SEEING AND HEARING TOO: AUDIO REPRESENTATION FOR VIDEO CAPTIONING

Shun-Po Chuang, Chia-Hung Wan, Hung-Yi Lee

National Taiwan University

ABSTRACT

Video captioning task has been a widely research topic of computer vision and machine learning. Most of the related works consider pure visual contents for description generation. On the other hand, auditory contents contain rich information for describing the scenes, such as human speech or environment sounds, but not widely explored in video caption generation yet. In this paper, we take full advantage of auditory contents in videos and experimented different approaches of exploiting auditory contents. Audio information improved caption generation in terms of popular evaluation methods in natural language generation such as BLEU, CIDer and METEOR. We also measured the semantic similarities between generated captions and human provided ground truth by sentence embedding, and found that machine generates captions more semantically related to the ground truth with good use of multi-modal contents. By analyzing the generated sentences, we found that some ambiguous situations for visual-only models which obtained incorrect results are resolved by the auditory-considering approaches.

Index Terms— Video caption generation

1. INTRODUCTION

Video captioning, in which machine generates one or multiple sentences to describe the content of a video clip, is a critical step towards machine intelligence, and it has many useful applications including video retrieval, automatic video subtitling, blind navigation, etc. Video captioning has been widely studied, and most of the related works exploit visual-contents only. However, humans understand the environments by not only seeing, but also hearing, so we believe auditory-contents in video also bring rich information for video captioning. For example, it is difficult to discriminate 'talking' and 'singing' by seeing, but they can be easily distinguished by hearing. For another example, it is hard to know whether two people are 'talking' or 'arguing' only by vision, but the two situations can be easily separated by the loudness of voice. Moreover, it is possible that the source of the sound does not appear in the video. In such situation, machine cannot know the existence of the sound source without hearing. Therefore, machine should consider both visual and auditory contents in video when doing captioning. There are already some attempts using auditory-content to improve video captioning [1, 2], but auditory-contents are not always shown to be helpful [3].

In this paper, we use a large variety of auditory-content representation techniques to improve video caption generation. Besides MFCC features which have been used in the previous work [1, 2, 3], we also use ASR system outputs, and the audio representation extracted by deep neural network including Audio Word2Vec [4] and SoundNet [5]. We compared the performance of different audio representations under different video caption generation models. The common evaluation measures like BLEU [6], CIDEr [7], ROUGE_L [8] and METEOR [8] evaluate the difference between generated captions and human-labeled sentences only on literal level. Besides evaluating results literally, in this paper, we further evaluate results on semantic level by computing the similarities by sentence embeddings [9].

2. RELATED WORKS

There are two categories of methods to achieve video captioning: template-based methods[10, 11, 12, 13] and sequence learning methods[14, 15, 16, 17, 18, 19, 20, 21]. For template-based methods, the objects are first recognized, and the names of the detected objects are filled into predefined language templates to generate captions. The diversity of generated sentences highly depends on the number of the predefined templates. Sequence learning methods generate sentences with more flexible syntactical structure by sequence-to-sequence model. The sequence-to-sequence model learns the probability of a word sequence given a video clip using the encoder-decoder architecture. In encoder-decoder architecture, the encoder takes a video clip as input and encodes it to the embedding space, then the decoder decodes the embedding vector into descriptions word by word.

Most of the video captioning task are based on visual contents. There is only a few examples exploiting auditory-contents. Ramanishka *et al.* proposed a multi-modal video description model [1], which exploited the category label and audio which was represented by MFCCs. Qin Jin *et al.* also considered multi-modal in video captioning [2], and the audio was represented by acoustic codebook and concatenated with visual features. Hori *et al.* proposed to expand the attention model to selectively attend on multi-modal features. MFCCs

This work was sponsored by MediaTek Inc., Hsin-chu, Taiwan.

