retrieval. Recently, benefiting from extraordinary abilities of convolutional neural networks (CNN) (Simonyan and Zisserman 2014; Szegedy et al. 2015; 2016), recurrent neural networks (RNN) (Hochreiter and Schmidhuber 1997) and large paired video language description datasets (Xu et al. 2016), video caption has achieved promising successes.

Most video caption frameworks can be simply split into an encoder stage and a decoder stage respectively. Conditioned on a fixed length of visual feature representation offered by encoder, decoder can generate a corresponding video description recurrently. To generate a fixed length video representation, several methods are proposed,

such as pooling over frames (Venugopalan et al. 2014), holistic video representations (Gua ; Rohrbach et al. 2015; 2013), sub-sampling on a fixed number of input frames (Yao et al. 2015) and extracting the last hidden state of recurrent visual feature encoder (Venugopalan et al. 2015).

Those feature encoding methods mentioned above are only based on visual cues. However, videos contain the visual modality and the audio modality. The resonance information underlying them is essential for video caption generation. We believe that the lack of arbitrary modality will result in the loss of information. For example, when a person is lying on the bed and singing a song, traditional video caption methods may generate an incomplete sentence, like "a person is lying on the bed", which may due to the loss of resonance information underling audio modality. If audio features can be integrated into video caption framework, precise sentence "a person is lying on the bed and singing" will be expected to generate.

To thoroughly utilize both visual and audio information, we propose and analyze three multimodal deep fusion strategies to maximize the benefits of visual-audio resonance information. The first one explores the impact on cross-modalities feature fusion from low to high order. The second establishes the visual-audio short-term dependency by sharing weights of corresponding front-end networks. The third extends the temporal dependency to long-term through sharing multimodal memory across visual and audio modalities. Furthermore, a dynamic multimodal feature fusion framework is also proposed to deal with audio modality absent problem during video caption generation.

The contributions of our paper include:

- We present three multimodal feature fusion strategies, to efficiently integrate audio cues into video caption.
- We propose an audio modality inference module to handle audio modality absent problem, through generating audio feature based on the corresponding visual feature of the video.
- Our experimental results based on Microsoft Research-Video to Text (MSR-VTT) and Microsoft Video Description (MSVD) datasets show that our multimodal feature fusion frameworks lead to the improved results in video caption.

\*Corresponding author. (zhaoxiang.zhang@ia.ac.cn)

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Related works

### Video Caption

Early works concerning on video caption can be classified into three groups.

The first category is template-based methods. They first identified the semantic attributes hidden in videos and then derived a sentence structure based on some predefined sentence templates (Krishnamoorthy et al. 2013; Thomason et al. 2014). Then, probabilistic graphical model was utilized to collect the most relevant contents in videos to generate the corresponding sentence. Although sentences generated by these models seemed to be grammatically correct, they were lack of richness and flexibility.

The second category treat video caption as a retrieval problem. They tagged videos with metadata (Aradhye, Toderici, and Yagnik 2009) and then clustered videos and captions based on these tags. Although the generated sentences were more naturally compared to the first group, they were subject to the metadata seriously.

The third category of video caption methods directly map visual representation into specific provided sentences (Venugopalan et al. 2014; Yao et al. 2015; Pan et al. 2016a; 2016b), which take inspiration from image caption (Vinyals et al. 2015; Donahue et al. 2015).

We argue that these video caption methods only rely on visual information while ignoring audio cues, which will restrict the performance of video caption. To handle this problem, we explore to incorporate audio information into video caption.

## Exploiting Audio Information from Videos

Audio sequence underlying videos always carry meaningful information. Recently, many researchers have tried to incorporate audio information into their specific applications. In (Owens et al. 2016), Owens et al. adopted ambient sounds as a supervisory signal for training visual models, their experiments showed that units of trained network supervised by sound signals carried semantic meaningful information about objects and scenes. Ren et al. (Ren et al. 2016) proposed a multimodal Long Short-Term Memory (LSTM) for speaker identification, which referred to locating a person who has the same identity with the ongoing sound in a certain video. Their key point was sharing weights across face and voice to model the temporal dependency over these two different modalities.

Inspired by (Ren et al. 2016), we propose to build temporal dependency across visual and audio modalities through sharing weights for video caption, aiming at exploring whether temporal dependency across visual and audio modalities can capture the resonance information among them or not.

## Memory Extended Recurrent Neural Network

Internal memory in RNN can preserve valuable information for specific tasks. However, it cannot well handle the tasks which need long-term temporal dependency.

To enhance the memory ability of RNN, an external memory has been utilized to extend RNN in some works, such as

Neural Turing Machine (NTM) (Graves, Wayne, and Danihelka 2014), memory network (Weston, Chopra, and Bordes 2014), which is simply dubbed as memory enhanced RNN (ME-RNN).

