INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multimodal Speech Summarization through Semantic Concept Learning

Shruti Palaskar[1], *Ruslan Salakhutdinov*[1], *Alan W Black*[1], *Florian Metze*[1,2]

[1] Carnegie Mellon University, USA
[2] Facebook AI, USA

spalaska@cs.cmu.edu, fmetze@fb.com

Abstract

We propose a cascaded multimodal abstractive speech summarization model that generates semantic concepts as an intermediate step towards summarization. We describe a method to leverage existing multimodal dataset annotations to curate groundtruth labels for such intermediate concept modeling. In addition to cascaded training, the concept labels also provide an interpretable intermediate output level that helps improve performance on the downstream summarization task. On the open-domain How2 data, we conduct utterance-level and video-level experiments for two granularities of concepts: Specific and Abstract. We compare various multimodal fusion models for concept generation based on the respective input modalities. We observe consistent improvements in concept modeling by using multimodal adaptation models over unimodal models. Using the cascaded multimodal speech summarization model, we see a significant improvement of 7.5 METEOR points and 5.1 ROUGE-L points compared to previous methods of speech summarization. Finally, we show the benefits of scalability of the proposed approaches on 2000 h of video data.

Index Terms: speech summarization, semantics, concept learning

1. Introduction

Summarization generates a condensed and comprehensive version of the input information and has been widely studied for textual documents [1, 2, 3]. Summarization assists users in understanding large content in a shorter time period while maintaining its informativeness. Most of the work on text summarization has focused on single-document news domain summarization [4, 5] with some work on multi-document summarization [6, 7]. Correspondingly, video summarization produces a compact version of the video (visual summary) by encapsulating the most informative parts either as a shorter video or a textual summary [8, 9, 10, 11].

With the abundance of videos uploaded online, there has been an increase in demand for efficient ways to search and retrieve relevant videos [12, 13, 14]. Cross-modal search applications often rely on text metadata associated with the video to find relevant content, but this is often missing or cannot represent subtle differences in related videos [13]. More importantly, the speech modality, which contains detailed information about the video, is not leveraged due to lack of availability of similar representation methods as for text or video.

Prior work has studied multimodal summarization [11] with speech as just an auxiliary modality. Speech summarization has been approached via pipeline models that first perform speech-to-text to convert spoken language into automatic speech recognition based predicted text, followed by textual summarization approaches mentioned above [15, 16, 17, 18]. Although this approach is widely used currently, there are considerable drawbacks to this approach such as the compounding of errors across ASR [19] and the loss of speech and audio-based context such as the prosody, audio events, speaker information, etc [20].

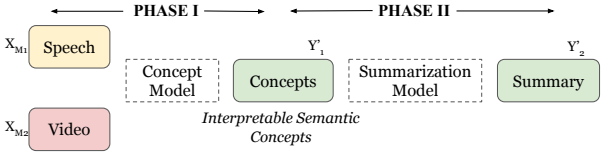

Figure 1: *Two phase cascaded model for speech summarization via multimodal semantic concept learning.*

Semantic concept learning is similar to other multimodal categorization tasks such as object recognition [21, 22], scene recognition [23, 24], action recognition [25, 26], or video tagging [27, 28]. Most of these methods are classification-based, with labels often human annotated from scratch or cleaned-up.

Instead of humanly-interpretable categories, there has also been work on latent semantic representation learning, especially to train general purpose embeddings for other downstream tasks [29, 30, 31]. While these embeddings have shown strong performance on downstream tasks [32, 31], being latent and non-observable, they do not provide the necessary controllability for generation-based tasks like summarization.

Text generation from multimodal data involves tasks such as video captioning [33, 34], question answering [35, 36], or summarization [37, 11] aimed towards generating a shorter, compressed textual description as compared with video transcription (ASR). Speech summarization by itself has largely been a unimodal effort that often represents input speech by auto-generated ASR transcripts [15, 16, 17, 18].

