Watch, Listen and Tell: Multi-modal Weakly Supervised Dense Event Captioning

Tanzila Rahman[1,2]   Bicheng Xu[1,2]   Leonid Sigal[1,2,3]
[1]University of British Columbia   [2]Vector Institute for AI   [3]Canada CIFAR AI Chair
{trahman8, bichengx, lsigal}@cs.ubc.ca

Abstract

Multi-modal learning, particularly among imaging and linguistic modalities, has made amazing strides in many high-level fundamental visual understanding problems, ranging from language grounding to dense event captioning. However, much of the research has been limited to approaches that either do not take audio corresponding to video into account at all, or those that model the audio-visual correlations in service of sound or sound source localization. In this paper, we present the evidence, that audio signals can carry surprising amount of information when it comes to high-level visual-lingual tasks. Specifically, we focus on the problem of weakly-supervised dense event captioning in videos and show that audio on its own can nearly rival performance of a state-of-the-art visual model and, combined with video, can improve on the state-of-the-art performance. Extensive experiments on the ActivityNet Captions dataset show that our proposed multi-modal approach outperforms state-of-the-art unimodal methods, as well as validate specific feature representation and architecture design choices.

1. Introduction

Humans often perceive the world through multiple sensory modalities, such as watching, listening, smelling, touching, and tasting. Consider two people sitting in a restaurant; seeing them across the table suggests that they maybe friends or coincidental companions; hearing, even the coarse demeanor of their conversation, makes the nature of their relationship much clearer. In our daily life, there are many other examples that produce strong evidence that multi-modal co-occurrences give us fuller perception of events. Recall how difficult it is to perceive the intricacies of the story from a *silent* film. Multi-modal perception has been widely studied in areas like psychology [10, 42], neurology [33], and human computer interaction [37].

In the computer vision community, however, the progress in learning representations from multiple modalities has been limited, especially for high-level perceptual tasks where such modalities (*e.g.*, audio or sound)
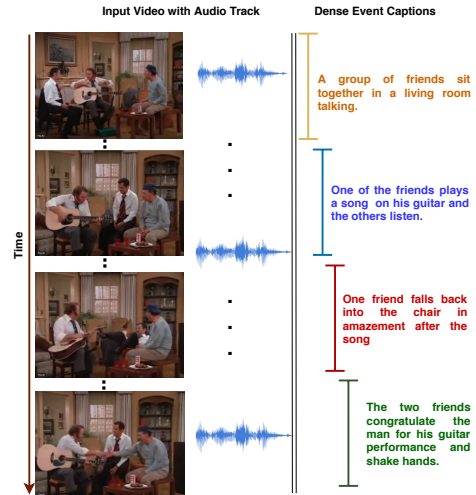

Figure 1: **Multi-modal Dense Event Captioning.** Illustration of our problem definition, where we use both audio features and visual information to generate the dense captions for a video in a weakly supervised manner.

can play an integral role. Recent works [27, 31] propose approaches for localizing audio in unconstrained videos (sound source localization) or utilize sound in video captioning [15, 16, 44, 38]. However, these approaches consider relatively short videos, *i.e.*, usually about 20 seconds, and focus on description of a single salient event [47]. More importantly, while they show that audio can boost the performance of visual models to an extent, such improvements are typically considered marginal and the role of audio is delegated to being secondary (or not nearly as important) as visual signal [16, 44].

We posit that sound (or audio) may in fact be much more important than the community may realize. Consider the previously mentioned example of a silent film. The lack of sound makes it significantly more difficult, if not impossible in many cases, to describe the rich flow of the story and constituent events. Armed with this intuition, we focus on dense event captioning [22, 43, 49] (*a.k.a.* dense-captioning of events in videos [20]) and endow our models with ability to utilize rich auditory signals for both event lo-

calization and captioning. Figure 1 illustrates one example of our multi-modal dense event captioning task. Compared with conventional video captioning, dense event captioning deals with longer and more complex video sequences, usually 2 minutes or more. To the best of our knowledge, our work is the first to tackle the dense event captioning with sound, treating sound as a first class perceptual modality.

Audio features can be represented in many different ways. Choosing the most appropriate representation for our task is challenging. To this end, we compare different audio feature representations in this work. Importantly, we show that audio signal alone can achieve impressive performance on the dense event captioning task (rivalling visual counterpart). The form of fusion needed to incorporate the audio with the video signal is another challenge. We consider and compare a variety of fusion strategies.

Dense event captioning provides detailed descriptions for videos, which is beneficial for in-depth video analysis. However, training a fully supervised model requires both caption annotations and corresponding temporal segment coordinates (*i.e.*, the start and end time of each event), which is extremely difficult and time consuming to collect. Recently, [12] proposes a method for dense event captioning in a weakly supervised setting. The approach does not require temporal segment annotation during training. During evaluation, the model is able to detect all events of interest and generate their corresponding captions. Inspired by and building on [12], we tackle our multi-modal dense event captioning in a weakly supervised manner.

Contributions. Our contributions are multiple fold. First, to the best of our knowledge, this is the first work that addresses dense event captioning task in a multi-modal setting. In doing so, we propose an attention-based multi-modal fusion model to integrate both audio and video information. Second, we compare different audio feature extraction techniques [4, 11, 23], and analyze their suitability for the task. Third, we discuss and test different fusion strategies for incorporating audio cues with visual features. Finally, extensive experiments on the ActivityNet Captions dataset [20] show that audio model, on its own, can nearly rival performance of a visual model and, combined with video, using our multi-modal weakly-supervised approach, can improve on the state-of-the-art performance.

