CMU Sinbad's Submission for the DSTC7 AVSD Challenge

Ramon Sanabria*, **Shruti Palaskar*** **and Florian Metze**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213; U.S.A.

Abstract

Audio-Visual Scene-Aware Dialog (AVSD) is understood as an extension of Visual Question Answering, the task of generating a textual answer in response to a textual question on multi-media content. The input typically consists of text features (either speech recognition output, or a summary describing the video contents), video features (object, scene, and/ or action features), and dialog history or context. In this paper, we describe our submission to the AVSD track of the 7^{th} Dialog State Tracking Challenge. We use hierarchical attention to fuse contributions from different modalities, and investigate transfer learning using a background corpus of 2,000 hours of how-to videos. Our approach uses dialog context, but we do not use dialog history explicitly. Our system achieves the best performance in both automatic and human evaluations.

1 Introduction

The goal of the Audio-Visual Scene-Aware Dialog (AVSD) task is to automatically answer questions about a visual stream (*i.e.*, videos or images). To do so, the algorithm needs to take into consideration the visual modality and the textual question to estimate the correct answer.

Motivated by the inherent multimodal integration that humans do, multiple language processing communities started considering more than one modality in their approaches. Some examples are (Palaskar, Sanabria, and Metze 2018) in Automatic Speech Recognition (ASR), (Specia et al. 2016) in Machine Translation (MT) and (Das et al. 2017) in Question Answering. Specifically, in AVSD, multimodality plays an important role because systems need to properly combine all modalities (*e.g.*, text, audio and video) to generate correct and fluent responses, while integrating information from all modalities. (Multi-modal) Sequence-to-Sequence (S2S) models are conceptually very simple, yet they outperformed traditional approaches in many language processing task such as QA (Lu et al. 2016) and MT (Specia et al. 2016). S2S model offer a great variety of options to integrate multimodal representation from different sources.

In this paper, we present Sinbad's systems for the audio-visual track of DSTC7 (Alamri et al. 2017; 2018), and an analysis of modality integration of different VQA S2S models. It is important to note that we did not use any ordering information provided in the challenge data (e.g., the order in which the questions were presented), so one could argue that our technical approach performs VQA rather than "dialog" tracking. The most important finding of our experiments is that multimodal integration (slighty) improves VQA, and our results support five additional interesting findings. *First*, hierarchical attention (Libovický and Helcl 2017; Hori et al. 2017) is the best mechanism to combine modalities. *Second*, visual features extracted from a 3-dimensional version of the traditional ResNet-101 trained for action recognition features are the most helpful representation for the visual modality. *Third*, we are able to perform VQA competitively by only providing the visual modality (*i.e.*, no text in the input). *Fourth*, by pretraining our model with a different task (*i.e.*, summarization) with a dataset from a different domain (*i.e.*, instructional videos from YouTube) our model slightly outperforms in-domain data-only models. *Finally*, automatic and manual rankings seem consistent for our models, with a hierarchical attention model with three-dimensional ResNet (Hara, Kataoka, and Satoh 2018) action features being ranked best overall.

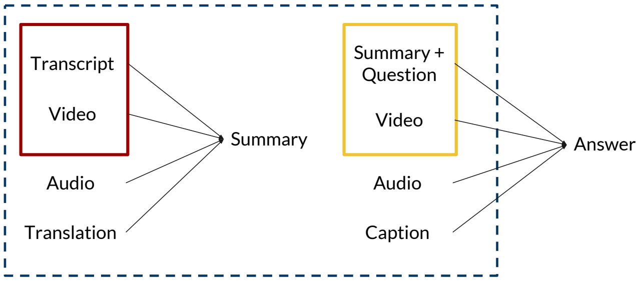


Figure 1: Our best performing model use the weights of a trained summarization model on the How2 dataset (left) to initialize the training of our DTSC7 challenge model (right).

2 Models

Baseline The multimodal baseline for this work is the Naiv̈e Fusion proposed in (Yu et al. 2016). Naiv̈e Fusion combines all modalities with a projection matrix. This pro-

[*] Equal contribution. Ordering determined by dice rolling.

