A Simple Baseline for Audio-Visual Scene-Aware Dialog

Idan Schwartz[1], Alexander Schwing[2], Tamir Hazan[1]

[1]Technion   [2]UIUC

idanschwartz@gmail.com, aschwing@illinois.edu, tamir.hazan@technion.ac.il

Abstract

The recently proposed audio-visual scene-aware dialog task paves the way to a more data-driven way of learning virtual assistants, smart speakers and car navigation systems. However, very little is known to date about how to effectively extract meaningful information from a plethora of sensors that pound the computational engine of those devices. Therefore, in this paper, we provide and carefully analyze a simple baseline for audio-visual scene-aware dialog which is trained end-to-end. Our method differentiates in a data-driven manner useful signals from distracting ones using an attention mechanism. We evaluate the proposed approach on the recently introduced and challenging audio-visual scene-aware dataset, and demonstrate the key features that permit to outperform the current state-of-the-art by more than 20% on CIDEr.

1. Introduction

We are interacting with a dynamic environment which constantly stimulates our brain via visual and auditory signals. Despite the huge amount of different information that is permanently occupying our nervous system, we are often easily able to quickly discern important cues from data that is irrelevant. Telling apart useful information from distracting aspects is also an important ability for virtual assistants, car navigation systems, or smart speakers. However present day technology uses a chain of components from speech recognition and dialog management to sentence generation and speech synthesis, making it hard to design a holistic and entirely data-driven approach.

For instance, in computer vision, a tremendous amount of recent work has focused on image captioning [68, 30, 11, 16, 75, 45, 77, 31, 69, 4, 15, 10], visual question generation [36, 48, 47, 28], visual question answering [5, 19, 59, 54, 44, 73, 74, 76, 57, 58, 49, 50], and very recently visual dialog [13, 14, 27, 46]. While those meticulously engineered algorithms have shown promising results in their specific domain, little is known about the end-to-end performance of an entire system. This is partly due to the fact that little data is publicly available to design such an end-to-end algorithm.

Recent work on audio-visual scene aware dialog [2, 25] partly addresses this shortcoming and proposes a novel
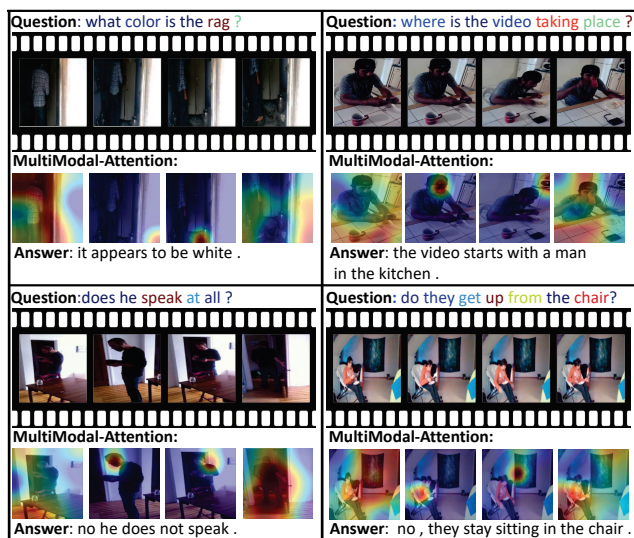

Figure 1: We present 4 different questions and the generated answer. Our attention unit is illustrated as well. Our model samples 4 frames, and attends to each frame separately, along with the question and the audio. We observe attention for each frame to differ, where first and fourth frames are widespread, while the second and third are more specific. Also, the question attention attends to relevant words. We also include the audio modality as input to the attention computation.

dataset. Different from classical datasets like MSCOCO [39], VQA [5] or Visual Dialog [13], this new dataset contains short video clips, the corresponding audio stream and a sequence of question-answer pairs. While development of an end-to-end data driven system isn't feasible just yet due to the missing speech signal, the new audio-visual scene aware dialog dataset at least permits to develop a holistic dialog management and sentence generation approach taking audio and video signals into account.

In recent work [2, 25], a baseline for a system based on audio, video and language data was proposed. Compelling results were achieved, demonstrating accurate question answering. The authors demonstrate that multimodal features based on I3D-Kinetics (RGB+Flow) [9] refined via a carefully designed attention-based mechanism improve the quality of the generated dialog.