ME-RNNs have been widely applied in many tasks. Besides handling single task which needs long temporal dependency, such as visual question answering (Xiong, Merity, and Socher 2016) and dialog systems (Dodge et al. 2015), ME-RNNs have been adopted for multi-tasks to model long temporal dependency across different tasks (Liu, Qiu, and Huang 2016).

To explore whether long visual-audio temporal dependency can capture the resonance information among two modalities, we first try to build a visual-audio shared memory across visual and audio modalities for video caption.

## Methods

In this section, we first introduce the basic video caption framework that our work is based on. Then, three multi-modal feature fusion strategies are depicted for video caption respectively. Meanwhile, dynamic multimodal feature fusion framework and its core component AMIN are also presented.

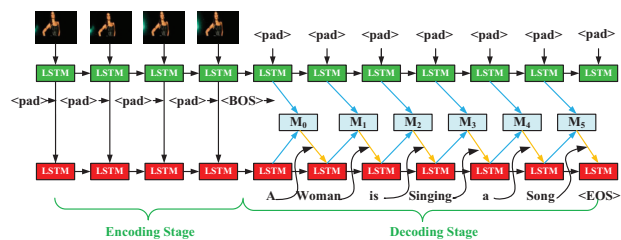


Figure 1: Basic pipeline of our video caption framework.

## Basic Video Caption Framework

Our basic video caption framework is extended from S2VT (sequence to sequence: video to text) model (Venugopalan et al. 2015) and M<sup>3</sup> (multimodal memory modeling) model (Wang et al. 2016), which is shown in Figure 1.

As in Figure 1, encoding stage encodes visual feature and decoding stage generates visual description. Specifically, visual feature inputs are constructed by the top LSTM layer (colored green). Intermediate multimodal memory (colored cyan) layer is shared by visual and textual modalities. Language is modeled by the bottom multimodal memory extended LSTM (colored red), which is conditioned on text sequence input and information reading from multimodal memory. <BOS> and <EOS> tags in Figure 1 indicate the begin-of-sentence and end-of-sentence respectively. <pad> hints that there is no input at the corresponding time step. In addition, the colored blue/orange lines denote writing/reading information into/from memory.

## Multimodal Feature Fusion strategies

### Concatenating Visual and Audio Features

In this section, we propose two different concatenation ways and

present them in Figure 2.

Specifically, one concatenation way is shown in Figure 2 (a). Before LSTM encoder, visual-audio feature pairs of corresponding video clips are directly concatenated together. Then the concatenated features are sent to LSTM encoder. The other concatenation way is presented in Figure 2 (b). Visual-audio feature pairs are first separately sent to the corresponding LSTM encoders. Then the last hidden states of these two LSTM encoders are concatenated together.

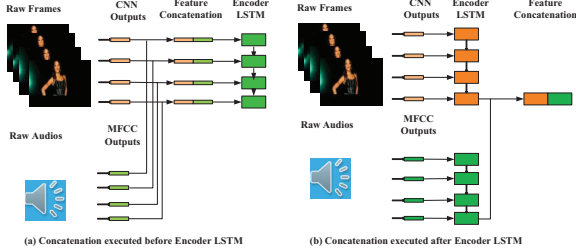


Figure 2: Visual and audio feature concatenation before and after LSTM encoder.

**Sharing Weights across Visual-Audio Modalities** Although concatenation is effective for visual and audio feature fusion, it can not capture the resonance information across them well. To handle this problem, we propose a multimodal LSTM via sharing weights across visual and audio modalities for video caption. Framework of this multimodal LSTM encoder is shown in Figure 3 (a) and formulated as:

$$i_t^s = \sigma(W_i^s x_{t-1}^s + U_i h_{t-1}^s + b_i^s), s = 0, 1 \quad (1)$$

$$f_t^s = \sigma(W_f^s x_{t-1}^s + U_f h_{t-1}^s + b_f^s), s = 0, 1 \quad (2)$$

$$o_t^s = \sigma(W_o^s x_{t-1}^s + U_o h_{t-1}^s + b_o^s), s = 0, 1 \quad (3)$$

$$\tilde{c}_t^s = \tanh(W_c^s x_{t-1}^s + U_c h_{t-1}^s + b_c^s), s = 0, 1 \quad (4)$$

$$c_t^s = f_t^s \odot c_{t-1}^s + i_t^s \odot \tilde{c}_t^s, s = 0, 1 \quad (5)$$

$$h_t^s = o_t^s \odot c_t^s, s = 0, 1 \quad (6)$$

where  $i_t$ ,  $f_t$ ,  $o_t$  and  $\tilde{c}_t$  are the input gate, forget gate, output gate and the updated memory content separately, the superscript  $s$  indexes visual and audio input sequences respectively. When  $s = 0$ , (1)-(6) denotes the LSTM-based visual feature encoder and  $x_t^s$  is the visual feature extracted by CNNs. When  $s = 1$ , (1)-(6) indicates the LSTM-based audio feature encoder and  $x_t^s$  is the audio MFCC (mel-frequency cepstral coefficients) feature. In addition,  $W^s (s = 0, 1)$  are weight matrices for inputting visual and audio features respectively.  $U$  is the weight matrix shared by hidden states of visual and audio encoders.  $b^s (s = 0, 1)$  are the corresponding biases.