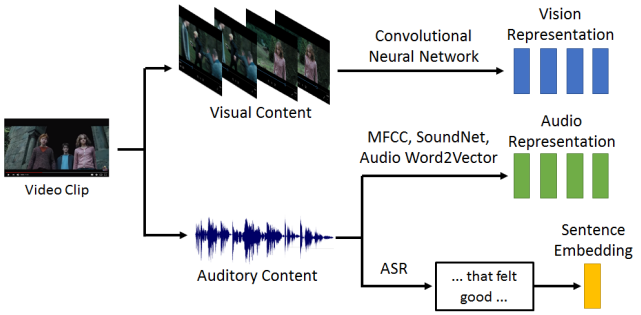

Fig. 1: *The features used in video captioning.*

were used to represent audio as well, and they were the inputs of a biLSTM which was jointly learned with the whole video caption models.

In video captioning, the visual features are usually extracted by a pre-trained deep convolution neural network like VGGnet [22], Alexnet [23] and GoogleNet [24], but audio is usually simply represented by MFCCs. On the other hand, there are several ways to pre-train a deep learning model for audio feature extraction [4, 25, 26, 27, 5], but not exploited in video captioning. Chung *et al.* proposed an unsupervised way for audio representation which is encoded by a sequence-to-sequence auto-encoder and named such features as Audio Word2Vec [4]. Aytar *et al.* proposed a deep convolution neural network named SoundNet for audio feature extraction [5] . By transfer knowledge from ImageCNN [28] and PlaceCNN [29], SoundNet features can detect the environment sounds and object sounds. The features extracted by sequence-to-sequence auto-encoder, that is, Audio Word2Vec, and SoundNet are used to enhance video captioning in this paper.

3. APPROACH

3.1. Feature Representation

The features used to represent a video clip in this paper is shown in Fig. 1. Here visual-contents are represented as a sequence of vectors extracted by a CNN. In this paper, we focus on exploiting the auditory-contents. Besides Mel-Frequency Cepstral Coefficient (MFCC), we use Audio Word2Vec and SoundNet for audio representations. ASR transcriptions are also considered. More details are shown below.

Audio Word2Vec. Audio Word2Vec encodes an audio segment into a fixed-length vector [4], which is learned from audio data without human annotation using Sequence-to-sequence Audoencoder (SA). SA consists of two LSTMs, one for encoding and the other for decoding. Encoder reads an audio segment represented as an acoustic feature (e.g. MFCC) sequence $X = (x_1, x_2, x_3...x_T)$ and maps it to a fixed-length vector z. Then decoder maps the fixed-length vector z to another sequence $Y = (y_1, y_2, y_3...y_T)$. Encoder and decoder are jointly trained to minimize the difference between sequence X and Y, measured by mean squared error $\sum_{t=1}^{T} ||x_t - y_t||^2$. Because the input X can be reconstructed from the fixed-length vector z, it will be the meaningful representation of input sequence X. Because training target of SA is the input of network, Audio Word2Vec does not need labeled data to train. In this paper, the auditory-content of a video clips are segmented into audio segments with equal time interval, and then each segment is represented by the fixed-length vector z. Hence, by Audio Word2Vec, the auditory-content is also represented as a sequence of vectors.

SoundNet. SoundNet [5] is a deep CNN for natural sound recognition. By transfer learning from the CNNs for object and scene recognition, SoundNet learns to classify object and scene by auditory-contents only. The hidden layer output of SoundNet can be used to represent a small segment of raw waveform, so we can represent the auditory-content of a video clip by a sequence of vectors based on SoundNet.

Sentence Embedding of ASR transcriptions. The content of human voice in a video clip is helpful for machine to generate more specific descriptions. Therefore, we utilize automatic speech recognition (ASR) system to transcribe human voice in videos, and using sentence embedding technique [30] to represent the transcription as a fixed-length vector.