2. Related Work

Audio Feature Representations. Recently computer vision community has begun to explore audio features for learning good representations in unconstrained videos. Aytar *et al.* [4] propose a sound network guided by a visual teacher to learn the representations for sound. Earlier works, [27, 31, 35], address sound source localization problem to identify which pixels or regions are responsible for

generating a specified sound in videos (sound grounding). For example, [31] introduces an attention based localization network guided by sound information. A joint representation between audio and visual networks is presented in [27, 35] to localize sound source. Gao *et al.* [14] formulate a new problem of audio source separation using a multi-instance multi-label learning framework. This framework maps audio bases, extracted by non-negative matrix factorization (NMF), to the detected visual objects. In recent year, audio event detection (AED) [8, 29, 36] has received attention in the research community. Most of the AED methods locate audio events and then classify each event.

Multi-modal Features in Video Analysis. Combining audio with visual features (*i.e.*, multi-modal representation) often boosts performance of networks in vision, especially in video analysis [2, 3, 16, 38, 44]. Ariav *et al.* [3] propose an end-to-end deep neural network to detect voice activity by incorporating audio and visual modalities. Features from both modalities are fused using multi-modal compact bilinear pooling (MCB) to generate a joint representation for speech signal. Authors in [2] propose a multi-modal method for egocentric activity recognition where audio-visual features are combined with multi-kernel learning and boosting.

Recently, multi-modal approaches are also gaining popularity for video captioning [38, 44]. In [16] a multi-modal attention mechanism to fuse information across different modalities is proposed. Hori *et al.* [17] extend the work in [16] by applying hypothesis-level integration based on minimum Bayes-risk decoding [21, 34] to improve the caption quality. Hao *et al.* [15] present multi-modal feature fusion strategies to maximize the benefits of visual-audio resonance information. Wang *et al.* [44] introduce a hierarchical encoder-decoder network to adaptively learn the attentive representations of multiple modalities, and fuse both global and local contexts of each modality for video understanding and sentence generation. A module for exploring modality selection during sentence generation is proposed in [38] with the aim to interpret how words in the generated sentences are associated with audio and visual modalities.

Dense Event Captioning in Videos. The task of dense event captioning in videos was first introduced in [20]. The task involves detecting multiple events that occur in a video and describing each event using natural language. Most of the works [26, 48] solve this problem in a two-stage manner, *i.e.*, first temporal event proposal generation and then sentence captioning for each of the proposed event segments. In [48], authors adopt a temporal action proposal network to localize proposals of interest in videos, and then generate descriptions for each proposal. Wang *et al.* [43] present a bidirectional proposal method that effectively exploits both past and future contexts to make proposal predictions. In [49], a differentiable masking scheme is used to ensure the consistency between proposal and captioning

modules. Li *et al.* [22] propose a descriptiveness regression component to unify the event localization and sentence generation. Xu *et al.* [46] present an end-to-end joint event detection and description network (JEDDi-Net) which adopts region convolutional 3D network [45] for proposal generation and refinement, and proposes hierarchical captioning.

Duan *et al.* [12] formulate the dense event captioning task in a weakly supervised setting, where there is no ground-truth temporal segment annotations during training and evaluation. They decompose the task into a pair of dual problems, event captioning and sentence localization, and present an iterative approach for training. Our work is motivated by [12] and builds on their framework. However, importantly, we fuse audio and visual features, and explore a variety of fusion mechanisms to address the multi-modal weakly supervised dense event captioning task. We note that [12] is thus far the only method for dense event captioning in the weakly supervised setting.

3. Multi-modal Dense Event Captioning

In this work, we consider two important modalities, audio and video, to generate dense captions in a weakly supervised setting. Weak supervision means that we do not require ground-truth temporal event segments during training. The overview of our multi-modal architecture is shown in Figure 2. The architecture consists of two modules, a *sentence localizer* and a *caption generator*. Given a set of initial random proposal segments in a video, caption generator produces captions for the specified segments. Sentence localizer then refines the corresponding segments with the generated captions. Caption generator is employed again to refine the captions. This process can proceed iteratively to arrive at consistent segments and captions; in practice we use one iteration following the observations in [12].

We extract features from audio, video, and captions first, and pass them as inputs to the sentence localizer during training. For each modality, an encoder is used to encode the input. We use recurrent neural networks (RNNs) with GRU [9] units as encoders. We then apply a *crossing attention* among the audio, video and caption features. Then an *attention feature fusion* mechanism followed by a fully-connected layer is applied to produce temporal segments.

The caption generator takes the encoded features of audio and video, along with the resultant temporal segments as inputs. It performs *soft mask clipping* on the audio and video features based on the temporal segments, and uses a *context fusion* technique to generate the multi-modal context features. Then a caption decoder, which is also an RNN with GRU units, generates one caption for each multi-modal context feature. We discuss and compare three different context fusion strategies to find the most appropriate one for our multi-modal integration.

In what follows, we first describe how to extract features from audio and video in Sec. 3.1. Then we present our weakly supervised approach in Sec. 3.2. Lastly, we demonstrate three different context fusion strategies in section 3.3.

3.1. Feature Representation

We consider both features from audio and video modalities for dense event captioning. It is generally challenging to select the most appropriate feature extraction process, especially for the audio modality. We describe different feature extraction methods to process both audio and video inputs.

3.1.1 Audio Feature Processing

ActivityNet Captions dataset [20] does not provide audio tracks. As such, we collected all audio data from the YouTube videos via the original URLs. Some videos are no longer available on YouTube. In total, we were able to collect around 15,840 audio tracks corresponding to ActivityNet videos. To process the audio, we consider and compare three different audio feature representations.