**Sharing Memory between Visual-Audio Modalities** To see whether long temporal dependency across visual and audio modalities is beneficial to video caption, we first build memory across visual and audio modalities. Concretely, an external memory is attached between visual and audio LSTM encoders. Framework of this multimodal memory encoder is presented in Figure 3 (b).

Basic procedures of our multimodal memory interactions between visual and audio modalities can be realized through the following steps: (1) read information from external memory. (2) fuse information from external memory into internal memories of visual and audio encoder LSTMs respectively. (3) update external memory.

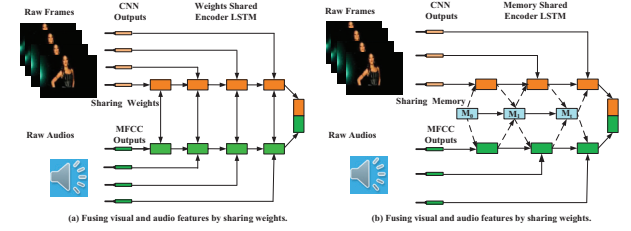


Figure 3: Visual and audio feature fusion via sharing weights and memory.

**a. External Memory** External memory adopted in our paper is defined as a matrix  $M \in R^{K \times D}$ , where  $K$  is the number of memory elements, and  $D$  is the dimension of each memory element.

At each time step  $t$ , an output  $h_t^s$  and three vectors, including key value  $key_t^s$ , erase vector  $e_t^s$  and add vector  $a_t^s$  are simultaneously emitted by visual and audio LSTM encoders respectively. They can be computed by

$$\begin{bmatrix} key_t^s \\ e_t^s \\ a_t^s \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \tanh \end{bmatrix} (W_e^s h_t^s + b_e^s), s = 0, 1 \quad (7)$$

where  $W_e^s, b_e^s (s = 0, 1)$  are the weights and bias for corresponding terms respectively.

**b. Reading Information from External Memory** We define the procedure as:

$$r_t^s = \alpha_t^s M_{t-1}, s = 0, 1 \quad (8)$$

where superscript  $s$  indexes the visual and audio input sequences,  $r_t^s \in R^D (s = 0, 1)$  indicate reading vectors for visual or audio streams respectively,  $\alpha_t \in R^K$  denotes attention distribution over the elements of memory  $M_{t-1}$ , which decides how much information will be read from the external memory.

Each element  $\alpha_{t,k}$  in  $\alpha_t$  can be obtained via the following calculation:

$$\alpha_{t,k}^s = \text{softmax}(g(M_{t-1,k}, key_{t-1}^s)), s = 0, 1 \quad (9)$$

where  $g(\cdot)$  is the similarity measure function, which is utilized to calculate the similarity between each element of memory and the key value  $key_t^s$  at time  $t$ . Here, we apply cosine similarity metric function.

**c. Fusing Information of External and Internal Memories** After we obtain information from external memory  $r_t$ , deep fusion strategy proposed in paper (Liu, Qiu, and Huang 2016) is utilized to comprehensively integrate  $r_t$  into

internal memories of visual and audio LSTMs respectively. In detail, states  $h_t^s$  of visual and audio LSTM encoders at step  $t$  conditioned not only on internal memory  $c_t^s$ , but also on information  $r_t^s$  reading from external memory, which can be computed via

$$h_t^s = o_t^s \odot (c_t^s + g_t^s \odot (W_l^s r_t^s)), s = 0, 1 \quad (10)$$

where  $W_l$  denotes the parameter matrix,  $g_t$  indicates the fusion gate, which controls how much information will flow from external memory into fused memory and can be obtained via

$$g_t^s = \sigma(W_p^s c_t^s + W_q^s r_t^s), s = 0, 1 \quad (11)$$

where  $W_p$  and  $W_q$  are the corresponding parameter matrices.