3.2. Model Architecture

The existing models [17, 20] are adjusted to integrate visual-content with auditory-content. Three model architectures used here are shown in Fig. 2, and respectively described in Section 3.2.1, Section 3.2.2 and Section 3.2.3. In Fig. 2, the feature sequence extracted from the visual-content is the blue vectors, while the green vectors can be MFCCs, or features from Audio Word2Vector or SoundNet. Here we assume the number of blue and green vectors is the same for the same video clip (it will be more clear how to achieve that in Section 4.2). We will describe how to use the sentence embedding of ASR system output in Section 3.2.4.

3.2.1. Bidirectional LSTM

The first model in Fig.3 (A) is modified from the model proposed by Bin *et al.* [20]. The input is the concatenation of visual features with MFCCs, or features from Audio Word2Vector or SoundNet. The model encodes the input with a bidirectional-LSTM (the red and pink blocks in Fig.3 (A)). The outputs of bidirectional-LSTM are further processed by a forward LSTM (the black blocks), whose final output is the whole video representation[1]. In decoding stage, the LSTM for generating description (the orange blocks) takes

[1]In the original paper [20], the outputs of bidirectional-LSTM are concatenated with original visual features as input of another LSTM for generating final video representation, but it did not show to be helpful in our preliminary experiments.

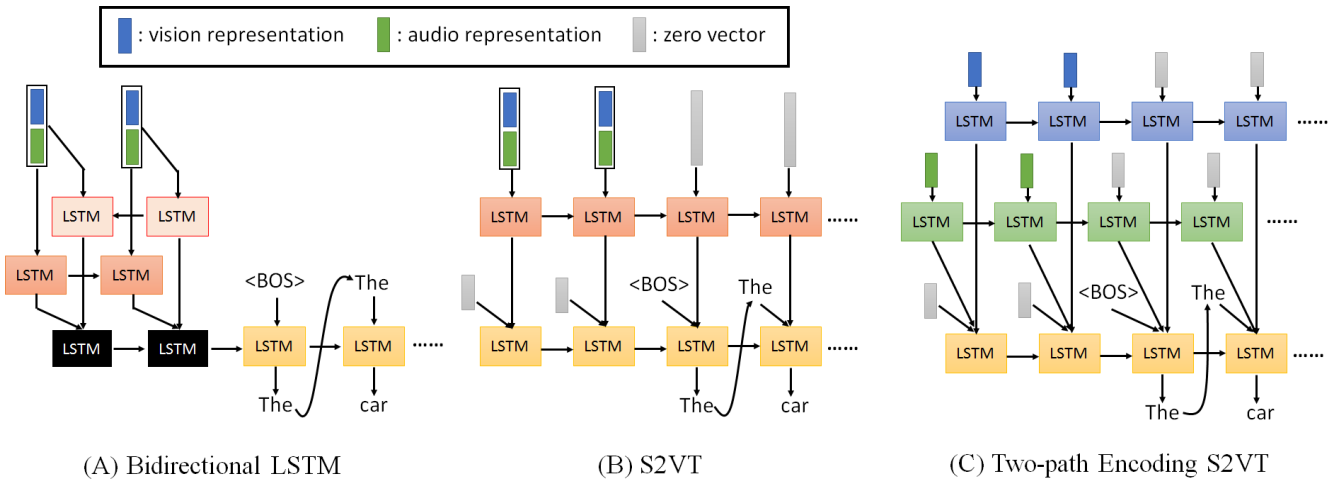

Fig. 2: *Model architecture for video captioning with visual- and auditory-contents.*

