# Multimodal Attention for Fusion of Audio and Spatiotemporal Features for Video Description

Chiori Hori Takaaki Hori Gordon Wichern
Jue Wang Teng-yok Lee Anoop Cherian Tim K. Marks

Mitsubishi Electric Research Laboratories (MERL)
{chori, thori, wichern, juewang, tlee, cherian, tmarks}@merl.com

## Abstract

*Video description approaches are based on encoder-decoder sentence generation using recurrent neural networks (RNNs) by integrating temporal attention mechanisms into each model, in which the decoder network predicts each word in the description by selectively giving more weight to encoded features from specific time frames. We incorporate audio features, in addition to image and motion features, for video description based on encoder-decoder recurrent neural networks (RNNs). To fuse these modalities, we introduce a multimodal attention model that can selectively utilize features from different modalities for each word in the output description. We apply our new framework for video description using MFCC and state-of-the-art audio features such as SoundNet. Results confirm that our attention-based multimodal fusion of audio features with visual features outperforms conventional video description approaches.*

## 1. Introduction

Recent work in video description has demonstrated the advantages of integrating temporal attention mechanisms into encoder-decoder neural networks, in which the decoder network predicts each word in the description by selectively giving more weight to encoded features from different times in the video. Typically, two different types of features, image features and motion features, are combined by naïve concatenation [14]. We recently proposed Attention-based Multimodal Fusion, in which we introduced a new use of attention: fusing information across different modalities [6, 7] (as well as over time). We use modality loosely to refer to different types of features, such as appearance (image), spatiotemporal, and audio features derived from the video, such as appearance, motion, or depth, as well as features from different sensors such as video and audio features.

A longstanding area of research addresses how to effectively combine information from multiple modalities for machine perception tasks [4]. As far as we know, our approach is the first to fuse multimodal information using attention across modalities in a neural network. Our method dynamically adjusts the relative importance of each modality to generate better descriptions. The benefits of our proposed multimodal attention include: (1) the modalities that are most helpful to discriminate each word in the description can dynamically receive a stronger weight, and (2) the network can detect interference (e.g., noise) and other sources of uncertainty in each modality and dynamically down-weight the modalities that are less certain. The multimodal attention mechanism for video description provides a means for introspection in the model, in the sense that the weights across modalities that are used in generating each word can be used to explore what features are useful in various contexts.

This work include expanding the feature set for video description to include audio (in addition to image and spatiotemporal features), and introducing a mechanism for selectively attending to different modalities. In this paper, we report the works in [6] and [7].

## 2. Attention-Based Multimodal Fusion

In [14], content vectors from VGG-16 (image features) and C3D (spatiotemporal motion features) are combined into one vector, which is used to predict the next word. This is performed in the fusion layer, in which the following activation vector is computed :

$$g_i = \tanh\left(W_s^{(\lambda_D)} s_{i-1} + d_i + b_s^{(\lambda_D)}\right), \quad (1)$$

where

$$d_i = W_{c1}^{(\lambda_D)} c_{1,i} + W_{c2}^{(\lambda_D)} c_{2,i}, \quad (2)$$

$c_{1,i}$ and $c_{2,i}$ are two feature vectors obtained using different input modalities, $s_i$ is the decoder state after word $y_i$ has been output, and $W_s^{(\lambda_D)}$ and $b_s^{(\lambda_D)}$ are respectively the weight matrix and bias for the decoder network $\lambda_D$.

Figure 1 illustrates this approach, which we call this approach Naïve Fusion, in which multimodal feature vectors are combined using one projection matrix $W_{c1}$ for the first modality (input sequence $x_{11}, \ldots, x_{1L}$), and a different projection matrix $W_{c2}$ for the second modality (input sequence $x'_{21}, \ldots, x_{2L'}$).

Our proposed method extends the attention mechanism to multimodal fusion. We call it *attentional fusion*, or *multimodal attention*. In our model, based on the current decoder state, the decoder network can selectively attend to specific modalities of input (or specific feature types) to predict the next word. Let $K$ be the number of modalities, i.e., the number of sequences of input feature vectors. Our attention-based feature fusion is performed using

$$g_i = \tanh\left(W_s^{(\lambda_D)} s_{i-1} + \sum_{k=1}^{K} \beta_{k,i} d_{k,i} + b_s^{(\lambda_D)}\right), \quad (3)$$

where

$$d_{k,i} = W_{ck}^{(\lambda_D)} c_{k,i} + b_{ck}^{(\lambda_D)}. \quad (4)$$

The multimodal attention weights $\beta_{k,i}$ are obtained in a similar way to the temporal attention mechanism:

$$\beta_{k,i} = \frac{\exp(v_{k,i})}{\sum_{\kappa=1}^{K} \exp(v_{\kappa,i})}, \quad (5)$$

where

$$v_{k,i} = w_B^{\mathsf{T}} \tanh(W_B s_{i-1} + V_{Bk} c_{k,i} + b_{Bk}). \quad (6)$$

Here $W_B$ and $V_{Bk}$ are matrices, $w_B$ and $b_{Bk}$ are vectors, and $v_{k,i}$ is a scalar. Unlike in Naïve multimodal fusion, the multimodal attention weights can change according to the decoder state and the feature vectors (shown in Figure 2). This enables the decoder network to attend to a different set of features and/or modalities when predicting each subsequent word in the description. Naïve fusion can be considered a special case of Attentional fusion, in which all modality attention weights, $\beta_{k,i}$, are constantly 1.