In this work, we address these two problems: (1) generating interpretable and semantically relevant concept representations for given speech (or video), Phase I, and (2) improvement over the pipeline approach for speech summarization via grounding in the said interpretable semantic concepts (cascaded model), Phase II. Figure 1 shows a pictorial depiction of the cascaded multimodal speech summarization model, trained in two phases.

2. Task Formulation

Figure 2 shows the flowchart of the proposed approach. Inputs to the *Concept Extraction Model* are the various modalities: images, videos, and speech. The outputs of this model are either the *Specific* or the *Abstract* concepts. These are then inputs to the *summarization model* which is used to generate a natural language video summary. To curate labels for semantic concept extraction, the annotated text is passed through a *concept curator*. Specific concepts are obtained from the human-annotated video transcript and Abstract concepts from the human-annotated abstractive summary.

http://dx.doi.org/10.21437/Interspeech.2021-1923

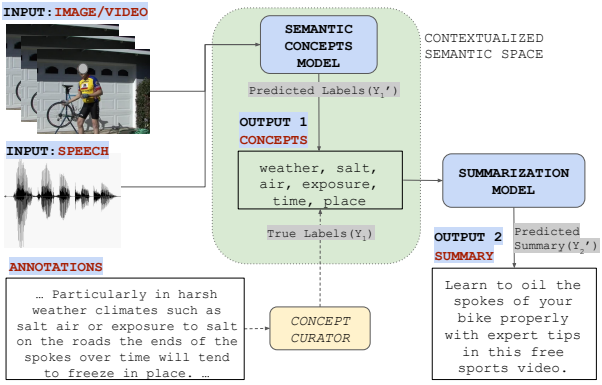

Figure 2: *Figure shows the system flowchart. The concept curator curates the groundtruth concepts (true labels Y_1) from existing annotated text. Speech and Video modalities are inputs to the Semantic Concepts Model that predicts concepts (Y_1'). The Summarization model takes the predicted concepts as inputs to generate a natural language Video Summary (Y_2').*

Specific concepts are semantically-rich content words that represent low-level fine-grained details of the task. These are curated from the utterance-level transcript of the video. These are highly domain-specific and their vocabulary may contain rare but important domain words (e.g. tensioning, spoke, etc.).

Abstract concepts are higher-level coarse-grained concepts that broadly represent the contents of the video, i.e. oil, spokes, bike, sports. These are curated from the human-annotated video summaries and are more generic and topic-based.

Video Summary is the natural language summary that provides a single sentence overview of the video. This summary consists of information gathered from all video modalities including speech, video, and text transcript.

2.1. Groundtruth Concept Curation

For the proposed concept extraction task, we use automatic methods to curate Specific and Abstract concept labels as collecting human annotation at this scale and granularity is expensive and difficult to standardize across annotators. For multimodal data such as videos that have multiple views of the same information, some information might be repeated (hence redundant) across modalities, for e.g. no necessity for captions if you understand the spoken language. This information redundancy brings forth an opportunity to automatically create groundtruth labels for tasks where data annotation is expensive. Part-of-speech tags in a sentence form a major portion of meaningful and domain-specific actions and content words, as used in [28]. We use spaCy to extract nouns and noun phrases as that worked best for both Specific (short segmented speech utterances) and Abstract concept modeling (long textual summaries of the video).

3. Models

We develop a cascaded input-to-concept and concept-to-summary model in two phases. Phase I is the concept generation model that takes various multimodal inputs. Phase II is the concept-to-summary model that takes as inputs the concepts generated in Phase I. Phase I Concept models are modality-specific fusion models that are trained towards generating contextual semantic concepts. Phase II Summarization models are text-to-text generation models for unstructured to structured generation.

3.1. Phase I: Concept Models

Speech-to-Concept (S2C) We use a Bidirectional Long Short Term Memory [38] encoder with pyramidal subsampling and an attention decoder (LSTM) for concept generation with speech inputs [39, 40]. This model learns to directly map speech to corresponding concepts, extending the work on direct acoustic-to-word speech recognizers [41] to directly generate concepts from speech (much like spoken language understanding). Speech inputs are very dense sequence vectors; using a pyramidal BiLSTM encoder converts low-level speech signals into higher-level features with input subsampling, a common technique for sequence-based speech models [40]. Additionally, we also use weights from a pre-trained acoustic-to-word speech recognizers in a transfer learning approach to boost speech-based model performance for direct speech-to-concept mapping.