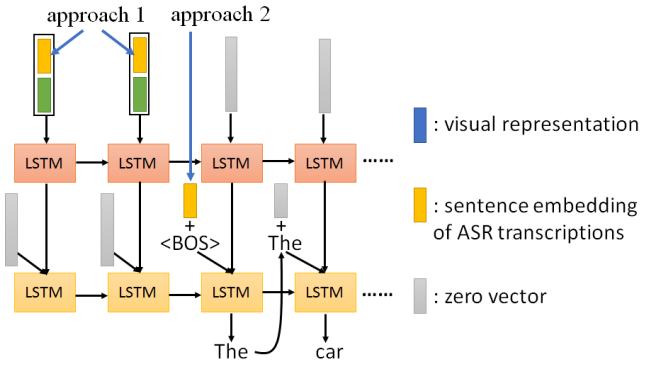

Fig. 3: Here we use S2VT model to describe the two approaches of exploiting ASR transcriptions. The same idea can be used in all models in Fig. 2.

the whole video representation as input, and output the description word-by-word until the EOS (End of Sentence) token is generated.

3.2.2. *S2VT*

The second model we used is S2VT model [17], shown in Fig. 2 (B). The model is a two layer LSTM, the first layer for visual content processing (the upper blocks in red), and the second layer for caption generation (the lower blocks in orange). The whole caption generation process based on S2VT has an encoding stage (the first two time steps in Fig. 2 (B)) and a decoding stage (the rest steps). In encoding stage, the concatenation of visual and audio features are sequentially fed into upper LSTM. The outputs of the upper LSTM will be concatenated with zero vectors in vocabulary size as the input of the lower LSTM. In decoding stage, the input of upper LSTM are replaced by zero vectors. The input of lower LSTM is the concatenation of the output of upper LSTM and one-hot encoding of the word generated in the last time step. For the first time step of decoding stage, the input would be BOS (Begin of Sentence) token. The output of lower LSTM are the words in the generated caption. We collect the words in sequence until EOS token is generated.

3.2.3. *Two-path Encoding S2VT*

In Section 3.2.2, the upper LSTM in S2VT has to process both visual and audio information. Here we modify S2VT model to reduce the work load of upper LSTM. The visual and acoustic features are encoded by separate LSTMs (blue and green blocks), and then the outputs of the two LSTMs are concatenated as the input of lower LSTM.

3.2.4. *How to Exploit ASR Transcriptions?*

The sentence embedding of ASR transcription is a single vector instead of a vector sequence, so it should be used differently from other audio representations. Here the two approaches to use sentence embedding are demonstrated by S2VT model in Fig. 3. For the first approach, we duplicate the sentence embedding, and concatenate the sentence embedding with the vision features. For the second approach, to let sentence embedding influence the caption generation process more directly, we concatenate the sentence embedding with the one-hot encoding of BOS token.

4. EXPERIMENTAL SETUP

4.1. Dataset

Microsoft Research - Video to Text (MSR-VTT) Corpus [31] was used in the experiments. It contains 7010 video clips for training and 2990 clips for testing. Each clip corresponds to