**d. Updating Memory** Memory is updated through the following procedures:

$$M_t^0 = M_{t-1}^0 [1 - \alpha_t^0 e_t^0] + \alpha_t^0 a_t^0 \quad (12)$$

$$M_t^1 = M_{t-1}^1 [1 - \alpha_t^1 e_t^1] + \alpha_t^1 a_t^1 \quad (13)$$

$$M_t = P M_t^0 + (1 - P) M_t^1 \quad (14)$$

where  $e_t^0/e_t^1$  and  $a_t^0/a_t^1$  are the erase and add vectors emitted by visual/audio encoder respectively. Final updating of memory is the combination of updated memory from visual and audio streams respectively. Parameter  $P$  is tuned on the validation set.

### Audio Modality Inference Framework

To still get benefits from audio features even when this modality is absent, we develop an audio modality inference framework (AMIN). AMIN is presented in Figure 4 (a).

AMIN can be formulated as follows:

$$\hat{y} = D_{AMIN}(E_{AMIN}(x, \theta), \vartheta) \quad (15)$$

where  $x$  indicates the visual feature and  $\hat{y}$  denotes the generated corresponding audio feature.  $E_{AMIN}/D_{AMIN}$  demonstrates the encoding/decoding function and  $\theta/\vartheta$  is the parameter set of the encodering/deconding stage.

We utilize  $\ell_2$  constraint as the training loss for AMIN model, which is dubbed as  $\mathcal{L}_{AMIN}$  and formulated:

$$\mathcal{L}_{AMIN} = \|y - \hat{y}\|^2 \quad (16)$$

where  $y$  is the ground truth audio MFCC feature.

### Dynamic Feature Fusion Framework

When AMIN is trained well, a dynamic feature fusion framework can be obtained by combining AMIN with our proposed feature fusion strategies, which is presented in Figure 4 (b).

Concerning videos which own both visual and audio sequences, they can be directly sent to multimodal feature fusion framework perform video caption (solid arrows). If offered videos have only visual sequence, AMIN model is adopted to generate audio features based on the corresponding video clip, then the visual and generated audio features are sent to multimodal feature fusion framework to perform video caption (dotted arrows).

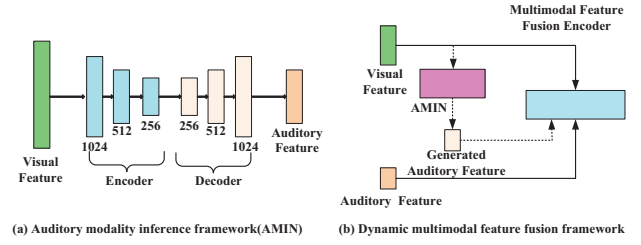


Figure 4: Dynamic multimodal feature fusion framework.

### Training and Inference

Assume the number of training video caption pairs  $(x^i, y^i)$  are  $N$  and the length of caption  $y^i$  is  $l_i$ , the averaged log-likelihood over the whole training dataset integrates a regularization term is treated as our objective function.

$$\mathcal{L}_{(\theta)} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{l_i} \log \rho(y_j^i | y_{1:j-1}^i, x^i, \theta) + \lambda \|\theta\|_2^2 \quad (17)$$

where  $y_j^i$  is adopted to represent the input word,  $\lambda$  indicates the regularization coefficient and  $\theta$  denotes all parameters needed to be optimized in the model.

Just as most LSTM language models, a softmax layer is employed to model the probability distribution over the whole vocabulary of the next generated word.

$$m_t = \tanh(W_x X_t + W_h h_t + W_y y_{t-1} + b_h) \quad (18)$$

$$\eta_t = \text{softmax}(U_p m_t + b_\eta) \quad (19)$$

where  $W_x, W_h, W_y, U_p, b_h$  and  $b_\eta$  are the parameters needed to be optimized. Depending on the probability distribution  $\eta_t$ , word sequence  $y_t$  can be recursively sampled until encountering the end of symbol in the vocabulary.

Concerning caption generation, a beam search strategy is chosen to generate word sequence (Yu et al. 2016).

## Experiments

### Data Representation and Video Caption Datasets

For a given video, we first sample it with a fixed number of frames/clips, then the pre-trained 3D CNNs are utilized to extract frame features. Meanwhile, MFCC features of each audio clip are extracted. Visual and audio feature representations can be denoted as  $x^s = \{x_1, x_2, \dots, x_n\}$ ,  $s = 0, 1$ , where  $s$  indexes visual and audio feature representations respectively and  $n$  is the number of sampled frames/clips.

To validate the performance of our model, we utilize the Microsoft Research-Video to Text Dataset (MSR-VTT) and Microsoft Video Description Dataset (MSVD) (Chen and Dolan 2011). Their split method can be found in (Xu et al. 2016) and (Yao et al. 2015) respectively.