MFCC Features. Mel-Frequency Cepstrum (MFC) is a common representation for sound in digital signal processing. Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC – a representation of the short-term power spectrum of sound [19]. We down-sample the audio from 44 kHz to 16 kHz and use 25 as the sampling rate. We choose 128 MFCC features, with 2048 as the FFT window size and 512 as the number of samples between successive frames (*i.e.*, hop length).

CQT Features. The Constant-Q-Transform (CQT) is a time-frequency representation where the frequency bins are geometrically spaced and the ratios of the center frequencies to bandwidths (Q-factors) of all bins are equal [7]. CQT is motivated from the human auditory system and the fundamental frequencies of the tones in Western music [30]. We perform feature extraction by choosing 64 Hz and 60 as the minimum frequency and the number of frequency bins respectively. Similar to the MFCC features described above, we use 2048 as the FFT window size and 512 as the hop length. We use VGG-16 [32] without the last classification layer to convert both MFCC and CQT features into 512-dimensional representations.

SoundNet Features. SoundNet [4] is a CNN that learns to represent raw audio waveforms. The acoustic representation is learned using two-million videos with their accompanying audios; leveraging the natural synchronization between them. We use a pretrained SoundNet [4] model to extract the 1000-dimension audio features from the 8-th convolutional layer (*i.e.*, conv8) for each video's audio track.

3.1.2 Video Feature Processing

Given an input video $\mathbf{V} = \{\mathbf{v}_t\}_{t=1}^{T_v}$, where $\mathbf{v}_t$ is the video frame at time t and T_v is the video length, a 3D-CNN model
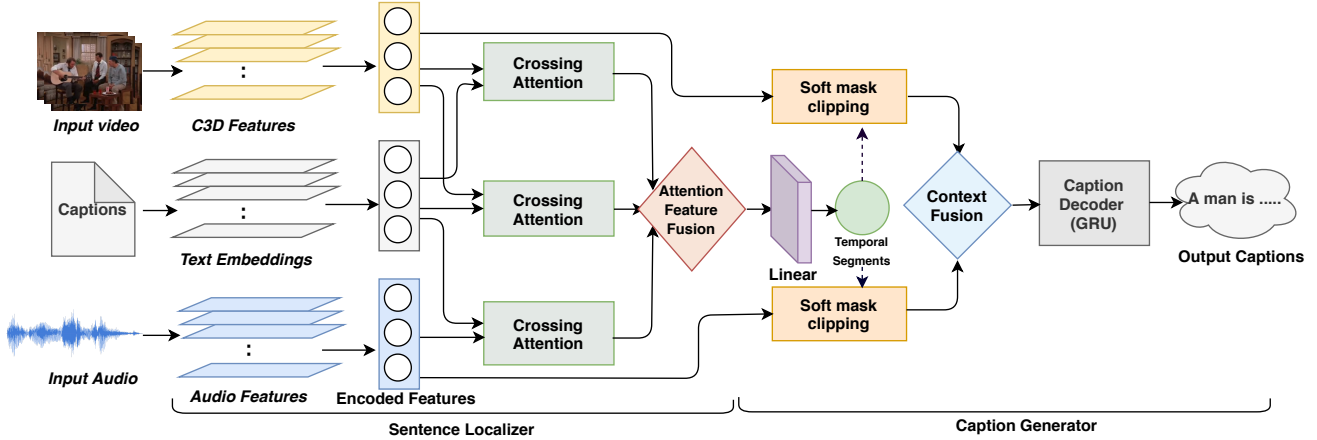
Figure 2: **Our Multi-modal Architecture.** The model has two parts, a *sentence localizer* and a *caption generator*. The *sentence localizer* takes audio, video, and captions as inputs and generates a temporal segment for each caption. The *caption generator* uses the resultant temporal segments, with audio and video features, to produce a caption for each segment.

is used to process the input video frames into a sequence of visual features $\{\mathbf{f}_t = F(\mathbf{v}_t : \mathbf{v}_{t+\delta})\}_{t=1}^{T_f}$. Here, δ means the time resolution for each feature $\mathbf{f}_t$ and T_f is the length of the feature sequence. We use features extracted from encoder F provided by the ActivityNet Captions dataset [20], where F is the pretrained C3D [18] network with $\delta = 16$ frames. The dimension of the resultant C3D features is a tensor of size $T_f \times D$, where $D = 500$ and $T_f = T_v/\delta$.

3.2. Weakly Supervised Model

Weak supervision means that we do not require ground-truth temporal alignments between the video (visual and audio collectively) and captions. We make a one-to-one correspondence assumption, meaning that we assume that each caption describes one temporal segment and each temporal segment corresponds to only one caption. This assumption holds in the current benchmark dataset and most real world scenarios. We employ two network modules, a sentence localizer and a caption generator. Given a caption, the sentence localizer will produce a temporal segment in the *context*, while the caption generator will generate a caption with a given temporal segment. We use *context* to refer an encoded video or audio.

Notations. We use GRU RNNs to encode visual and audio streams of the video. This results in a sequence of output feature vectors, one per frame, $\mathbf{O} = \{\mathbf{o}_t \in \mathbb{R}^k\}_{t=0}^{T_o}$ and the final hidden state $\mathbf{h}^o \in \mathbb{R}^k$, where T_o is the length of the video. While in practice we get two sets of such vectors (one set for video and one set for corresponding audio "frames"), we omit the subscript for clarity of formulation that follows. A caption is encoded similarly by the output features of the RNN, $\mathbf{C} = \{\mathbf{c}_t \in \mathbb{R}^k\}_{t=0}^{T_c}$ with the last hidden state being $\mathbf{h}^c \in \mathbb{R}^k$, where T_c is the length of the caption in words. We use *context* to refer the encoding of the

full visual or audio information in videos. A context segment $\mathbf{S}$ is represented by (c, l), where c and l denote segment's temporal center and length respectively within $\mathbf{O}$.