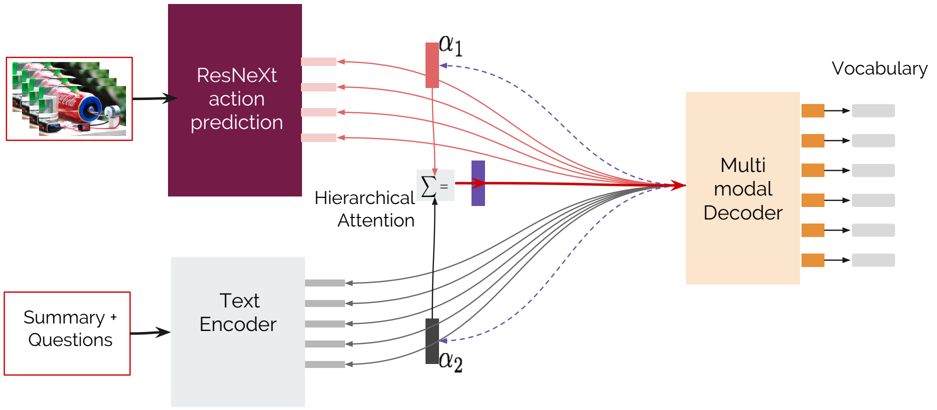


Figure 2: Text-and-Video Dialog generation models with Hierarchical Attention.

jection matrix maps a vector that contains all modalities concatenated to a target dimension.

Video-RNN This model, detailed in (Libovický et al. 2018), only uses video as input modality and use the same architecture as in (Bahdanau, Cho, and Bengio 2014). More specifically, each frame of the video is represented as a vector and the sequence of vectors are fed into a bidirectional RNN. Then, a decoder applies attention to the sequence of encoded frame.

Hierarchical Attention For multimodal summarization we follow the hierarchical attention approach (Hori et al. 2017; Libovický and Helcl 2017) to combine textual and visual modalities. The model computes the context vector independently for each of the input modalities. In the next step, the context vectors are treated as hidden states of another encoder and a new vector is computed. The hierarchical attention computation is shown in Figure 2. This type of attention mechanism performed better for multimodal summarization than a text-only model (Libovický et al. 2018) (see Section 4).

3 Related Work

First approaches on Image QA (IQA) were done in a very limited environment and with unrealistic data (Geman et al. 2015). Antol *et al.* proposed the first approach on IQA in a real-world scenario (Antol et al. 2015). In (Antol et al. 2015), the authors released a dataset where they collected questions about images of realistic scenes. In addition, Antol *et al.* also proposed some IQA classification-based baseline models. Inspired by (Bahdanau, Cho, and Bengio 2014), Xu *et al.* proposed to use attention for image captioning. Attention allowed the model to focus dynamically on different parts of the image (Xu et al. 2015). More recently, Zhu *et al.* ported the idea on (Xu et al. 2015) and applied spatial attention to a QA model (Zhu et al. 2016). The main difference of the mentioned approaches and our work is that we attend to multiple modalities while they only focus on the image. Also, our visual stream is video-based instead of image-based.

Yu *et al.* presented an approach to model dialog conditioned on the video (Yu et al. 2016). In this case, their approach used concatenation to combine the different modalities. In (Hori et al. 2017; Libovický and Helcl 2017), Libovický and Helcl and *Hori et al.* presented Hierarchical Attention, a technique explained in Section 2. Hierarchical Attention offers a solution to attend to multiple modalities and combine both modalities again with attention. This technique obtained strong results in multimodal MT (Libovický and Helcl 2017), video description (Hori et al. 2017) and video summarization (Libovický et al. 2018). Another approach to jointly attend to multiple modalities is co-attention (Lu et al. 2016).In this case, the co-attention mechanism performs question-guided visual attention and image-guided question attention. We will leave the co-attention model for future work.

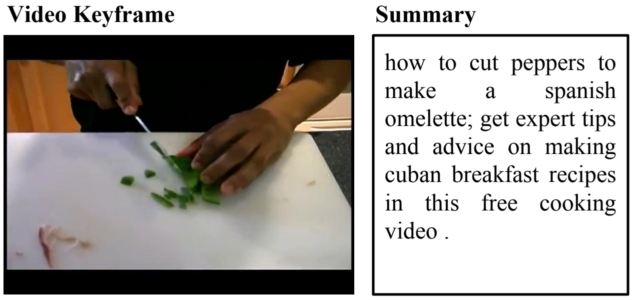
4 Experiments

To test the architectures explained in Section 2, we use the implementations developed during the Jelinek Memorial Summer Workshop 2018 (Caglayan et al. 2017). We evaluate and compare our approaches with the baselines described in Section 2 provided by the organizers of DTSC7[1]. For training, validation and testing we use the subsets defined in Table 1. The best-performing systems were evaluated by the organizers using an undisclosed evaluation test set of 6745 questions and 1710 videos.