However, since much effort was dedicated to collecting the dataset, little analysis of such a holistic system was provided. Moreover, due to tremendous amounts of available data (certainly a ten-fold increase compared to classical vi-

sual dialog data) this is by no means trivial. To provide this missing information and to share some insights with the community about how and where to improve, in this paper, we follow the spirit of [26] and demonstrate (1) that simply using the question as a signal already permits to outperform the current state-of-the-art; (2) that it is crucial to maintain spatial features for the video signal (either VGG19 [63] or I3D-Kinetics [9]). Reducing every video frame into a single representation drops performance significantly; (3) that temporally subsampling the video frames improves the accuracy; (4) that using attention over all available data (including different frames) is beneficial. To this end we analyze how to fuse the attended vectors for different data modalities.

Our simple baseline, which consists of three jointly trained components (data representation extraction, attention and answer generation) outperforms state-of-the-art by a large margin of 20% on CIDEr. Improvements of the proposed approach are largely due to the aforementioned four points. Results of generated answers are contrasted to the current state-of-the-art in Fig. 1. We observe plausible answers to many questions and attention that focuses on important parts in both video and text.

2. Related Work

A significant amount of research has been conducted regarding image captioning, visual question generation, visual question answering, visual dialog, video data, audio data and multimodal attention models. We briefly review those related areas in the following.

Image Captioning: Originally image captioning was formulated as a retrieval problem. The best fitting caption from a set of considered options was found by matching features obtained from the available textual descriptions and the given image. Importantly, the matching function is typically learned using a dataset of image-caption pairs. While such a formulation permits end-to-end training, assessing the fit of image descriptors to a large pool of captions is computationally expensive. Moreover, it's likely prohibitive to construct a database of captions that is sufficient for describing even a modestly large fraction of plausible images.

To address this challenge, recurrent neural nets (RNNs) decompose captions into a product space of individual words. This technique has recently found widespread use for image captioning because remarkable results have been demonstrated which are, despite being constructed word by word, syntactically correct most of the time. For instance, a CNN to extract image features and a language RNN that shares a joint embedding layer was trained [45]. Joint training of a CNN with a language RNN to generate sentences one word at a time was demonstrated in [75], and subsequently extended [75] using additional attention parameters which identify salient objects for caption generation.A bi-directional RNN was employed along with a structured loss function in a shared vision-language space [31]. Diversity was considered,

e.g., by Wang *et al*. [69] and Deshpande *et al*. [15].

Visual Question Answering: Beyond generating a caption for an image, a large amount of work has focused on answering a question about a given image. On a plethora of datasets [43, 54, 5, 19, 81, 29], models with multi-modal attention [41, 76, 3, 12, 18, 59, 74, 57, 58], deep net architecture developments [8, 44, 42] and memory nets [73] have been investigated.

Visual Question Generation: In spirit similar to question answering is the task of visual question generation, which is still very much an open-ended topic. For example, Ren *et al*. [54] discuss a rule-based method, converting a given sentence into a corresponding question which has a single word answer. Mostafazadeh *et al*. [48] learned a question generation model with human-authored questions rather than machine-generated descriptions. Vijayakumar *et al*. [67] have shown results for this task as well. Different from the two aforementioned techniques, Jain *et al*. [28] argued for more diverse predictions and use a variational auto-encoder approach. Li *et al*. [36] discuss VQA and VQG as dual tasks and suggest a joint training. They take advantage of the state-of-the art VQA model by Ben-younes *et al*. [8] and report improvements for both VQA and VQG.

Visual Dialog: Visual dialog [13] combines the three aforementioned tasks. Strictly speaking it requires both generation of questions and corresponding answers. Originally, visual dialog required to only predict the answer for a given question, a given image and a provided history of question-answer pairs. While this resembles the VQA task, different approaches, *e.g.*, also based on reinforcement learning, have been proposed recently [35, 14, 27, 46, 72].