3.2.1 Sentence Localizer

Sentence localizer attempts to localize a given caption in a video by considering the caption and the encoded complete video (*context*). Formally, given a (video or audio) context $\mathbf{O}$ and an encoded caption $\mathbf{C}$, sentence localizer will regress a temporal segment $\mathbf{S}$ in $\mathbf{O}$. With the context and caption features, it first applies *crossing attention* among them. Then an *attention feature fusion*, followed by one layer fully-connected neural network, is used to generate the temporal segment. Following [10], we use 15 predefined temporal segments and generate 15 offsets in sentence localization using fully connected layer. The final segments are the sum of temporal segments and offsets value. The purpose is to fine-tune the offset value for best localization.

Crossing Attention. The crossing attention consists of two sub-attentions, one caption attention $\mathbf{Att}_c$, and one context attention $\mathbf{Att}_o$. For a context $\mathbf{O}$ and a caption $\mathbf{C}$, we first compute the attention between $\mathbf{h}^o$ and $\mathbf{C}$ as:

$$\mathbf{Att}_c = \text{softmax}((\mathbf{h}^o)^T \boldsymbol{\alpha}_c \mathbf{C})\mathbf{C}^T, \qquad (1)$$

and then calculate the attention between $\mathbf{h}^c$ and $\mathbf{O}$ as:

$$\mathbf{Att}_o = \text{softmax}((\mathbf{h}^c)^T \boldsymbol{\alpha}_o \mathbf{O})\mathbf{O}^T, \qquad (2)$$

where $\boldsymbol{\alpha}_c \in \mathbb{R}^{k \times k}$ and $\boldsymbol{\alpha}_o \in \mathbb{R}^{k \times k}$ are the learnable attention weights, and $()^T$ is the matrix transpose operation. We note that $\mathbf{Att}_o$ is a vector of size $1 \times k$ comprising of attention weighted features for the visual/audio frames; similarly $\mathbf{Att}_c$ is a vector of size $1 \times k$ of attended caption features.
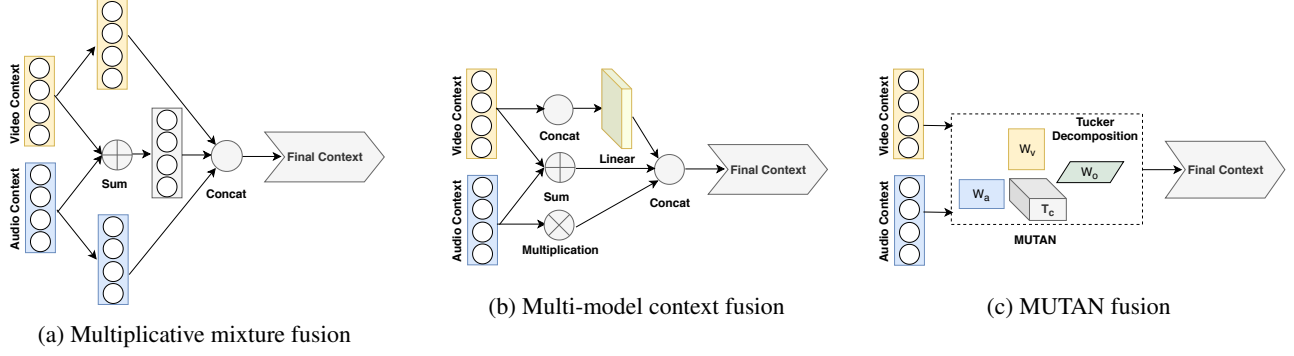
Figure 3: **Context Fusion Strategies.** Three fusion strategies are illustrated: (a) multiplicative mixture fusion, (b) multimodal context fusion, and (c) MUTAN fusion.

When training our multi-modal approaches, the caption attention $\mathbf{Att}_c$ is calculated only between the visual modality and the captions, and we generate video attention $\mathbf{Att}_v$ and audio attention $\mathbf{Att}_a$ using Eq. 2. While we are training our unimodal approaches which either use audio (or video) information to generate captions, the caption attention $\mathbf{Att}_c$ is calculated between the audio (or video) and captions.

Attention Feature Fusion. After obtaining the sub-attentions, we use the multi-model feature fusion technique [13] to fuse them together:

$$\mathbf{Att}_{sum} = \mathbf{Att}_c + \mathbf{Att}_v + \mathbf{Att}_a \qquad (3)$$

$$\mathbf{Att}_{dot} = \mathbf{Att}_c \cdot \mathbf{Att}_v \cdot \mathbf{Att}_a \qquad (4)$$

$$\mathbf{Att}_{fc} = \mathrm{fc}(\mathbf{Att}_c||\mathbf{Att}_v||\mathbf{Att}_a) \qquad (5)$$

$$\mathbf{Att}_{fusion} = \mathbf{Att}_{sum}||\mathbf{Att}_{dot}||\mathbf{Att}_{fc} \qquad (6)$$

where $+$ and $\cdot$ are the element-wise addition and multiplication, $||$ is the column-wise concatenation, and $\mathrm{fc}(\cdot)$ is a one-layer fully-connected neural network.

3.2.2 Caption Generator

Given a temporal segment $\mathbf{S}$ in a context $\mathbf{O}$, the caption generator will generate a caption based on $\mathbf{S}$. With the temporal segments generated by the sentence localizer (Sec. 3.2.1), the caption generator first applies *soft mask clipping* on the contexts, and then uses a *context fusion* mechanism (Sec. 3.3) to fuse the clipped contexts together. The fused contexts are then fed to a caption decoder, which is also a GRU RNN, to generate the corresponding captions.