(A) Bi-LSTM		BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGH-L	METEOR
(A-1)	Visual-Only	0.636	0.483	0.360	0.260	0.252	**0.517**	0.223
(A-2)	+ MFCC	0.649	0.482	0.355	0.255	0.249	0.510	0.221
(A-3)	+ A2V-Speech	0.646	0.485	0.356	0.254	0.241	0.509	0.220
(A-4)	+ A2V-Video	0.641	0.486	**0.362**	**0.261**	**0.258**	0.515	0.220
(A-5)	+ SoundNet	**0.653**	**0.492**	**0.362**	0.260	0.247	0.516	**0.226**
(A-6)	+ ASR-Con	0.649	0.487	0.360	0.259	0.245	0.513	0.222
(A-7)	+ ASR-Dec	0.600	0.442	0.320	0.224	0.220	0.499	0.210
(B) S2VT		BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGH-L	METEOR
(B-1)	Visual-Only	0.666	0.513	0.386	0.279	0.260	0.529	0.226
(B-2)	+ MFCC	0.676	0.527	0.403	0.297	0.282	0.534	0.231
(B-3)	+ A2V-Speech	0.675	0.525	0.399	0.293	0.290	0.537	0.232
(B-4)	+ A2V-Video	0.679	0.527	0.399	0.292	0.289	0.537	0.233
(B-5)	+ SoundNet	**0.682**	**0.536**	**0.412**	**0.305**	0.279	**0.539**	**0.233**
(B-6)	+ ASR-Con	0.673	0.521	0.393	0.289	**0.288**	0.534	0.232
(B-7)	+ ASR-Dec	0.681	0.530	0.405	0.297	0.265	0.536	0.231
(C) 2-S2VT		BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGH-L	METEOR
(C-1)	Visual-Only	0.666	0.513	0.386	0.279	0.260	0.529	0.226
(C-2)	+ MFCC	0.624	0.458	0.329	0.225	0.170	0.489	0.204
(C-3)	+ A2V-Speech	0.660	0.493	0.362	0.254	0.224	0.515	0.218
(C-4)	+ A2V-Video	0.664	0.515	0.392	0.289	**0.280**	**0.536**	0.231
(C-5)	+ SoundNet	0.681	0.525	0.394	0.286	0.260	0.531	**0.232**
(C-6)	+ ASR-Con	**0.688**	**0.530**	0.397	0.285	0.267	0.534	0.231
(C-7)	+ ASR-Dec	0.681	**0.530**	**0.405**	**0.297**	0.265	**0.536**	0.231

Table 1: Evaluation scores of different models and features.

20 natural language descriptions labeled by AMT workers. Because some clips are not available, and some do not have audio, we only used 5928 clips for training and 2623 for testing. The training and testing json files containing video id and corresponds captions we used in the experiments are provided here for easy reproduction[2].

4.2. Feature Representation

Visual Content Representation. We extracted visual features via **fc7** layer of VGG 19 layers model [22], which are 4096-dimensional vectors. Following the setting in the previous work [17], in each video clip, visual features were only extracted from 80 sampled frames, which resulted in 80 feature vectors for each clip. The frames were sampled by the same time interval in each clip, but the lengths of the time intervals were different in different clips. This paper focuses on the auditory features, so we do not use other kinds of visual features like optical flow, C3D, etc [17, 1, 3, 19]. Because 80 feature vectors were extracted from the visual content, the auditory content of a video clip was also represented by 80 vectors no matter the representation approach used, except sentence embedding of ASR system output, which is only a single vector.

MFCC: 39-dimensional MFCCs were extracted by Kaldi toolkit [32], and cepstral mean and variance normalization were applied. For each clip, we sampled 80 MFCC features by the same time interval for caption generation.

Audio Word2Vec: We first segmented the audio of each clip into 80 segments with equal length. As in the previous work [4], we used sequence-to-sequence autoencoder to encode each segment into a 300-dimensional vector. The Audio Word2Vec model was either learned from Librispeech ASR corpus [33] or audio content of the training video clips in MSR-VTT corpus, so we have two sets of Audio Word2Vec in the following experiments. In this way, we can observe the influence of the training data domains of Audio Word2Vec on video captioning. Librispeech ASR corpus contains approximately 1000 hours english speech data. By training on this corpus, Audio Word2Vec maintained the characteristics of human speech [4]. The video clips of MSR-VTT contain sound other then human speech, so Audio Word2Vec learned from MSR-VTT may include information besides human speech in the extracted feature vectors.

SoundNet: With over two million videos downloaded from Flickr, which resulted in over one year continuous natural sound and video, SoundNet learned good representation of audio from large unlabeled video. 5-layers and 8-layers models are available [29]. We used **pool5** layer in 8-layers model which performed best in classification tasks [29]. A sequence of 256-dimensional vectors is extracted from a video clip. Because the number of feature vectors extracted by SoundNet can be more/less than 80, we have to down/up sample