Video Data: A variety of tasks like video paragraph captioning [78], video object segmentation [53], pose estimation [79], video classification [32], and action recognition [62] have used video data for a long time. Probably most related to our approach are video classification and action recognition since both techniques also extract a representation from a video. While the extracted representation is subsequently used for either classification or action recognition, we employ the representation to more accurately answer a question. Commonly used feature representations for either video classification or action recognition are I3D-based features by Carreira *et al*. [9], extracted from an action recognition dataset. With proper fine-tuning the I3D-based features proved to be better than the classical approaches, such as C3D [65] that capture spatiotemporal information via a 3D CNN. In this work, we assess a naïve feature extractor based on VGG [63], and demonstrate that for video-reasoning, careful reduction of the spatial dimension is more crucial than the type of extracted features used to embed the video frames. Wang *et al*. [70] showed that working with video frame samples, achieves not only efficiency, but also improves performance compared to a conservative dense temporal representation. Recently, Zhou *et al*. [80] further
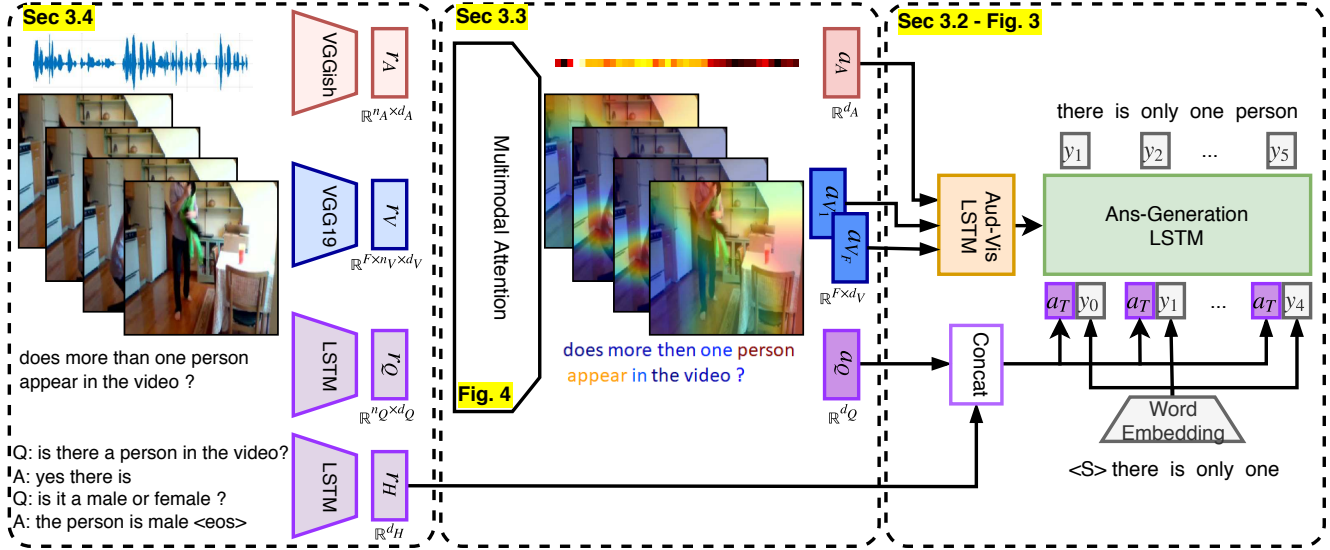
Figure 2: Overview of our approach for the AVSD task. More details can be found in Sec. 3.

extended those ideas, and suggested to capture relational temporal relationships between the sampled frames, relying on the relational-networks concept [56]. We follow those ideas by also sub-sampling a small set of frames uniformly. Our model further advances those concepts, by exploiting spatial relationships between sampled temporal frames via a high-order multimodal attention module, where each video frame is treated as a separate modality. Li *et al*. [37] propose the Video-LSTM model, which uses attention to emphasis relevant locations, during LSTM video encoding. Our approach differs in that attention on one frame can influence attention on other frames which isn't the case in their model.

Audio Data: Audio data gained popularity in the vision community recently. For instance, prediction of pose given audio input [60], learning of audio-visual object models from unlabeled video for audio source separation in novel videos [20, 51], use of video and audio data for acoustic scene/object classification [6], source separation was also considered in [17] and learning to see using audio [52].

Multimodal Attention: Multimodal attention has been a prominent component in tasks which operate on different input data. Xu *et al*. [75] showed an encoder decoder attention model for image captioning, which was extended to visual question answering [74]. Yang *et al*. [76] propose a multi-step reasoning system using an attention model. Multimodal pooling methods were also explored [18, 33]. Lu *et al*. [41] suggest to produce co-attention for the image and question separately, using a hierarchical and parallel formulation. Schwartz *et al*. [57, 58] later extend this approach to high-order attention applied over image, question and answer modalities via potentials. Similarly, in the visual dialog task, co-attention models have held the state-of-the-art [71, 40] attending over image, question and history in hierarchical manner. For audio-visual scene-aware dialog, [25] also use a sum-pooling type of attention, using the question feature
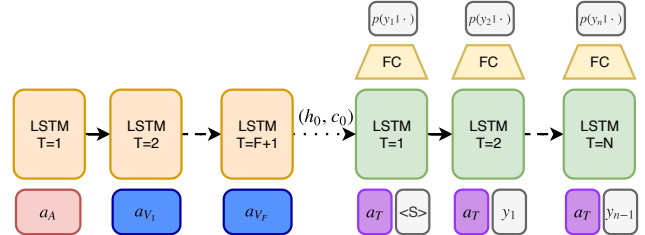