Soft Mask Clipping. Getting a temporal segment $\mathbf{S}$ from a context, *i.e.*, the clipping operation, is non-differentiable, which makes it difficult to handle in end-to-end training. To this end, we utilize a continuous mask function with regard to the time step t to perform soft clipping. The mask M to obtain an $\mathbf{S}$ is defined as follows:

$$\mathrm{M}(t, \mathbf{S}) = \sigma(-L(t - c + \frac{l}{2})) - \sigma(-L(t - c - \frac{l}{2})), \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function, and L is a scaling factor. When L is large enough, this mask function becomes a step function which performs the exact clipping. We use the normalized weighted sum of the context features (weighted by the mask) as a feature representing $\mathbf{S}$. This operation approximates traditional mean-pooling over clipped frames.

3.3. Context Fusion

Because audio and visual representations are from two different modalities, merging them together is a crucial task in a multi-modal setting. We use three different context merging techniques (Fig. 3) to fuse the video $\mathbf{V}'$ and audio $\mathbf{A}'$ features obtained after the normalized soft mask clipping operation. We treat $\mathbf{V}'$ and $\mathbf{A}'$ as row vectors.

Multiplicative Mixture Fusion. The multiplicative mixture fusion can make the model automatically focus on information from a more reliable modality and reduce emphasis on the less reliable one [25]. Given a pair of features $\mathbf{V}'$ and $\mathbf{A}'$, the multiplicative mixture fusion first adds these two contexts and then concatenates the added context with the two original ones. That is, it produces a final context as follows,

$$\mathbf{C}_{final} = (\mathbf{V}' + \mathbf{A}')||\mathbf{V}'||\mathbf{A}' \qquad (8)$$

where $+$ and $||$ are the element-wise addition and column-wise concatenation respectively.

Multi-modal Context Fusion. This fusion strategy is similar to Eq. 6. But here, we apply the fusion technique on $\mathbf{A}'$ and $\mathbf{V}'$ (segments as opposed to full video context),

$$\mathbf{C}_{final} = (\mathbf{V}' + \mathbf{A}')||(\mathbf{V}' \cdot \mathbf{A}')|| \mathrm{fc}(\mathbf{V}'||\mathbf{A}'). \qquad (9)$$

MUTAN Fusion. MUTAN fusion was first proposed in [6] to solve visual question answering tasks by fusing visual and linguistic features. We adopt the fusion scheme to fuse $\mathbf{V}'$ and $\mathbf{A}'$. With the idea of Tucker decomposition [39], we first reduce the dimension of $\mathbf{V}'$ and $\mathbf{A}'$,

$$\mathbf{V}'' = \tanh(\mathbf{V}' \times \mathbf{W}_v) \qquad (10)$$

$$\mathbf{A}'' = \tanh(\mathbf{A}' \times \mathbf{W}_a) \qquad (11)$$

where $\mathbf{W}_v$ and $\mathbf{W}_a$ are learnable parameters and $\tanh(\cdot)$ is the hyperbolic tangent function. Then we produce final context as folows:

$$\tilde{\mathbf{C}} = ((\mathbf{T}_c \times_1 \mathbf{V}'') \times_2 \mathbf{A}'') \tag{12}$$

$$\mathbf{C}_{\text{final}} = \text{squeeze}(\tilde{\mathbf{C}}) \times \mathbf{W}_o, \tag{13}$$

where $\mathbf{T}_c$ and $\mathbf{W}_o$ are learnable parameters. $\times_i, i \in \{1, 2\}$ denotes the mode-i product between a tensor and a matrix, and $\times$ is the matrix multiplication operation. $\mathbf{T}_c$ models the interactions between the video and the audio modalities, which is a 3-dimension tensor; squeeze operator squeezes $\tilde{\mathbf{C}}$ into a row vector.

3.4. Training Loss

We follow the training procedure and loss function presented in [12] to train our networks. We employ the idea of cycle consistency [50] to train the sentence localizer and the caption generator, and treat the temporal segment regression as a classification problem. The final training loss is formulated as

$$L = L_{\text{c}} + \lambda_s L_{\text{s}} + \lambda_r L_{\text{r}} \tag{14}$$

where λ_s and λ_r are tunable hyperpramaters. L_{c} is the caption reconstruction loss, which is a cross-entropy loss measuring the similarity between two sentences. L_{s} is the segment reconstruction loss, which is an L2 loss. It measures the similarity between two temporal segments. L_{r} is the temporal segment regression loss, which is also a cross-entropy loss, because we regard the temporal segment regression as a classification problem.

4. Experiments

In this section, we first describe the dataset used in our experiments, which is an extension of the ActivityNet Captions Dataset [20] (Sec. 4.1). Then we present the experimental setup and implementation details (Sec. 4.2). Lastly, we discuss the experimental results for both unimodal (*i.e.*, trained using either audio or video modality) and multimodal approaches (Sec. 4.3).

4.1. Dataset

ActivityNet Captions dataset [20] is a benchmark for large-scale dense event captioning in videos. The dataset consists of 20,000 videos where each video is annotated with a series of temporally aligned captions. On average, one video corresponds to 3.65 captions. However, besides the captions, the current dataset only provides C3D features [18] for visual frames, no original videos. To obtain the audio tracks for those videos, we needed to find the original videos on YouTube and download the audios via the provided URLs. Around 5,000 videos are unavailable on YouTube now. We are able to find 8026 videos

Features	M	C	R	B@1	B@2	B@3	B@4	S
Pretrained model								
MFCC	2.70	6.46	6.74	5.52	1.74	0.67	0.21	3.51
CQT	2.38	5.60	5.72	4.37	1.57	0.46	0.13	2.90
SoundNet	2.63	5.76	6.99	6.28	1.81	0.38	0.12	3.44
Final model								
MFCC	3.36	9.56	8.51	6.68	2.55	1.23	0.60	4.20
CQT	3.25	8.97	7.43	6.34	2.69	0.93	0.32	3.63
SoundNet	3.41	9.21	8.50	7.19	2.15	0.49	0.13	4.22

Table 1: **Audio Only Results.** Illustrated are dense captioning results of pretrained and final models using audio only.