Visual Adaptive Training (VAT) Visual Adaptive Training is a multimodal adaptation model to combine the speech and vision modalities. We adapt the S2C model using the VAT strategy previously applied to multimodal speech recognition [42, 43]. The VAT model learns an embedding shift transform between the low-level input speech signals and the corresponding video features leading to a transformed low-level audio-visual multimodal signal, which then proceeds through the pyramidal BiLSTM encoder (and decoder) of the S2C model. The VAT submodule is trained end-to-end with the S2C model.

VideoRNN The given sequence of features from multiple frames for every utterance or video are represented into higher-level features using a BiLSTM encoder.

Hierarchical Attention (HierAttn) Hierarchical Attention is a multimodal adaptation model to combine text and vision modalities applied to machine translation [44] and summarization [11]. In this model, there are separate BiLSTM encoders for each input, with an encoder-specific attention layer for each. This is followed by another attention layer, the hierarchical attention layer, applied on top of the encoder-specific attention layers, generating a multimodal context vector. Via this the hierarchical attention layer learns to weigh each input modality. The output of this hierarchical attention layer is fed into an LSTM decoder.

Pred. Text-to-Concept (S′2C) For video-level concept generation, the speech lengths are too long to build a single S2C model. For current computational limitations, we represent long speech by predicted text (S′) using an off-the-shelf ASR [45]. This is the predicted-text-to-concept model (S′2C).

3.2. Phase II: Summarization Models

S2S This model takes the Specific and Abstract concepts predicted by concept extraction models as inputs and converts them into a natural language summary (video Summary in Figure 2). Outputs of the semantic concept extraction model pass through a standard BiLSTM encoder followed by an attention layer and an LSTM decoder [39].

4. Experimental Setup

4.1. Dataset

The How2 dataset [46], statistics in Table 1, is an open-source open-domain instructional videos corpus that contains 4 paral-

Table 1: *Table shows dataset statistics, available modalities, vocabulary and average number (#) of concepts for How2-300h and How2-2000h datasets. Note the large target vocabulary space, which is uncommon for such tasks.*

Dataset	Concept	Modalities	Split	Samples	Vocab	Avg. #
How2-300h-Utt	Specific	Speech, Image, Transcript	Train	184,286	**9,014**	2.9
			Test	2,361	-	3.0
How2-300h-Video	Abstract	Speech, Video, Transcript, Summary	Train	13,172	**2,611**	5.9
			Test	127	-	5.6
How2-2000h-Video	Abstract	Speech, Video, Transcript, Summary	Train	73,993	**5,227**	5.9
			Test	2,156	-	5.8

Table 2: *Abstract concepts generation on the How2-300h-Video and How2-2000h-Video data.*

		How2-300h-Video			How2-2000h-Video		
Inputs	**Models**	**P**	**R**	**F1**	**P**	**R**	**F1**
Pred. Text	S′2C	16.4	40.6	23.5	52.5	57.3	54.8
Video	VideoRNN	40.8	46.7	43.6	58.6	**64.0**	61.2
Pred. Text + Video	HierAttn	**47.9**	**47.0**	**47.4**	**66.2**	63.2	**64.7**

lel modalities: speech, video, human-annotated transcription, and a summary. In this work, we use the following subsets of How2: How2-300h-Utt, How2-300h-Video, and How2-2000h-Video. The How2-300h-Utt contains speech utterances, corresponding transcripts, and images, used for modeling Specific concepts. How2-300h-Video is the 300h video equivalent used for modeling Abstract concepts, which is then scaled to the larger How2-2000h-Video. The dataset has a large vocabulary of 9,014 Specific concepts and 2611 (300 h), and 5,227 (2000 h) Abstract concepts[1].