4.1 Data

In this work, we use two datasets. The first dataset, collected by the DSTC7 organizers, is composed by crowdsourced dialogues conditioned on videos from the Charades3 dataset (Sigurdsson et al. 2016). The second one is a recently released multimodal multitask dataset of instructional videos called How2 (Sanabria et al. 2018). Table 1 summarizes the amount of data for each dataset.

[1] `https://github.com/dialogtekgeek/AudioVisualSceneAwareDialog`



how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

Figure 3: How2 dataset example with different modalities. “Cuban breakfast” and “free online video” is not mentioned in the transcript, and has to be derived from other sources.

Charades DSTC7 organizers crowdsourced human annotated questions, answers, captions, and summaries from videos belonging to the Charades dataset (Sigurdsson et al. 2016). The original videos of this dataset contain untrimmed and multi-action videos. In the DSTC7 dataset, each video has ten questions and answers pairs. The dataset statistics for training, validation, disclosed test and undisclosed evaluation test set are given in Table 1. More details about the dataset collection are described in (Alamri et al. 2017; 2018; Sigurdsson et al. 2016).

How2 How2 (Sanabria et al. 2018) is a large-scale dataset of instructional videos. It covers a wide variety of topics in 2,000 hours of clips. It provides word-level time alignments and the ground-truth English subtitles. It also provides human-generated summaries of the videos and crowdsourced Portuguese translations of the subtitles based on the video. In this work, we will only use the subtitles, the summaries, and the videos to train a summarization model to use its weights to initialize the training of a VQA system.

	Charades		How2
Split	Sentences	Videos	Videos
train	76590	7659	73993
val	17870	1787	2965
test	7330	733	2156
held_out	6745	1710	169*

Table 1: Dataset statistics for Charades and How2. The number of videos in the held_out test set of How2 is from the 300 hours subset of the data (*).

4.2 Multimodal Features

To fully exploit the information provided in the videos we extract different representations from each modality. To do so, we use DNNs trained for a particular task to extract their internal representation. Based on empirical observations, we know that pretrained DNNs capture specific characteristics to solve a specific task. Therefore we use DNNs trained for object recognition, place recognition, action recognition, and audio event detection to extract a meaningful representation of the data. We hypothesize that each of this features will capture information of the video that will be useful to answer each question.

Object Features These features are an intermediate representation of a CNN ResNet-50 trained with the ImageNet dataset (Deng et al. 2009)[2]. ImageNet is a dataset for object recognition with more than one million of images annotated with one thousand classes.

Place Features (Nallapati et al. 2016) extract scene feature representations from a static image. In this case, the network is trained to recognize scenes from an image. More specifically, (Nallapati et al. 2016) trained the network with Place365 dataset that contains 10 million images comprising with more than 400 classes.

I3D Flow (Carreira and Zisserman 2017) are video features extracted from an spatiotemporal CNN architecture trained for action recognition. The network is trained to recognize 400 different human actions. (Carreira and Zisserman 2017) use a optical flow representation of the Kinetics Human Action Video dataset that contains 400 samples for class. We extract a 2048 dimensional representation from the `Mixed_5c` layer.

I3D RGB I3D RGB is also a video feature from (Carreira and Zisserman 2017) but instead of using optical flow, the network uses video frames with three channels as the input stream.

3D ResNeXt (Hara, Kataoka, and Satoh 2018) is a 3-dimensional version of the traditional ResNet-101. The third dimensionality of the convolution allows us to extract features from the video instead of an image. The network, similar I3D RGB and I3D Flow, is trained with the Kinetics Human Action Video dataset. From 3D ResNeXt, we extract a 2048 dimensional vector. These representations are shown in Figure 2.

VGGish (Hershey et al. 2017) are audio features that have been extracted from a CNN to perform audio even detection network. The network architecture is inspired by the traditional image classification network: VGG. This network works with log Mel spectrograms features extracted from 16 KHz audio recordings. The network was trained with 70M training videos (5.24 million hours) with a total of target 30,871 labels. We use a 128-dimensional embedding.