(out of 10009 videos) for training and 3880 videos (out of 4917 videos) for validation. We use those available training/validation videos throughout our experiments.

4.2. Experiment Setup and Implementation Details

We follow the experiment protocol in [12] to train and evaluate all the models. We consider the models proposed in [12] as our baselines, *i.e.*, unimodal models that only utilize audio or visual features. Due to the difference in the number of videos for training and validation from the original dataset, we run all the experiments from scratch using the PyTorch implementation provided by [12][1]. The dimensions of the hidden and output layers for all GRU RNNs (audio/video/caption encoders and caption decoders) are set to 512. We also follow [12] to build the word vocabulary (containing 6,000 words) and preprocess the words.

Training. Weak supervision means that we do not have ground-truth temporal segments. We first train the caption generator only (*pretrained model*), and then train the sentence localizer and caption generator together (*final model*). To train the pretrained model, we input the entire context sequence (*Fake Proposal*, $\mathbf{S} = (0.5, 1)$). We use the weights of the pretrained model to initialize the relevant weights in the final model. For both pretrained model and final model, we train them in both unimodal and multi-modal settings. To train unimodal models, we use initial learning rates 0.0001 and 0.01 for audio and video respectively with a cross-entropy loss. While training our multi-modal models, we set the initial learning rates to 0.0001 for the network parts that have been initialized with the pretrained weights, and 0.01 for other network components. λ_s and λ_r in Eq. 14 are both set to 0.1. We train the networks using stochastic gradient descent with a momentum factor of 0.8.

Testing. To test the pretrained models, we select one random ground truth description as well as random temporal segment instead of entire video unlike training. For the final models, following [12], we start from 15 randomly guessed temporal segments, and apply one round of fixed-point iteration and the IoU filtering mechanism to obtain a set of filtered segments. Caption generators are applied to the filtered segments together with context features to produce the dense event captions.

---

[1]https://github.com/XgDuan/WSDEC

Fusion Strategies	M	C	R	B@1	B@2	B@3	B@4	S	mIoU
Pretrained model									
Multiplicative mixture fusion	3.59	8.12	7.51	7.12	2.74	1.22	0.56	4.58	-
Multi-modal context fusion	3.55	7.91	7.54	7.24	2.78	1.28	0.62	4.45	-
MUTAN fusion	**3.71**	**8.20**	**7.71**	**7.45**	**2.92**	**1.31**	**0.63**	**4.78**	-
Final model									
Multiplicative mixture fusion	4.89	**13.97**	10.39	9.92	4.17	1.85	0.88	5.95	29.87
Multi-modal context fusion	**4.94**	13.90	10.37	9.95	4.20	**1.86**	0.89	5.98	29.91
MUTAN fusion	4.93	13.79	**10.39**	**10.00**	**4.20**	1.85	**0.90**	6.01	**30.02**

Table 2: **Fusion Strategies.** Testing results for different context fusion strategies for integrating audio and video modalities are illustrated for both pretrained and final models. We use MFCC audio features and C3D video features for all experiments.

Model	M	C	R	B@1	B@2	B@3	B@4	S	mIoU
Pretrained model									
Unimodal (C3D video feature) [12]	3.66	8.20	7.42	7.06	2.76	1.29	0.62	4.41	-
Unimodal (SoundNet audio feature)	2.63	5.76	6.99	6.28	1.81	0.38	0.12	3.44	-
Unimodal (MFCC audio feature)	2.70	6.46	6.74	5.52	1.74	0.67	0.21	3.51	-
Multi-modal (SoundNet audio + C3D video feature)	3.72	8.02	7.50	7.12	2.74	1.23	0.58	4.46	-
Multi-modal (MFCC audio + C3D video feature)	3.71	8.20	7.71	7.45	2.92	1.31	0.63	4.78	-
Final model									
Unimodal (C3D video feature) [12]	4.89	13.81	9.92	9.45	3.97	1.75	0.83	5.83	29.78
Unimodal (SoundNet audio feature)	3.41	9.21	8.50	7.19	2.15	0.49	0.13	4.22	25.57
Unimodal (MFCC audio feature)	3.36	9.56	8.51	6.68	2.55	1.23	0.60	4.20	27.16
Multi-modal (SoundNet audio + C3D video feature)	**5.03**	**14.27**	10.35	9.75	4.19	**1.92**	**0.94**	**6.04**	29.96
Multi-modal (MFCC audio + C3D video feature)	4.93	13.79	**10.39**	**10.00**	**4.20**	1.85	0.90	6.01	**30.02**