Table 1: Comparison results of different models for video caption with C3D frame features.

| Models         | B@3          | B@4          | METEOR       |
|----------------|--------------|--------------|--------------|
| M <sup>3</sup> | 0.472        | 0.351        | 0.257        |
| audio          | 0.370        | 0.268        | 0.192        |
| Visual         | 0.473        | 0.363        | 0.257        |
| V-CatL-A       | 0.480        | 0.369        | 0.258        |
| V-CatH-A       | 0.485        | 0.374        | 0.258        |
| V-ShaMem-A     | 0.493        | 0.375        | 0.259        |
| V-ShaWei-A     | <b>0.494</b> | <b>0.383</b> | <b>0.261</b> |

## Experimental Setup

During model training, start and end tags are added to each sentence respectively. Words that not existed in vocabulary are replaced with UNK token. Furthermore, masks are added to sentences, visual and auditory features separately for better batch training. Parameters are set as follows, beam search size, word embedding dimension and LSTM hidden state dimension are 5, 468 and 512 respectively. Size of visual-auditory and visual-textual shared memories are  $64 \times 128$  and  $128 \times 512$  respectively. To avoid overfitting, dropout (Srivastava et al. 2014) with 0.5 rate are utilized on both the output of fully connected layer and the output layers of LSTM, but not on the intermediate recurrent transitions. In addition, gradients are clipped into range  $[-10, 10]$  to prevent gradient explosion. Optimization algorithm utilized for our deep feature fusion frameworks is ADADELTA (Zeiler 2012).

Concerning auditory modality supplemental network, it contains 3 fully connected layers for encoder and decoder respectively. Units' numbers of encoder hidden layers are 1024, 512 and 256 separately and 256, 512, 1024 for decoder hidden layers.

## Evaluation of Multimodal Feature Fusion Models

**Evaluation the Performance of Various Multimodal Feature Fusion Frameworks** To validate the effectiveness of integrating audio modality into video caption framework, we develop several feature fusion models and denote them as follows: V-CatL-A/V-CatH-A: Concatenating features of visual and audio modalities before/after encoder LSTM. V-ShaWei-A/V-ShaMem-A: Sharing weights/memeory across visual and audio modalities during encoding stage.

We compare our feature fusion models with several video caption models, including M<sup>3</sup> (Wang et al. 2016), Visual model (our basic video caption model), Audio model (our basic video caption model with audio features instead of visual features) respectively. Comparison results based on C3D visual features are shown in Table 1.

Table 1 reveals that performances of our visual and audio feature fusion models, including V-CatL-A, V-CatH-A, V-ShaMem-A and V-ShaWei-A models, are uniformly better than those of models which only conditioned on visual or audio features. Moreover, performance of V-CatH-A is better than that of V-CatL-A, which indicates concatenating visual and audio features in higher layer is more

efficient than that in low layer. In addition, results of V-ShaMem-A and V-ShaWei-A models are superior to those of V-CatL-A and V-CatH-A models, which hints the temporal dependency across visual and audio modalities can further boost the performance of video caption. Moreover, performances of V-ShaWei-A model surpass those of V-ShaMem-A model, demonstrating short temporal dependency is more efficient. It may be because that short temporal dependency can capture the resonance information among visual and audio modalities more efficiently.

Our best model can make a great improvement over M<sup>3</sup> by  $\frac{38.3-35.1}{35.1} = 9.1\%$  in BLUE@4 score and by  $\frac{26.1-25.7}{25.7} = 1.5\%$  in METEOR score based on C3D feature.

**Evaluation of the Generated Sentences of Various Multimodal Feature Fusion Strategies** Figure 5 presents some sentences generated by different models and human-annotated ground truth based on the test set of MSR-VTT. We can see that audio model always generates wrong sentences, which may be because the absence of visual modality leads to serious information loss. On the other hand, V-ShaWei-A model can generate sentences with more related objects, actions and targets.

Concerning the first video, sentence generated by Visual model focuses more on visual cues while ignores audio information. As a result, it generates wrong content ("to a man" vs. "news"). V-Cat-A model generates accurate object "a man" and action "talking" while lossing the content "news". It is because that directly concatenating visual and audio features may lead to the collapse of information. Both V-ShaMem-A and V-ShaWei-A models can generate more related sentences with the help of audio cues. Concerning V-ShaMem-A model, it focuses more on the summation of longer period of information, which blurs the resonance among visual and audio modalities and offers a more abstract word "something". Concerning V-ShaWei-A model, it pays more attention to the event in a finer granularity which real matters, indicating it can capture the resonance information among two modalities effectively.

Concerning the second video, all models can generate the related action "swimming" and target "in the water". While only V-ShaWei-A model generates precise object ("fish" vs. "man" and "person"). Reason is that V-ShaWei-A model can capture both motion and sound sensitive object (the resonance information among visual and audio modalities), other than static object which looks like a man.