4.3 Transfer Learning with How2 Dataset

There are many common modalities between the Charades dataset and the How2 dataset as described in Section 4.1. To exploit this fact and increase the training data for this

[2] `https://github.com/KaimingHe/deep-residual-networks`

task, we first train models on the How2 data and then fine-tune (FT) them on the Charades dataset. The pipeline and respective input modalities are shown in Figure 1. The models trained on How2 data use transcription of video (and/or video features) in the input and generate an abstractive textual summary of the video in the output. The methods used for training these are described in (Libovický et al. 2018). We initialize the training of a sequence-to-sequence model for the Charades data with the weights of this learned model, using summary+question (and/or video features) in the input and generating the answer in the output. While fine-tuning, we share the vocabulary for the two datasets and randomly initialize words that do not occur in both.

Although the two datasets have the same modalities, there are differences in the outputs. The main difference between the two datasets is that the summaries of the How2 dataset (usually 2-3 sentences) follow a particular pattern or template as described in (Sanabria et al. 2018), while the pattern of answers (usually single sentence) in the DSTC7 dataset are more stochastic. The input to the summarization model is the video, that is longer in duration than the videos in the Charades dataset and the transcript which. We will observe the effects of these differences in Section 4.

4.4 Experimental Setup

In all our experiments, the text encoder consists of 2 bidirectional layers of encoder with 256 Gated Recurrent Units (GRU) (Cho et al. 2014) and 2 layers of decoder with Conditional Gated Recurrent Units (CGRU) (Firat and Cho 2016). The models are optimized with an Adam Optimizer (Kingma and Ba 2014) with learning rate $4 \cdot 10^{-4}$ halved after each epoch when the validation performance does not increase. We use separate vocabulary for source and target text and restrict it to containing words that occur at least 5 times. We add 4 extra tokens for padding, start, and end of sentence, and unknown words (*'pad', 'bos', 'eos' and 'unk'*) to the dictionary. We restrict the maximum input length to 200 tokens. We train the model for 30 epochs but choose the checkpoint with the best performance on ROUGE-L score on the validation data. For summarization models with How2 dataset, we follow the same training regime as in (Libovický et al. 2018). During fine-tuning, we share the vocabulary of How2 and Charades dataset, leading to 30440 words in total.

Evaluation We evaluated our models using the metrics and toolkit[1] proposed by the competition organizers. We report the common natural language processing metrics like BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin and Och 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). In addition to these objective evaluation metrics for this task, the organizers also evaluated some models on crowdsourced human scores. The human evaluators were asked to score model outputs based on how semantically, grammatically and factually correct the generated answers are.

4.5 Results

Table 2 presents our different models trained using Charades and How2 data, and using various modalities one at a time (text-only, video-only) or together (text-and-video). First, we report the baseline results using the model architecture and code-base provided by the competition organizers (Alamri et al. 2017). We replicate their results using I3D RGB, I3D Flow and VGGish features. To compare the performance of different visual features, we use Objects, Places and 3D ResNet in the baseline and observe similar or slightly worse performance showing all features are equally rich representations.

For text-only models (models 7 and 8), the input is a concatenation of summary of the video followed by the question. The summary is repeated for every question following the assumption that it has relevant input information for each question. Further, we will see that improvements in the text-and-video models over text-only models show that only using the summary in the input may not be sufficient and visual features are useful in such scenarios. In the video-only models (models 9 and 10), we observe lower performance than the text-only model as expected. It is interesting to note that the video-only model is worse only by about 3-4 ROUGE-L points than the text-only model showing the richness of the visual features, 3D ResNet, here. In text-and-video models (models 11-15), we use Hierarchical attention for multimodal adaptation with different visual features. We observe that 3D ResNet performs the best for adaptation models.

We fine-tune each of these models on summarization models trained using the How2 data. In the text-only (model 8) and video-only (model 10) we see substantial gains using fine-tuning over models trained only on Charades data. For the text-and-video model, the gains are not too high, and further exploration of this behavior is needed.

Table 3 shows the 4 best models from Table 2 which we submitted to the challenge. These were evaluated on the undisclosed evaluation test set by the organizers. The baselines (model 1 and 2) are same as those in the previous table but evaluated on the undisclosed test set. The trends we observe on the prototype test set are same as those observed on the undisclosed test set. Additionally, this table also contains the human evaluation scores. The evaluators were asked to rate even the groundtruth references which scored 3.938. Our best model scores 3.491 while the baseline scores 2.848. This further shows that our models score well not only in quantitative scores but also in qualitative scores.