4.2. Multimodal Features

Speech The speech features are extracted as dense time-series data following the standard feature extraction pipeline [47]. We extract 80-dimensional filterbank features and 3-dimensional pitch features for every frame of the utterances sampled at 30 frames/second.

Video [48] propose a 3-dimensional version of the traditional ResNet-101 model [49], 3D ResNeXt, with a third dimension of convolution that represents the sequential video information. The network is trained with the Kinetics Human Action Video dataset [26]. From 3D ResNeXt, we extract a 2048-dimensional vector for every keyframe.

Text Predicted Text is generated through the widely used off-the-shelf English speech recognizer, the ASPIRE model [45]. This is an out-of-domain speech recognizer trained for utterance-level prediction. Utterances are decoded independently and concatenated to create video-level transcript.

4.3. Evaluation

We evaluate the quality of the concepts as well as summaries. For concept evaluation, Precision (P), Recall (R), and F1 metrics are reported. This is at the corpus level to remove any input/output length variation dependencies. For summaries, we use standard text generation metrics: METEOR [50] and ROUGE [51].

[1]This vocabulary size for this task is much larger than prior work in speech/image/video classification tasks.

Table 3: *Specific concepts generation using the How2-300h-Utt data. † represents pre-trained model initialization.*

Inputs	Models	P	R	F1
Video	VideoRNN	13.3	4.7	7.0
Speech	S2C	26.0	21.5	23.5
Speech	S2C †	62.8	62.7	62.7
Speech + Video	VAT †	**66.6**	**64.7**	**65.7**

5. Results & Discussion

Table 3 contains results for Specific concept generation on the How2-300h-Utt dataset. Speech-based S2C model outperforms the video-based VideoRNN model on all metrics. The S2C model achieves a huge boost in performance by transfer learning using a pre-trained ASR. On top of this improvement, the VAT model results in further improvement of 3 F1 points (absolute). As speech is a noisy signal in the How2 dataset, grounding with the vision modality improves performance.

Table 2 shows the Abstract concept generation at video-level on the How2-300h-Video and How2-2000h-Video sets. The video-only model for Abstract concept generation performs much better with the video-level context instead of utterance-level. Overall, the Hierarchical Attention (HierAttn) model for predicted text and video achieves significantly higher performance than either modality by itself. Using more training data with How2-2000h-Video boosts the performance of all models while maintaining the same trends as the How2-300h-Video set.

Table 4 contains results for Specific and Abstract concept to summary generation on the How2-300h-Video data. For the summarization task, our two baselines are, (1) a language model (LM) trained on the groundtruth summaries, as done in prior work [11], and (2) a strong sequence-to-sequence (S2S) abstractive summarization model [11] which takes the complete ASR predicted video transcript as the input and summarizes it without any intermediate concepts. Both Specific and Abstract concept models outperform the LM baseline significantly. The VideoRNN and HierAttn Abstract concepts to summary models outperform LM as well as the abstractive summarization

Table 4: *Summarization with Specific or Abstract concept models evaluated via METEOR (MET) and ROUGE-L (RG-L).*

Concept	Model	MET	RG-L
None	LM [11]	15.0	32.3
None	S2S - Pred. Text [11]	22.9	46.1
Specific	S2C †	20.9	43.6
Specific	VAT †	21.8	45.9
Abstract	S′2C	21.2	44.6
Abstract	VideoRNN	24.3	49.2
Abstract	HierAttn	**30.4**	**51.2**

Table 5: *Example model outputs showing utility of interpretable intermediate semantic concepts.*

Model	Output Concepts/Summary
Groundtruth	side, stretch, exercise, video
S′2C	exercise, **fitness**, **trainer**, video
VideoRNN	**press**, exercise, video
Groundtruth	learn a side stretch exercise with small weights for your pilates routine in this free exercise video .
S2C	learn how to do the weekend yoga pose with tips from a **fitness** instructor in this free yoga lesson video .
VAT	learn more about this exercise with tips from a **fitness instructor** in this free exercise video .
S′2C	learn how to do pilates exercise with tips from a **fitness trainer** in this free exercise video .
VideoRNN	learn a chest **press** exercise with tips from a pilates instructor in this free exercise video .

baseline by a significant margin (7.5 points on METEOR and 5.1 points absolute on ROUGE-L), demonstrating the benefit of modeling concept generation as an intermediate task. A significant improvement in METEOR with this model suggests the generation of creative summaries, containing semantically relevant words, which might not be present in the groundtruth, but learned through the contextualized semantic space.