Concerning the third video, only V-ShaWei-A model generates more related action ("showing" vs. "playing"), which indicates V-ShaWei-A model can capture the nature of an action.

Concerning the forth video, V-Cat-A and V-ShaWei-A model can generate more related actions ("knocking on a wall", "using a phone") with the aid of audio information. However, V-ShaMem-A model focus more on global event and generates a sentence "lying on bed". Moreover, Visual model pays more attention on visual information and also generates description "lying on bed".

Concerning the fifth video, event happened in this video is more related to visual information. Consequently, Visual, V-



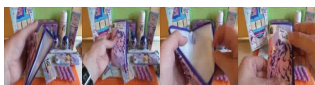
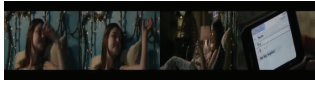

|   |   |   |  |
|---|---|---|--|
|  | Aduio: a man is <i>showing how to make a dish</i><br>Visual: a man in a suit is <i>talking to a man</i><br>V-Cat-A: a man is <i>talking</i><br>V-ShaMem-A: a man in a suit is <i>talking about something</i><br>V-ShaWei-A: a man in a suit <i>talks about the news</i> | <b>Generated Sentence:</b><br>a man is <i>showing how to make a dish</i><br>a man in a suit is <i>talking to a man</i><br>a man is <i>talking</i><br>a man in a suit is <i>talking about something</i><br>a man in a suit <i>talks about the news</i> | <b>Reference Sentence:</b><br>1. a man with a blue shirt is talking<br>2. a man talks about matters of science<br>3. a man in a suit is talking about psychological              |
|  | Aduio: a <i>man</i> is swimming<br>Visual: a <i>person</i> is swimming in the water<br>V-Cat-A: a <i>man</i> is swimming in the water<br>V-ShaMem-A: a <i>man</i> is swimming in the water<br>V-ShaWei-A: a <i>fish</i> is swimming in the water                        | <b>Generated Sentence:</b><br>a <i>man</i> is swimming<br>a <i>person</i> is swimming in the water<br>a <i>man</i> is swimming in the water<br>a <i>man</i> is swimming in the water<br>a <i>fish</i> is swimming in the water                        | <b>Reference Sentence:</b><br>1. gold fishes are swimming in the blue water<br>2. fish swimming in the fish tank<br>3. some red and white fish are swimming in a tank            |
|  | Aduio: someone is <i>playing</i> with toys<br>Visual: a person is <i>playing</i> a video game<br>V-Cat-A: a person is <i>playing</i> with toys<br>V-ShaMem-A: a person is <i>playing</i> a video game<br>V-ShaWei-A: a person is <i>showing</i> with toys               | <b>Generated Sentence:</b><br>someone is <i>playing</i> with toys<br>a person is <i>playing</i> a video game<br>a person is <i>playing</i> with toys<br>a person is <i>playing</i> a video game<br>a person is <i>showing</i> with toys               | <b>Reference Sentence:</b><br>1. person shows of disney merchandising<br>2. a person shows off a wallet<br>3. someone is showing some art  |
|  | Aduio: a person is <i>knocking</i><br>Visual: a girl is <i>laying on bed</i><br>V-Cat-A: a girl is <i>using a phone</i><br>V-ShaMem-A: a woman is <i>laying on a bed</i><br>V-ShaWei-A: a girl is <i>knocking</i> on a wall   | <b>Generated Sentence:</b><br>a person is <i>knocking</i><br>a girl is <i>laying on bed</i><br>a girl is <i>using a phone</i><br>a woman is <i>laying on a bed</i><br>a girl is <i>knocking</i> on a wall   | <b>Reference Sentence:</b><br>1. a girl in bed knocking on the wall<br>2. a girl is knocking on the wall<br>3. a girl lays in bed and uses her phone                             |
|  | Aduio: someone is <i>playing</i> a game<br>Visual: a <i>cartoon character</i> is <i>dancing</i><br>V-Cat-A: a girl is <i>singing</i><br>V-ShaMem-A: a <i>group of people</i> are <i>dancing</i><br>V-ShaWei-A: a <i>group of cartoon characters</i> are <i>dancing</i>  | <b>Generated Sentence:</b><br>someone is <i>playing</i> a game<br>a <i>cartoon character</i> is <i>dancing</i><br>a girl is <i>singing</i><br>a <i>group of people</i> are <i>dancing</i><br>a <i>group of cartoon characters</i> are <i>dancing</i>  | <b>Reference Sentence:</b><br>1. cartoon characters dance in the rain<br>2. animated characters are dancing in the rain<br>3. a bunch of anime and cartoon character are dancing |

Figure 5: Descriptions generated by Visual, audio, V-Cat-A, V-ShaMem-A, V-ShaWei-A models and human-annotated ground truth based on the test set of MSR-VTT.