[2] https://github.com/alex82528/video_captioning_data

	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGH-L	METEOR
(A) Visual-Only + Soundnet	0.682	0.536	0.412	0.305	0.279	0.539	0.233
(B) +A2V-Speech	0.698	0.553	0.430	0.323	0.315	0.552	0.241
(C) +ASR-Con	0.701	0.557	0.432	0.327	0.319	0.553	0.244
(D) +A2V-Speech+ASR-Con	**0.703**	**0.559**	**0.436**	**0.331**	**0.331**	**0.557**	**0.245**

Table 2: Ensembling the results of different audio representations using S2VT model. The results in row (A) is the results in (B-5) in Table 1.

the features. Each dimension in a feature vector extracted by SoundNet corresponds to a specific pattern. Because the feature vectors are the outputs of a pooling layer, and ReLU activation function is used, all the values in the extracted vectors are non-negative. We believe that each non-negative value revealing the existence of a specific pattern, so more non-negative values in the vectors means that there are richer information in the corresponding time span. We define the importance of a feature vector by the summation of the elements in the feature. When a clip has $80 + K$ frames, we dumped the least important K frames; when a clip has $80 - K$ frames, we duplicated the most important K frames[3].

Sentence Embedding: We used Microsoft Azure Bing Speech API to generate ASR transcriptions, and applied a sentence embedding model [30] to encode it. The model was trained on English tweets, which output a 700-dimensional vectors given a sentence. For clips not containing speech, the ASR system would not generate any results. In such case, we used a zero vector to represent the sentence.

4.3. Parameter Setting

All LSTM were initialized by uniform distribution in range of -0.1 to 0.1. S2VT-based models used 256-dimensional LSTM, and 512-dimensional LSTM were used in bidirectional LSTM-based model in the following experiments[4]. Vocabulary size were set to 3000 and we did not use pre-trained language model on decoding stage. We used a linear transformation to reduce one-hot representation of a word to 300-dimensional vector as the input of LSTM, the weights of the transformation were jointly trained with the model. We trained models for 200 epochs with batch size 100 and Adam optimizer.

[3]Although this methods make visual and auditory contents asynchronous, it leads to slightly better performance than sampling SoundNet feature at equal intervals. We do not show the experiments of the comparison due to space limitation.

[4]In the preliminary experiments, we found that 256- and 512- dimensional LSTM achieved the best results for S2VT and BiLSTM respectively. We do not show the results due to space limitation.

Table 3: Semantic evaluation by the similarities between the sentence embedding of the generated captions and ground truth. The table shows the results of S2VT with different features (part (b) of Table 1 and Table 2).

Features	Similarity
(A) Visual-only	0.443
(B) +MFCC	0.452
(C) +A2V-Speech	**0.456**
(D) +A2V-Video	**0.456**
(E) +SoundNet	0.450
(F) +ASR-Con	0.454
(G) +ASR-Dec	0.448
(H) +A2V-Speech+SoundNet	0.470
(I) +A2V-Speech+ASR-Con	0.471
(J) +SoundNet+ASR-Con	0.472
(K) +A2V-Speech+SoundNet+ASR-Con	**0.477**

4.4. Evaluation Methods

BLEU, METEOR, ROUGH-L and CIDEr were used[5]. These metrics are widely utilized on natural language generation task such as machine translation. We also proposed a new evaluation measure. We trained a Sent2Vec model on the testing set of MSR-VTT Corpus. We encoded the ground truth descriptions and generated descriptions into a 700-dimensional vector, and then computed the cosine similarity between them. Higher value of similarity means the generated description is semantically close to the ground truth. The front method can be seem as literal evaluation, and the latter is a semantic evaluation.