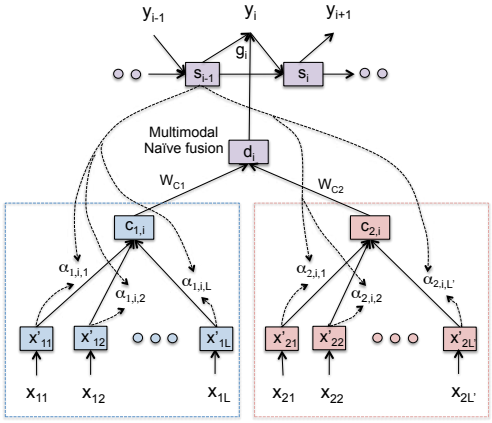


Figure 1. Naïve Fusion of multimodal features.

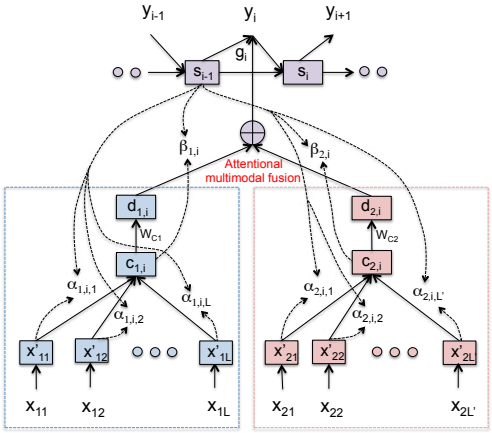


Figure 2. Our Attentional Fusion of multimodal features.

# 3. Experiments

## 3.1. Datasets

We evaluated our proposed feature fusion using the MSVD (YouTube2Text) [3] and MSR-VTT [12]. MSVD (YouTube2Text) has 1,970 video clips with multiple natural language descriptions. There are 80,839 sentences in total, with about 41 annotated sentences per clip. Each sentence on average contains about 8 words. The words contained in all the sentences constitute a vocabulary of 13,010 unique lexical entries. The dataset is open-domain and covers a wide range of topics including sports, animals, and music. Following [3], we split the dataset into a training set of 1,200 video clips, a validation set of 100 clips, and a test set of the remaining 670 clips.

MSR-VTT [12] consists of 10,000 web video clips with 41.2 hours and 200,000 clip-sentence pairs in total, covering a comprehensive list of 20 categories and a wide variety of video content. Each clip was annotated with about 20 natural sentences. The dataset is split into training, validation, and testing sets of 65%, 5%, 30%, corresponding to 6,513, 497, and 2,990 clips respectively. However, because the video clips are hosted on YouTube, some of the MSR-VTT videos have been removed due to content or copyright issues. When we downloaded the videos (February 2017), approximately 12% were unavailable. Thus, we trained and tested our approach using just the subset of the MSR-VTT dataset that was available, which consists of 5,763, 419, and 2,616 clips for train, validation, and test respectively.

## 3.2. Video Processing

The image data are extracted from each video clip at 24 frames per second and rescaled to 224×224-pixel images. In our experiments, we use a temporal stride of 16. For extracting image features, we use a VGG-16 network [10]

that was pretrained on the ImageNet dataset [8]. The output of the fully-connected fc7 layer of a VGG-16 network [10] pretrained on the Imagenet dataset is used for the image features, which produces a sequence of 4096-dimensional feature vectors. To model motion and short-term spatio-temporal activity, we use the pretrained C3D [11] model. The C3D network reads sequential frames in the video and outputs a fixed-length feature vector every 16 frames. We extracted activation vectors from fully-connected fc6-1 layer, which has 4096-dimensional features.

### 3.3. Audio Processing

The previous methods that used the MSVD (YouTube2Text) dataset did not use audio features [13, 9, 14]. In this paper, we additionally incorporate audio features. Since the packaged MSVD (YouTube2Text) dataset does not include the audio track from the YouTube videos, we extracted the audio data via the original video URLs. Although some of the videos were no longer available on YouTube, we were able to collect audio data for 1,649 video clips, which covers 84% of the dataset. The 44 kHz-sampled audio data are downsampled to 16 kHz, and mel-frequency cepstral coefficients (MFCCs) are extracted from each 50 ms time window with 25 ms shift. The sequence of 13-dimensional MFCC features are then concatenated into one vector for every group of 20 consecutive frames, which results in a sequence of 260-dimensional vectors. The MFCC features are normalized so that the mean and variance vectors are $\mathbf{0}$ and $\mathbf{1}$ in the training set. The validation and test sets are also adjusted using the original mean and variance vectors from the training set. Unlike for the image features, we apply a BLSTM encoder network to the MFCC features, which is trained jointly with the decoder network. If audio data were not available for a video clip, then we feed in a sequence of dummy MFCC features (zero vectors).