ShaMem-A and V-ShaWei-A models all generate more precise actions ("dancing" vs. "playing" and "singing"). Moreover, V-ShaMem-A and V-ShaWei-A models offer more precise number of objects ("a group of" vs. "a girl", "a cartoon character" and "someone"), indicating temporal dependency across visual and audio modalities is helpful for the identification of the object. Moreover, V-ShaWei-A model provides more accurate object ("cartoon characters" vs. "people"), which validates short temporal dependency is more effective in capturing resonance information among two modalities.

## Evaluation of Dynamic Multimodal Feature Fusion

**Evaluation of Supplemental Audio Modality based on MSR-VTT** To validate whether the supplemental audio modality has comparable effects with the original one, we compare models V-ShaMem-GA (similar with V-ShaMem-A model except utilizing generated audio features instead of original audio features), V-ShaMem-Zero (similar with V-ShaMem-A model except utilizing zeros to replace audio features) and Visual model based on MSR-VTT dataset. V-ShaWei-GA, V-ShaCatH-GA, V-ShaWei-Zero, V-ShaCatH-zero share the similar meanings with corresponding terms.

Comparison results are shown in Table 2. Models with visual and generated audio features (V-CatH-GA, V-ShaMem-GA and V-ShaWei-GA), are superior to corresponding models with visual and zero filled audio features (V-CatH-Zero, V-ShaMem-Zero and V-ShaWei-Zero) and Visual model, which indicates supplemental audio features convey useful information.

Table 2: Comparison results of different models for video caption with C3D frame features based on MSR-VTT.

| Models        | B@3          | B@4          | METEOR       |
|---------------|--------------|--------------|--------------|
| Visual        | 0.473        | 0.363        | 0.257        |
| V-CatH-Zero   | 0.447        | 0.343        | 0.241        |
| V-CatH-GA     | 0.479        | 0.372        | 0.255        |
| V-ShaMem-Zero | 0.450        | 0.338        | 0.251        |
| V-ShaMem-GA   | 0.479        | 0.374        | 0.256        |
| V-ShaWei-Zero | 0.463        | 0.354        | 0.252        |
| V-ShaWei-GA   | <b>0.487</b> | <b>0.379</b> | <b>0.259</b> |

**Evaluation of Supplemental Audio Modality based on MSVD** To further verify the effectiveness of supplemental audio features, we evaluate video caption based on MSVD dataset which has no audio cues. Audio features are first generated by audio modality inference network (AMIN), then these features are fused with visual information through our multimodal feature fusion frameworks for video caption.

To validate whether supplemental audio features contain useful information or not, we compare models V-ShaWei-GA, V-ShaMem-GA with Visual model. In addition, to verify whether pretraining based on a big dataset MSR-VTT dataset will further boost the performance or not, V-ShaWei-GA-Pre, V-ShaMem-GA-Pre models (similar with V-ShaWei-GA and V-ShaMem-GA models respectively, except that before training on MSVD dataset, models are first




|   |  |   |
|---|--|---|
|  | <b>Generated Sentence:</b><br>GA: a man is <i>playing a guitar</i><br>Visual: a man is <i>playing a Piano</i><br>V-ShaWei-GA: a man is <i>playing a violin</i>                 | <b>Reference Sentence:</b><br>1. a kid is playing a violin<br>2. a boy plays a violion<br>3. a boy is playing the violin on stage   |
|  | <b>Generated Sentence:</b><br>GA: a man is <i>pouring water</i><br>Visual: the person is cooking the <i>something</i><br>V-ShaWei-GA: a man is pouring sauce into <i>a pot</i> | <b>Reference Sentence:</b><br>1. someone is pouring tomato sauce from a can into a saucepan containing meat pieces<br>2. a person pours tomato sauce in a pot<br>3. a man pours tomato sauce into a pan with meat |
|  | <b>Generated Sentence:</b><br>GA: a <i>girl</i> is riding a horse<br>Visual: a <i>man</i> is riding a horse<br>V-ShaWei-GA: a <i>girl</i> is riding a horse                    | <b>Reference Sentence:</b><br>1. a girl is horseback riding through a course<br>2. a girl is riding a horse<br>3. a woman is riding a horse   |

Figure 6: Descriptions generated by Visual, Generated audio (GA), V-ShaWei-GA models and human-annotated ground truth based on the test set of MSVD.

pretrained based on MSR-VTT) are utilized as comparisons. Comparison results are presented in Table 3.