5. EXPERIMENTAL RESULTS

5.1. Evaluation Scores

Table 1 shows the results of literal evaluation. Parts (A), (B) and (C) are respectively the results for bidirectional-LSTM, S2VT and two-path encoding S2VT (2-S2VT). Row (1) in each part is the baseline using visual information only. The results in (B-1) and (C-1) are the same because without considering auditory-content two-path encoding S2VT reduced to the original S2VT. Rows (2) to (7) exploiting auditory-

[5]implemented by `https://github.com/vsubhashini/caption-eval`

Table 4: Generated results with different features.

https://www.youtube.com/watch?v=1DQwhuFhcJk start time: 168.17, end time: 179.48	
Ground Truth	person singing a song
Visual-Only	a man is **talking** about a UNK
+A2V-Speech	a man is **singing** a song
+SoundNet	a man is **singing**

https://www.youtube.com/watch?v=KMydT2yve3k start time: 955.61, end time: 972.57	
Ground Truth	people on tv show are talking to a caller
Visual-Only	a **woman** is talking about the lady
+A2V-Speech	a **man** is talking to a **woman**
+SoundNet	a **man** is talking about a **woman** s UNK

https://www.youtube.com/watch?v=13iHUdS3Qmo start time: 223.44, end time: 234.01	
Ground Truth	a man is in a rap music video
Visual-Only	a man and woman are food
+A2V-Speech	a **music video** with a band
+SoundNet	a man and woman are talking

content in different ways. A2V-Speech and A2V-Video stand for Audio Word2Vec learned from Librispeech ASR corpus and MSR-VTT corpus respectively. ASR-Con and ASR-Dec represent the two approaches of using sentence embedding: concatenating with visual features or considering as the input at decoding stage. The results in rows (B-7) and (C-7) are the same because when ASR sentence embedding only used in the decoding state, S2VT and two-path encoding S2VT are exactly the same. The results in bold is the best results among all kinds of features with the same model architecture, while the results with bottom line means it is the best result across all models and features.

Auditory contents make unapparent influence on bidirectional LSTM model. MFCC is not helpful in terms of all measures, except BLEU@1 (rows (A-2) v.s. (A-1)). A2V-Video and SoundNet increased the scores slightly in terms of some evaluation measures, but not all of them (rows (A-4), (A-5) s.v. (A-1)). This may be because bidirectional LSTM model used a single vector to represent the whole video, and it may be difficult to use a vector to represent both vision and audio information[6]. The obvious improvements were achieved by S2VT and 2-S2VT. With S2VT model (part (B)), all auditory contents, especially SoundNet, enhanced the scores. In 2-S2VT model (part (C)), audio features enhanced the performance, except MFCC features and A2V-Speech (rows (C-2), (C-3) v.s. (C-1)). Compared the results of S2VT and 2-S2VT models (parts (C) v.s. (B)), we found that with audio information, S2VT outperformed 2-S2VT in all cases, except using ASR-Con (rows (C-6) v.s. (B-6)). The results shows that the interaction between the vision and audio information in the upper LSTM of S2VT is helpful, and 2-S2VT outperformed S2VT with ASR-Con probably because ASR result and the features extracted by VGG are not at the same level, and little interaction is needed between them. Compared across all model and features in Table 1, SounNet feature plus S2VT model performed the best (row (B-5)). Because S2VT model achieved the best performance in Table 1, it was used in all the following experiments.

We further integrated the results from S2VT models (part (B) of Table 1). The results of ensemble are listed in Table 2. The likelihood of the output of each model[7] was first computed, and normalized with its length. We consider the normalized likelihoods as the confidence of the generated sentences. Among the results to be integrated, the generated sentence with the highest normalized likelihood is selected for evaluation. Because S2VT plus SoundNet (row (B-5) in Table 1) achieved the best results, we integrated it with the results of other models. Many different combination were tested[8], and we found that ASR-Con and A2V-Speech are most complementary with SoundNet, which are rows (B), (C) and (D) of Table 2. Because SoundNet aims at detecting audio events instead the content of human speech[9], it is reasonable that ASR-Con and A2V-Speech, which contain the information of the content of speech, improved the performance after integration.