Figure 3: Our decoder for audio-visual scene-aware dialog. We start with encoding of attended audio and video vectors using the Aud-Vis LSTM (orange colored), followed by the Ans-Generation LSTM that receives the textual data concatenated with the previous answer word (green colored).

along with audio and video modalities separately. In contrast, here we compute attention over each modality via local and cross data evidence, letting all the modalities interact with each other.

3. Audio Visual Scene-Aware Dialog Baselines

Our method has three building blocks: answer generation, attention and data representation as shown in Fig. 2.

3.1. Answer Generation

We are interested in predicting an answer $y = (y_1, \ldots, y_n)$ consisting of n words $y_i \in \mathcal{Y}_i = \{1, \ldots, |\mathcal{Y}_i|\}$ each arising from a vocabulary of possible words $\mathcal{Y}_i$. Given data $x = (Q, V, A, H)$ which subsumes, a question Q, a sub-sampled video $V = (V_1, \ldots, V_F)$ composed of F frames, the corresponding audio signal A, and a history of past question-answer pairs H, we construct a probability model over the set of possible words for the answer generation task. To this end, we formulate prediction of the answer as inference in a recurrent model where the joint probability is
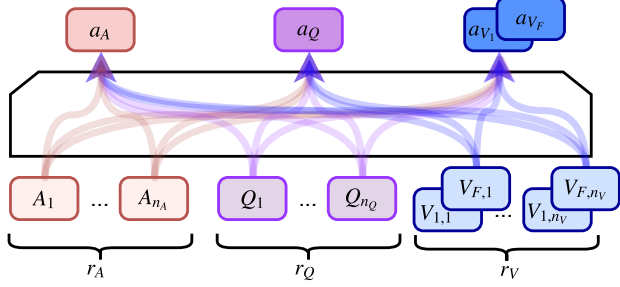
Figure 4: Multimodal Attention model for audio-visual scene-aware dialog. We treat each frame as a modality, along with audio and question modality, to total of 6 modalities. Each element attention score is affected not only from local evidence, but also via cross-data interactions of all other elements.

given by the product of conditionals, *i.e.*,

$$p(y|x) = \prod_{i=1}^{n} p(y_i|y_{<i}, x).$$

Note that, for now, we condition on all the data x for readability and provide details later. Instead of conditioning the probability of the current word $p(y_i|y_{<i}, x)$ on its entire past $y_{<i}$, we combine two recurrent nets: an audio-visual recurrent net that generates the temporal information which is fed as an initialization to the answer generating recurrent net. See Fig. 3 for a schematic.

Audio-visual LSTM-net: It operates on an attended audio embedding a_A and attended video embeddings $a_{V_1}, ..., a_{V_F}$ for each of the F frames $f \in \{1, \ldots, F\}$. This LSTM-net has $F+1$ units, the first unit's input is the attended audio vector, and the input to the F subsequent units are the attended video representations $a_{V_1}, \ldots, a_{V_F}$. The context vector that is generated from this LSTM, *i.e.*, (h_0, c_0) summarizes the audio-visual attention and is provided as input to the answer generation LSTM-net.

Answer generation LSTM-net: It computes conditional probabilities for the possible words $y_i \in \mathcal{Y}_i$ of the answer $y = (y_1, \ldots, y_n)$. This probability considers the last word and captures context via a representation h_{i-1} obtained from the previous time-step.

$$p(y_i|y_{i-1}, h_{i-1}, x) = g_w(y_i, y_{i-1}, h_{i-1}, x).$$

We illustrate the LSTM-net g_w in Fig. 3. Using the initial state (h_0, c_0), the LSTM-net g_w predicts in its i-th step a probability distribution $p(y_i|y_{i-1}, h_{i-1}, x)$ over words $y_i \in \mathcal{Y}_i$ using as input y_{i-1} and the textual attention vector $a_T = (a_Q, r_H)$: the attended textual vector is a concatenation of the attended question vector a_Q and the history vector r_H, which represents information about question and history data. The output of the LSTM-net is transformed via a FC-layer with a dropout and a softmax to obtain the probability distribution $p(y_i|y_{i-1}, h_{i-1}, x)$.