5 Qualitative Analysis

We perform certain qualitative analysis of the different models: text-only, video-only and text-and-video, each with fine-tuning, to better understand the quality of the results and the model behavior. We compute the number of unique words in the answers generated by each of the three models. We observe that multimodal fusion and fine-tuning with How2 both help increase the number of unique words. Another metric we use is the average length of outputs (avg.). Fine-tuning leads to longer outputs in text-only and video-only models. These models also led to higher gains over Cha-

		BLEU						
Sr. No.	**Description**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**METEOR**	**ROUGE-L**	**CIDEr**
	Input: Text and Video (different features), Model: Baseline (Alamri et al. 2017)							
1	Charades & I3D RGB & I3D Flow	0.273	0.173	0.118	0.084	0.117	0.291	0.766
2	Charades & I3D RGB & I3D Flow & VGGish	0.271	0.172	0.118	**0.085**	0.116	**0.292**	**0.791**
3	Charades & Objects	0.272	0.173	0.117	0.083	**0.118**	0.287	0.742
4	Charades & Places	0.269	0.171	0.116	0.082	0.116	0.286	0.727
5	Charades & 3D ResNet	0.264	0.166	0.112	0.079	0.116	0.284	0.711
6	Charades & 3D ResNet & Objects & Places	**0.276**	**0.176**	**0.120**	**0.085**	**0.118**	0.287	0.752
	Input: Text Only, Model: S2S							
7	Charades	0.297	0.200	0.142	0.105	0.138	0.330	1.079
8	How2 `FT` Charades	**0.311**	**0.212**	**0.152**	**0.114**	**0.146**	**0.337**	**1.169**
	Input: Video Only (3D ResNet features), Model: Video-RNN							
9	Charades	0.264	0.170	0.118	0.085	0.116	0.294	0.804
10	How2 `FT` Charades	**0.279**	**0.179**	**0.122**	**0.086**	**0.122**	**0.300**	**0.833**
	Input: Text and Video (different features), Model: Hierarchical Attention							
11	Charades & Objects	0.274	0.179	0.125	0.091	0.121	0.301	0.876
12	Charades & Places	0.287	0.191	0.136	0.101	0.133	0.320	1.036
13	Charades & VGGish	0.303	0.206	0.148	0.110	0.144	0.338	1.150
14	Charades & 3D ResNet	**0.306**	**0.209**	**0.150**	**0.112**	**0.144**	**0.338**	**1.161**
15	How2 `FT` Charades & 3D ResNet	**0.307**	**0.210**	**0.151**	**0.113**	**0.145**	**0.339**	**1.180**

Table 2: Automatic evaluation metrics on the test set provided by the organizers (groundtruth available). Models 1-6 are trained using the methods described in (Alamri et al. 2017) with different modaltities. We treat them as our baselines. Models 7 and 8 are trained on text-only, models 9 an 10 on video-only and models 11-15 on text-and-video. Models 8, 10 and 15 are first trained on the How2 data and then fine-tuned `FT` on the Charades data.

		BLEU							**Human**
ref. Table 2	**Description**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**METEOR**	**ROUGE-L**	**CIDEr**	**Rating**
	Input: Text and Video (different features), Model: Baseline (Alamri et al. 2017)								
1	Charades & I3D RGB & I3D Flow	0.621	0.480	0.379	0.305	0.217	0.481	0.733	-
2	Charades & I3D RGB & I3D Flow & VGGish	0.626	0.485	0.383	0.309	0.215	0.487	0.746	2.848
	Input: Text Only, Model: S2S								
7	Charades *	0.692	0.555	0.447	0.364	0.254	0.543	1.006	-
8	How2 `FT` Charades *	**0.711**	**0.570**	**0.461**	**0.376**	**0.264**	**0.554**	**1.076**	3.394
	Input: Text and Video (3D ResNet features), Model:Hierarchical Attention								
9	Charades *	0.718	0.584	**0.478**	**0.394**	**0.267**	**0.563**	**1.094**	**3.491**
15	How2 `FT` Charades *	**0.723**	**0.586**	0.476	**0.387**	**0.266**	**0.564**	1.087	3.459
-	Groundtruth	-	-	-	-	-	-	-	3.938

Table 3: **Automatic and Human evaluation** scores on the undisclosed evaluation test set prepared by DTSC7 organizers (we do not have access to groundtruth). Models 1 and 2 are the same baselines as in Table 2. Models 3 and 4 are trained on text-only. Models 5 and 6 are trained on text-and-video using Hierarchical attention. Models 4 and 6 are first trained on the How2 data and then fine-tuned `FT` on the Charades data. Systems marks with an asterisk (*) were the ones submitted to the challenge. Model 6 i.e. 'How2 `FT` Charades' was the best performing model. Note that the first column has a reference number to the model in Table 2.