Then we evaluate the semantic correctness of the generated sentences. Each video has several ground truth descriptions, so we computed the cosine similarities between the generated sentence and each ground truth description, and took the maximum similarity as the evaluation score. We averaged the maximum similarity of each video clip over the testing set. Table 3 shows the semantic evaluation results of S2VT with different features (the literal evaluations of the same set of results have been shown in part (b) of Table 1 and Table 2). Row (A) is the results of using vision feature only, and rows (B) to (G) are the results for exploiting auditory-contents. In terms of semantic evaluation, auditory-contents also increase the evaluation scores no matter the representation approaches. Rows (H) to (K) are the results of ensemble. Integrating the results based on A2V-Speech, SoundNet and ASR-Con improved the performance in terms of semantic evaluation.

5.2. Observation

Table 4 lists some example results from the testing set of MSR-VTT. The YouTube link and the start and end time of each example are shown. In each example, we display one of the ground truth descriptions, and the generated captions

[6] In the future work, we are trying to use attention mechanism to deal with the problem.

[7] summing up the logarithm probability of each generated word in sentence

[8] We cannot show the whole results due to space limitation.

[9] It is very likely all human speech is considered as the same event and has very similar features based on SoundNet.

of vision only, and with A2V-Speech or SoundNet. It can be inferred that auditory features are sensitive to the sound of gender and auditory-related action. In the fist example video clip, there is a person sinning. Without hearing, the machine outputted "a man is talking ... ". By considering auditory-content, machine can generate the description "a man is singing ...". In the second example, a male host is talking to a female caller. Compared the results without and with audio features, we found that machine aware there is a man in the scene only when it hears. The last example is a music video, and only the model based on A2V-Speech produced related caption.

6. CONCLUSION

In this paper, we utilized different kinds of acoustic feature with different video captioning models. We found that considering auditory contents truly enhances the task. S2VT plus SoundNet achieved the best performance, and it can be further improved by integrating with models exploiting human speech. In the future work, we will try more different models and investigate how vision and audio information interact with each other in video caption generation.

7. REFERENCES

[1] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko, "Multimodal video description," in *Proceedings of the 2016 ACM on Multimedia Conference*, New York, NY, USA, 2016, MM '16, pp. 1092–1096, ACM.

[2] Qin Jin, Junwei Liang, and Xiaozhu Lin, "Generating natural video descriptions via multimodal processing," in *INTERSPEECH*, 2016.

[3] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R Hershey, and Tim K Marks, "Attention-based multimodal fusion for video description," *arXiv preprint arXiv:1701.03126*, 2017.

[4] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[6] 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*. Association for Computational Linguistics, 2002, pp. 311–318.

[7] 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*, 2015, pp. 4566–4575.

[8] 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*, 2005, vol. 29, pp. 65–72.

[9] Quoc Le and Tomas Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[10] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 433–440.

[11] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele, "Coherent multi-sentence video description with variable level of detail," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 184–195.

[12] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework.," in *AAAI*, 2015, vol. 5, p. 6.

[13] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2712–2719.

[14] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[15] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Describing videos by exploiting temporal structure,"

in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[16] 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*, 2016, pp. 5288–5296.

[17] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4534–4542.

[18] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4584–4593.

[19] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.

[20] Yi Bin, Yang Yang, Zi Huang, Fumin Shen, Xing Xu, and Heng Tao Shen, "Bidirectional long-short term memory for video description," *CoRR*, vol. abs/1606.04631, 2016.

[21] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1029–1038.

[22] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[25] Herman Kamper, Weiran Wang, and Karen Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4950–4954.

[26] Wanjia He, Weiran Wang, and Karen Livescu, "Multi-view recurrent neural acoustic word embeddings," *CoRR*, vol. abs/1611.04496, 2016.

[27] Merlijn Blaauw and Jordi Bonada, "Modeling and transforming speech using variational autoencoders.," in *INTERSPEECH*, 2016, pp. 1770–1774.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[29] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[30] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," *arXiv*, 2017.

[31] 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 (CVPR)*, 2016.

[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.