Among Table 3, performances of V-ShaWei-GA models are better than those of Visual model ( $M^3$ , state-of-art Visual model), which again verifies that supplemental audio features carry meaningful information for video caption. In addition, models with pretraining obtain best performance, which demonstrates knowledge learned from other big dataset can further enhance our specific task.

**Evaluation of the Generated Sentences of Dynamic Multimodal Feature Fusion Framework** Figure 6 presents some sentences generated by GA (models with only generated audio features),  $M^3$  (Wang et al. 2016), V-ShaWei-GA models and human-annotated ground truth based on the test set of MSVD.

Concerning the first video, sentence generated by Visual model focuses more on visual cues. Consequently, it generates wrong content "piano", which is because the object behind the boy is like a piano and takes a large space in the image. V-ShaMem-GA model equipped with generated audio cues captures more related object "violin", which further verifies that the supplemental audio modality is useful and the V-ShaWei-GA model can capture temporal related visual and audio cues. GA model generates more similar term "guitar" than "piano", compared to the precise term "violin", which validates the effectiveness of generated audio cues.

Concerning the second video, V-ShaWei-GA model can generate more accurate action ("pouring sauce into a pot" vs. "cooking the something"), which reveals that the V-ShaWei-GA model can capture the resonance information underlying visual and audio modalities effectively. Similar with V-ShaWei-GA model, GA model also generates precise action "pouring", further demonstrating that the generated audio features is meaningful.

Concerning the third video, V-ShaWei-GA and GA model can generate more related object ("girl" vs. "man").

## Conclusions

In this paper, we propose three multimodal feature fusion strategies to integrate audio information into models for en-

Table 3: Comparison results of different models for video caption with C3D frame features based on MSVD.

| Models                   | B@3          | B@4          | METEOR       |
|--------------------------|--------------|--------------|--------------|
| $M^3$ (Wang et al. 2016) | 0.563        | 0.455        | 0.299        |
| GA                       | 0.482        | 0.381        | 0.281        |
| V-ShaMem-GA              | 0.569        | 0.467        | 0.304        |
| V-ShaMem-GA-Pre          | 0.570        | 0.471        | 0.307        |
| V-ShaWei-GA              | 0.571        | 0.468        | 0.307        |
| V-ShaWei-GA-Pre          | <b>0.584</b> | <b>0.479</b> | <b>0.309</b> |

hanced video caption. Each of these three strategies can uniformly boost the performance of video caption, which denotes the valuableness of audio cues underlying videos. In addition, fusion models via sharing weights across visual and audio modalities can well model the reason information among them and obtains the best results. Moreover, based on our multimodal feature fusion model, we propose a dynamic multimodal feature fusion framework to handle audio modality absent problem. It can generate promising audio features based on the corresponding visual features when the audio modality is missing.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61773375, No. 61375036, No. 61602481, No. 61702510), and in part by the Microsoft Collaborative Research Project.

## References

- Aradhye, H.; Toderici, G.; and Yagnik, J. 2009. Video2text: Learning to annotate video content. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, 144–151. IEEE.
- Chen, D. L., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational*



- Linguistics: Human Language Technologies-Volume 1*, 190–200. Association for Computational Linguistics.
- Dodge, J.; Gane, A.; Zhang, X.; Bordes, A.; Chopra, S.; Miller, A.; Szlam, A.; and Weston, J. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.
- Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.
- Graves, A.; Wayne, G.; and Danihelka, I. 2014. Neural Turing machines. *arXiv preprint arXiv:1410.5401*.
- Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R. J.; Saenko, K.; and Guadarrama, S. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, volume 1, 2.
- Liu, P.; Qiu, X.; and Huang, X. 2016. Deep multi-task learning with shared memory. *arXiv preprint arXiv:1609.07222*.
- Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2016. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, 801–816. Springer.
- Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; and Zhuang, Y. 2016a. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1029–1038.
- Pan, Y.; Mei, T.; Yao, T.; Li, H.; and Rui, Y. 2016b. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4594–4602.
- Ren, J. S.; Hu, Y.; Tai, Y.-W.; Wang, C.; Xu, L.; Sun, W.; and Yan, Q. 2016. Look, listen and learn-a multimodal lstm for speaker identification. In *AAAI*, 3581–3587.
- Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 433–440.
- Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3202–3212.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
- Thomason, J.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; and Mooney, R. J. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Coling*, volume 2, 9.
- Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.
- Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 4534–4542.
- Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.
- Wang, J.; Wang, W.; Huang, Y.; Wang, L.; and Tan, T. 2016. Multimodal memory modelling for video captioning. *arXiv preprint arXiv:1611.05592*.
- Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.
- Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, 2397–2406.
- Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 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*, 5288–5296.
- Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, 4507–4515.
- Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4584–4593.
- Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.