Sr. No.	Model	# unique words	Avg. output length	% sent. changed	% sent changed in content word
1	Text Only Charades	384	8.98	-	-
2	Text Only How2 FT Charades	726	9.23	79.46%	65.30%
3	Video Only Charades	269	9.22	83.60%	72.35%
4	Video Only How2 FT Charades	331	9.37	87.00%	74.91%
5	Text and Video Charades	488	8.95	76.37%	59.00%
6	Text and Video How2 FT Charades	740	8.98	77.72%	60.21%

Table 4: Qualitative evaluation of different systems. % sentences (sent) changed are with respect to text-only Charades model.

rades only models 2. Our final metric is the percentage (%) of sentences changed in a given system when compared with the text-only model trained only on Charades data. We compute this metric by counting all tokens changed, as well as by counting only content-based tokens, *i.e.* not counting stop words or punctuation as changed. We see the maximum percentage of changed sentences are in the video-only models. The difference percentage change by considering only content words is approximately 10-15% absolute.

6 Conclusions

In this paper, we present our submission to the Audio-Visual Scene-Aware Dialog (AVSD) track of the 7th Dialog State Tracking Challenge (DSTC7). Our final submission achieved the best performance in both human (mean opinion score) and automatic (BLEU, ROUGE, METEOR, and CIDEr) evaluation metrics.

We cast the task as a multi-modal video summarization problem, in which the input is given by video features as well as text context concatenated with the question, while the summary provides the desired “answer”. We applied hierarchical attention to fuse contributions of the text and image modalities, using an nmtpytorch implementation, which provided small improvements over the baseline. We experimented with additional visual features, which again slightly improved performance over the provided ones. However, we did not achieve significant improvements by pre-training our model on a large corpus of 2,000 hours of how-to videos at this time.

We are currently performing more analysis to understand the capabilities of this model better, document the behavior at different operating points, and extend it to tasks such as identifying and describing differences in videos.

References

Alamri, H.; Hori, C.; Marks, T. K.; Parikh, D.; and Batra, D. 2017. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. -.

Alamri, H.; Cartillier, V.; Lopes, R. G.; Das, A.; Wang, J.; Essa, I.; Batra, D.; Parikh, D.; Cherian, A.; Marks, T. K.; et al. 2018. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Caglayan, O.; García-Martínez, M.; Bardet, A.; Aransa, W.; Bougares, F.; and Barrault, L. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics* 109:15–28.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4724–4733. IEEE.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111. Doha, Qatar: Association for Computational Linguistics.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

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

Firat, O., and Cho, K. 2016. Conditional gated recurrent unit with attention mechanism. https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences* 201422953.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. 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)*, 6546–6555.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R.; and Wilson, K. 2017.

Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Hori, C.; Hori, T.; Lee, T.-Y.; Zhang, Z.; Harsham, B.; Hershey, J. R.; Marks, T. K.; and Sumi, K. 2017. Attention-based multimodal fusion for video description. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 4203–4212. IEEE.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Libovický, J., and Helcl, J. 2017. 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)*, 196–202.

Libovický, J.; Palaskar, S.; Gella, S.; and Metze, F. 2018. Multimodal abstractive summarization of open-domain videos. In *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*.

Lin, C.-Y., and Och, F. J. 2004. 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*, 605–612.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, 289–297.

Nallapati, R.; Zhou, B.; dos Santos, C.; aglar Gulehre; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Computational Natural Language Learning*.

Palaskar, S.; Sanabria, R.; and Metze, F. 2018. Visual features for context-aware speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Sanabria, R.; Caglayan, O.; Palaskar, S.; Elliott, D.; Barrault, L.; Specia, L.; and Metze, F. 2018. How2: a large-scale dataset for multimodal language understanding. In *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*.

Specia, L.; Frank, S.; Sima'an, K.; and Elliott, D. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, 543–553.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4584–4593.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4995–5004.