3.2. Attention

The attention step provides an attended representation for the data components, *i.e.*, $a_{V_f} \in \mathbb{R}^{d_V}$ for frame $f \in \{1, \ldots, F\}$ of the video data, $a_A \in \mathbb{R}^{d_A}$ for the audio data, and $a_T \in \mathbb{R}^{d_T}$ for the textual data. These attended representations are obtained by transforming the representations extracted from the raw data, *i.e.*, $r_{V_f} \in \mathbb{R}^{n_V \times d_V}$ for the video data, $r_A \in \mathbb{R}^{n_A \times d_A}$ for the audio data, and for the textual data, $r_Q \in \mathbb{R}^{n_Q \times d_Q}$ as well as $r_H \in \mathbb{R}^{d_H}$ which capture signals from the question and history respectively. We outline the general procedure in Fig. 4.

Formally, we obtain the attended representation

$$a_\alpha = \sum_{k=1}^{n_\alpha} \alpha_k p_\alpha(k),$$

where $\alpha \in \{A, Q, V_1, \ldots, V_F\}$ is used to index the available data components (audio, question, visual frames), n_α is the number of entities in a data component (*e.g.*, the number of words in a question), and $p_\alpha(k) \geq 0 \,\forall\alpha$ is a probability distribution ($\sum_{k=1}^{n_\alpha} p_\alpha(k) = 1 \,\forall\alpha$) over the n_α entity representations of data α. For instance, if we let $\alpha = A$ we obtain the attended audio representation $a_A = \sum_{k=1}^{n_A} A_k p_A(k)$.

We compute the attention via a factor graph attention approach [57, 58]. The attention probability distribution over a data source α consists of a log-prior distribution π_α, a local evidence l_α that relies solely on its data representation r_α and a cross data evidence c_α that accounts for correlations between the different data representations r_α, r_β, for $\beta \in \{A, Q, V_1, \ldots, V_F\}$. This probability distribution takes the form:

$$p_\alpha(k) \propto \exp\left(\hat{w}_\alpha \pi_\alpha(k) + l_\alpha(k) + c_\alpha(k)\right).$$

The local evidence is $l_\alpha(k) = w_\alpha\left(v_\alpha^\top \mathrm{relu}(V_\alpha \alpha_k)\right)$, the log-prior is $\pi_\alpha(k)$ and the cross data evidence is

$$c_\alpha(k) = \sum_{\beta \in \mathcal{D}} \frac{w_{\alpha,\beta}}{n_\beta} \sum_{j=1}^{n_\beta} \left(\left(\frac{L_\alpha \alpha_k}{\|L_\alpha \alpha_k\|}\right)^\top \left(\frac{R_\beta \beta_j}{\|R_\beta \beta_j\|}\right) \right).$$

The set $\mathcal{D} = \{A, Q, V_1, \ldots, V_F\}$ consists of the possible data types. The trainable parameters of the model are: (1) $V_\alpha, L_\alpha, R_\alpha$ which re-embed the data representation to tune the attention; (2) v_α which scores the local modality; and (3) $\hat{w}_\alpha, w_\alpha, w_{\alpha,\beta}$ which weight the three components with respect to each other.

We found the use of attention for history to not yield improvements. Therefore, we obtain the attended textual representation $a_T \in \mathbb{R}^{d_T}$ by concatenating the attended question representation $a_Q \in \mathbb{R}^{d_Q}$ with the history representation $r_H \in \mathbb{R}^{d_H}$. Consequently, $d_T = d_Q + d_H$.

3.3. Data Representation

The proposed approach relies on representations r_α obtained for a variety of data components which we briefly discuss subsequently.

Video: Containing both temporal and spatial information, video data is among the most memory consuming. Common practice is to reduce the spatial information while maintaining attention over the temporal dimension. Instead, we first reduce the temporal dimension, maintaining the ability for spatial attention to reason about the video content. To ensure fast training, we reduce the temporal dimension by sampling F frames uniformly. For each sampled frame we extract a representation from a deep net trained on ImageNet (in our case VGG19). We then fine tune the representation of each frame using a 1D conv layer with a bias term. This conv layer is identical for all the F frames. Consequently, we obtain the video representation $r_V \in \mathbb{R}^{F \times n_V \times d_V}$, where F is the number of sampled frames, n_V is the spatial dimension and d_V is the embedding dimension.