Table 3: **Multi-modal Results.** Comparison among unimodal and our multi-modal models using MUTAN fusion.
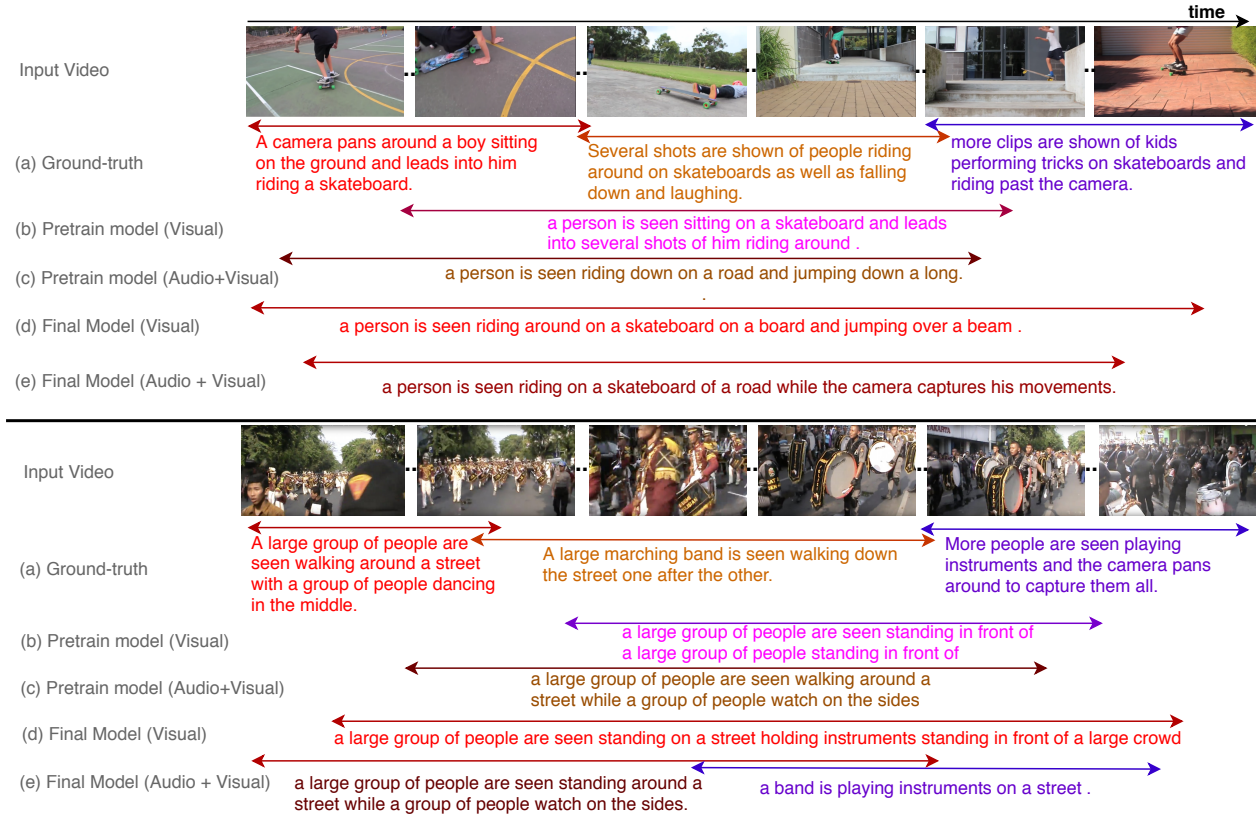

Figure 4: **Qualitative Results.** Both pretrained and final model results are illustrated of two videos. Captions are from (a) ground-truth; (b) pretrained model trained only using visual features; (c) multi-modal pretrained model; (d) final model trained with video features only; (e) our multi-modal final model for dense event captioning in videos.

Model	M	C	R	B@1	B@2	B@3	B@4	S
Unimodal (C3D)[10]	7.09	24.46	14.79	14.32	6.23	2.89	1.35	8.22
Multi-modal (SoundNet audio feature + C3D video feature)	7.02	24.22	14.66	14.18	6.13	2.88	1.41	7.89
Multi-modal (MFCC audio feature + C3D video feature)	**7.23**	**25.36**	**15.37**	**15.23**	**6.58**	**3.04**	**1.46**	**8.51**

Table 4: Results with ground-truth temporal segments.

Model	M	C	R	B@1	B@2	B@3	B@4	S
Unimodal (C3D) [10]	4.58	10.45	9.27	8.7	3.39	1.50	0.69	-
Multi-modal (SoundNet + C3D)	4.70	10.32	9.40	8.95	3.40	1.53	0.73	5.51
Multi-modal (MFCC + C3D)	**4.78**	**10.53**	**9.60**	**9.23**	**3.62**	**1.69**	**0.82**	**5.56**

Table 5: Pretrained model results on the full dataset.

Evaluation metrics. We measure the performance of captioning results using traditional evaluation metrics: METEOR (M) [5], CIDEr (C) [40], Rouge-L (R) [24], Spice (S) [1] and Bleu@N (B@N) [28]. For score computations, we use official scripts provided by [20][2]. Where appropriate, we use mean Intesection over Union (mIoU) to measure segment localization performance.

4.3. Experiment Results

Since audio features can be represented in a variety of ways [4, 30, 41], finding the best representation is challenging. We conduct experiments on both pretrained models and final models using different audio representations, *i.e.*, MFCC [19], CQT [7], and SoundNet [4], which are described in Sec. 3.1.1. Table 1 shows the experiment results of pretrained models and final models using only audio features. We can see that both MFCC and soundNet can generate comparable results.

As discussed in Sec. 3.3, in the multi-modal setting, choosing a good fusion strategy to combine both audio and video features is another crucial point. Table 2 shows comparison of different context fusion techniques using MFCC audio representations and C3D visual features (Sec. 3.1.2) for both pretrained models and final models. Among all fusion techniques, we find that MUTAN fusion is the most appropriate one for our weakly supervised multi-modal dense event captioning task. Therefore, we decide to use MUTAN fusion technique for our multi-modal models when comparing to unimodal models. Tab. 3 shows the testing results for comparison among unimodal and multi-modal approaches. We can see that our multi-modal approach (both MFCC and SoundNet audio with C3D video features) outperforms state-of-the-art unimodal method [12] in most evaluation metrics. Specifically on the Bleu@3 and Bleu@4 scores, it leads to 9% and 13% improvement respectively. Comparing among unimodel approaches, we are surprised to find that only using audio features achieves competitive performance. We trained our caption generator with GT segments to remove the effect of localization. The results are shown in Table 4. We also conduct experiment on pretrain caption generator using the full dataset where for some videos, au-

dio data is not available (treated as missing data). We consider zero feature vectors for missing audios. The results are shown in Table 5. In addition, we randomly selected 15 validation videos and invited 20 people to conduct human evaluation for comparing our multi-modal model to the visual-only one. The forced choice preference rate for our multi-modal model is 60.67%.