Table 5 shows model outputs for certain concept generation and summarization models. In the S′2C and VideoRNN models, concept models predict novel and semantically relevant concepts such as *press, fitness, trainer* which are not in the groundtruth but match the topic of the video. Similarly, in other examples, we see higher word diversity in generated summaries by using the learned semantic concepts as inputs for e.g. *chest press exercise, pilates exercise, weekend yoga pose*, etc.

6. Conclusion

We presented a cascaded multimodal abstractive speech summarization model that learns semantic concepts as an intermediate step before summarization. We demonstrate strong performance of this model compared to prior work on abstractive speech summarization and observe significant gains in the automatic summarization evaluation metrics. We evaluate the intermediate concepts as well and find consistent gains with using multimodal inputs rather than unimodal. Upon analysis, we find the interpretable intermediate concepts help generate more creative summaries.

7. References

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[2] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 68–73.

[3] I. Mani, *Advances in automatic text summarization*. MIT press, 1999.

[4] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 379–389.

[5] R. Nallapati, B. Zhou, C. dos Santos, Ç. glar Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *CoNLL 2016*, p. 280, 2016.

[6] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[7] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. R. Radev, "Graph-based neural multi-document summarization," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017, pp. 452–462.

[8] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[9] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[10] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux, and R. Ptucha, "Semantic text summarization of long videos," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 989–997.

[11] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6587–6596. [Online]. Available: https://www.aclweb.org/anthology/P19-1659

[12] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 423–432.

[13] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 975–985, 2012.

[14] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Learning joint representations of videos and sentences with web image search," in *European Conference on Computer Vision*. Springer, 2016, pp. 651–667.

[15] P. Manakul, M. J. Gales, and L. Wang, "Abstractive spoken document summarization using hierarchical model with multi-stage attention diversity optimization," *submission to Interspeech, Shanghai*, 2020.

[16] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2190–2196.

[17] C. Zhu, R. Xu, M. Zeng, and X. Huang, "A hierarchical network for abstractive meeting summarization with cross-domain pretraining," *arXiv preprint arXiv:2004.02016*, 2020.

[18] D. Rezazadegan, S. Berkovsky, J. C. Quiroz, A. B. Kocaballi, Y. Wang, L. Laranjo, and E. Coiera, "Automatic speech summarisation: A scoping review," *arXiv preprint arXiv:2008.11897*, 2020.

[19] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "Asr error correction and domain adaptation using machine translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6344–6348.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[23] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[24] J. Ray, H. Wang, D. Tran, Y. Wang, M. Feiszli, L. Torresani, and M. Paluri, "Scenes-objects-actions: A multi-task, multi-label video dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 635–651.

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[27] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[28] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 046–12 055.

[29] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[30] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3591–3600.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13–23.

[32] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv preprint arXiv:1909.11740*, 2019.

[33] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[34] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.

[35] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[36] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[37] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, C. Zong *et al.*, "Msmo: multimodal summarization with multimodal output," 2018.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 3104–3112.

[40] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[41] S. Palaskar and F. Metze, "Acoustic-to-word recognition with sequence-to-sequence models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 397–404.

[42] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach." in *Interspeech*, 2016, pp. 3414–3418.

[43] O. Caglayan, R. Sanabria, S. Palaskar, L. Barraul, and F. Metze, "Multimodal grounding for sequence-to-sequence speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8648–8652.

[44] J. Libovický and J. Helcl, "Attention strategies for multi-source sequence-to-sequence learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 196–202.

[45] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 539–546.

[46] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NIPS, 2018.

[47] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[48] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[51] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 605–612.