We also extracted SoundNet features using a pre-trained CNN [1]. We extracted 1024-dimensional feature vectors (using fully connected layer conv7) from each video's audio track. Unlike for MFCC features, we do not apply a BLSTM encoder for SoundNet features.

### 3.4. Experimental Setup

The caption generation model, i.e., the decoder network, is trained to minimize the cross entropy criterion using the training set. Image features and deep audio features (SoundNet ) are fed to the decoder network through one projection layer of 512 units, while MFCC audio features are fed to a BLSTM encoder (one projection layer of 512 units and bidirectional LSTM layers of 512 cells) followed by the decoder network. The decoder network has one LSTM layer with 512 cells. Each word is embedded to a 256-dimensional vector when it is fed to the LSTM layer. In this video description task, we used L2 regularization for all experimental conditions and used RMSprop optimization.

## 4. Results and Discussion

Tables 1 and 2 show the evaluation results on the MSVD (YouTube2Text) and MSR-VTT Subset. On each dataset, we compare the performance of unimodal systems to that of Attentional multimodal fusion systems. Unimodal system results show that image-only and motion-only features provide significantly better BLEU4 and METEOR scores than audio-only features. Since video description mainly relies on objects and background scene in the video, it seems to be difficult to generate appropriate descriptions only using audio features. Furthermore, some YouTube videos include unrelated sound that was not in the original scene, such as overdubbed music that was added to the video in post-production, and some video clips have no audio track. In such cases, it is almost impossible to generate related sentences.

However, by performing Attentional fusion of audio features (MFCC, SoundNet) along with the image and motion features, both BLEU4 and METEOR scores improved over unimodal systems and over multimoal systems based only on image and motion features. This result demonstrates that audio features are useful for video description when they are used as additional information.

## 5. Conclusion

We proposed a new modality-dependent attention mechanism, which we call multimodal attention, for video description based on encoder-decoder sentence generation using recurrent neural networks (RNNs). In this approach, the attention model selectively attends not just to specific times, but to specific modalities of input such as image features, spatiotemporal motion features, and audio features. This approach provides a natural way to fuse multimodal information for video description. In addition, Attentional Fusion enables us to analyze the attention weights for each word to examine how each modality contributes to each word. We evaluated our attention-based multimodal fusion method on the MSVD (YouTube2Text) and MSR-VTT. achieving results that are competitive with current state-of-the-art methods that employ temporal attention models.

Future work consists of incorporating a new state-of-the-art audio features derived from the audio tracks of videos, known as Audio Set VGGish [5], as well as a new state-of-the-art video feature developed for action recognition, I3D [2], into our framework for multimodal (and temporal) attention.

Table 1. Results of feature integration on MSVD (YouTube2Text) dataset

| | feature type | | | | Evaluation metric | |
|---|---|---|---|---|---|---|
| | Image | Motion | Audio | | BLEU4 | METEOR |
| Unimodal systems | VGG-16 | | | | 0.464 | 0.309 |
| | | C3D | | | 0.464 | 0.304 |
| | | | MFCC | | 0.267 | 0.228 |
| | | | | SoundNet | 0.216 | 0.177 |
| Attentional fusion | VGG-16 | C3D | | | 0.507 | 0.318 |
| | VGG-16 | C3D | MFCC | | 0.517 | **0.320** |
| | VGG-16 | C3D | | SoundNet | 0.517 | 0.315 |
| | VGG-16 | C3D | MFCC | SoundNet | **0.519** | 0.312 |

Table 2. Results of feature integration on MSR-VTT data set

| | feature type | | | Evaluation metric | |
|---|---|---|---|---|---|
| | Image | Motion | Audio | BLEU4 | METEOR |
| Single features | VGG-16 | | | 0.361 | 0.244 |
| | | C3D | | 0.362 | 0.246 |
| | | | MFCC | 0.248 | 0.209 |
| | | | SoundNet | 0.218 | 0.198 |
| Attentional fusion | VGG-16 | C3D | | 0.394 | 0.257 |
| | VGG-16 | C3D | MFCC | **0.397** | **0.258** |
| | VGG-16 | C3D | SoundNet | 0.395 | 0.253 |
| | VGG-16 | C3D | MFCC & SoundNet | 0.390 | 0.254 |

## References


[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 3

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3

[3] 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 *ICCV*, 2013. 2

[4] M. E. Hennecke, D. G. Stork, and K. V. Prasad. Visionary speech: Looking ahead to practical speechreading systems. In *Speechreading by Humans and Machines*. 1996. 1

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. CNN architectures for large-scale audio classification. In *ICASSP*, 2017. 3

[6] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *ICCV*, 2017. 1

[7] C. Hori, T. Hori, T. K. Marks, and J. R. Hershey. Early and late integration of audio features for automatic video description. In *ASRU*, 2017. 1

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 3

[9] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. *CoRR*, abs/1505.01861, 2015. 3

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2, 3

[11] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3

[12] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2

[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville. Describing videos by exploiting temporal structure. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4507–4515, 2015. 3

[14] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. *CoRR*, abs/1510.07712, 2015. 1, 3