Figure 4 demonstrates some qualitative results for both pretrained models and final models. It displays the ground-truth captions along with the ones generated by unimodal models and our multi-modal models. The arrow segments indicate the ground-truth or detected temporal event segments. We utilize C3D visual features along with audio features. We can see that our multi-modal approaches outperform unimodal ones, both on caption quality and temporal segment accuracy.

Similar to [12], we are suffering from two limitations. One is that sometimes our multi-modal model can not detect the beginning of an event correctly. The other is that most of the time our final model only generates around 2 event captions, which means that the multi-modal approach is still not good enough to detect all the events in the weakly supervised setting. Overcoming of these two limitations is the focus of our future work.

5. Conclusion

Audio is a less explored modality in the computer vision community. In this paper, we propose a muti-modal approach for dense event captioning in a weakly supervised setting. We incorporate both audio features with visual ones to generate dense event captions for given videos. We discuss and compare different feature representation methods and context fusion strategies. Extensive experiments illustrate that audio features can play a vital role, and combining both audio and visual modalities can achieve performance better than the state-of-the-art unimodal visual model.

---

[2]https://github.com/ranjaykrishna/densevid_eval

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 8

[2] Mehmet Ali Arabacı, Fatih Özkan, Elif Surer, Peter Jančovič, and Alptekin Temizel. Multi-modal egocentric activity recognition using audio-visual features. *arXiv preprint arXiv:1807.00612*, 2018. 2

[3] Ido Ariav and Israel Cohen. An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. *IEEE Journal of Selected Topics in Signal Processing*, 2019. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 2, 3, 8

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 8

[6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2612–2620, 2017. 5

[7] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991. 3, 8

[8] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Polyphonic sound event detection using multi label deep neural networks. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2015. 2

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[10] Richard K Davenport, Charles M Rogers, and I Steele Russell. Cross modal perception in apes. *Neuropsychologia*, 11(1):21–28, 1973. 1

[11] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980. 2

[12] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3063–3073, 2018. 2, 3, 6, 7, 8

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017. 5

[14] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[15] Wangli Hao, Zhaoxiang Zhang, and He Guan. Integrating both visual and audio cues for enhanced video caption. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[16] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017. 1, 2

[17] Chiori Hori, Takaaki Hori, Tim K Marks, and John R Hershey. Early and late integration of audio features for automatic video description. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 430–436. IEEE, 2017. 2

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 4, 6

[19] Wenxin Jiang, Alicja Wieczorkowska, and Zbigniew W Raś. Music instrument estimation in polyphonic sound based on short-term spectrum match. In *Foundations of Computational Intelligence Volume 2*, pages 259–273. Springer, 2009. 3, 8

[20] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017. 1, 2, 3, 4, 6, 8

[21] Shankar Kumar and William Byrne. Minimum bayes-risk decoding for statistical machine translation. Technical report, JOHNS HOPKINS UNIV BALTIMORE MD CENTER FOR LANGUAGE AND SPEECH PROCESSING (CLSP), 2004. 2

[22] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500, 2018. 1, 3

[23] Thomas Lidy and Alexander Schindler. Cqt-based convolutional neural networks for audio scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, volume 90, pages 1032–1048. DCASE2016 Challenge, 2016. 2

[24] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics, 2004. 8

[25] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018. 5

[26] Yuan Liu and Moyini Yao. Best vision technologies submission to activitynet challenge 2018-task: Dense-captioning events in videos. *arXiv preprint arXiv:1806.09278*, 2018. 2

[27] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 8

[29] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE, 2016. 2

[30] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64, 2010. 3, 8

[31] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[33] Barry E Stein and M Alex Meredith. *The merging of the senses.* The MIT Press, 1993. 1

[34] Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. Explicit word error minimization in n-best list rescoring. In *Fifth European Conference on Speech Communication and Technology*, 1997. 2

[35] Yingxiang Sun, Jiajia Chen, Chau Yuen, and Susanto Rahardja. Indoor sound source localization with probabilistic neural network. *IEEE Transactions on Industrial Electronics*, 65(8):6403–6413, 2018. 2

[36] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160*, 2016. 2

[37] M Iftekhar Tanveer, Ji Liu, and M Ehsan Hoque. Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 863–866. ACM, 2015. 1

[38] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872*, 2018. 1, 2

[39] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 5

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 8

[41] Rivarol Vergin, Douglas O'shaughnessy, and Azarshid Farhat. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7(5):525–532, 1999. 8

[42] Jean Vroomen and Beatrice de Gelder. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance*, 26(5):1583, 2000. 1

[43] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018. 1, 2

[44] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*, 2018. 1, 2

[45] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5783–5792, 2017. 3

[46] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 396–405. IEEE, 2019. 3

[47] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016. 1

[48] Ting Yao, Yehao Li, Zhaofan Qiu, Fuchen Long, Yingwei Pan, Dong Li, and Tao Mei. Msr asia msm at activitynet challenge 2017: Trimmed action recognition, temporal action proposals and densecaptioning events in videos. In *CVPR ActivityNet Challenge Workshop*, 2017. 2

[49] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 1, 2

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 6