Audio: For audio, we extracted features from a strong audio classification model (*i.e.*, VGGish [24]) by taking the last representation before the final FC-layer. This representation has adaptive temporal length. For each batch we find the maximal temporal length of the audio signal, and zero-padded the shorter audio representations. We then fine-tune each audio file using a 1D conv layer with a bias. We obtain the audio representation $r_A \in \mathbb{R}^{n_A \times d_A}$, where n_A is the maximal temporal length of a given batch and d_A is the embedding dimension.

Question: We start with an adaptive-length list of 1-hot word-representations. For each batch we find the longest sentence, and zero-pad shorter ones. We embed each word using a linear-embedding layer, followed by a single layer LSTM-net with dropout. The last hidden state of the LSTM is the question representation $r_Q \in \mathbb{R}^{n_Q \times d_Q}$, where n_Q is the length of the maximal sentence for the given batch and d_Q is the embedding dimension.

History: The history data source consists of the past T question-answer pairs, which we denote by $H = (Q, A)_{t \in \{1,\dots,T\}}$. The history embedding consists of two components: we first embed each question-answer pair $(Q, A)_t$ using a LSTM-net to get T representations of the history. We then feed these representations into another LSTM-net to obtain the vector representation $r_H \in \mathbb{R}^{d_H}$, where d_H is the history embedding dimension.

We embed each question-answer pair $(Q, A)_t$ following the question embedding above. A question-answer pair starts with a list of 1-hot word-representations of the words in the question followed by 1-hot word-representations of the words in the answer. For each batch we find the longest question-answer sequence, and zero-pad the shorter ones. We embed each 1-hot vector using a linear-embedding layer, followed by a two layer LSTM-net with a dropout. The last hidden state of this LSTM-net is the vector representation of $(Q, A)_t$, which we denote by r_t.

We embed the history by feeding $r_1, \dots, r_T$ to a one layer LSTM-net with dropout, in order to capture the temporal aspect of the question-answer history. To deal with the
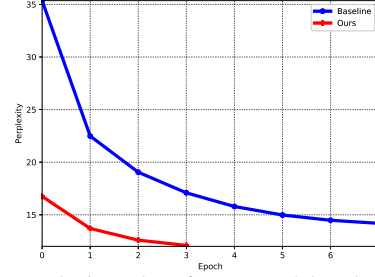

Figure 5: Perplexity values for our model *vs.* baseline [25].

adaptive length of history interactions, for each batch we find the interaction with the longest history, and zero-pad question-answer pairs with shorter history. The final LSTM-net hidden state is the history representation $r_H \in \mathbb{R}^{d_H}$, where d_H is the history embedding dimension.

4. Results

In the following we evaluate the discussed baseline on the Audio Visual Scene-Aware Dialog (AVSD) dataset. We follow the proposed protocol and assess the generated answers to a user question given a dialog context [2, 25]. This context consists of a dialog history (previous questions and answers) in addition to video and audio information about the scene. Our code is publicly available[1].

4.1. AVSD v0.1 Dataset

The AVSD dataset consists of annotated conversations about short videos. The dataset contains 9,848 videos taken from CHARADES, a multi-action dataset with 157 action categories [61]. Each dialog is obtained from two Amazon Mechanical Turk (AMT) workers, who discuss about events in a video. One of the workers takes the role of an answerer who had already watched the video. The answerer replies to questions asked by another AMT worker, the questioner.

The questioner was not shown the whole video but only the first, middle and last frames of the video. The dialog revolves around the events in and other aspects of the video. The AVSD v0.1 dataset is split into 7,659 train dialogs, 1,787 validation and 1,710 test dialogs. Because the test set doesn't currently include ground truth, we follow [25] and evaluate on the 'prototype test-set' with 733 dialogs. Because the 'prototype test-set' is part of the 'v0.1 validation-set,' we use the 'prototype validation-set' with 732 dialogs, which doesn't overlap with the 'prototype test-set.'

4.2. Implementation Details

Our system relies on textual, visual and audio data representations, *i.e.*, r_α for $\alpha \in \{A, Q, V_1, \dots, V_F\}$. For the **video** representation we randomly sample $F = 4$ equally spaced frames, and use the last conv layer of a VGG19 having a dimensions of $7 \times 7 \times 512$. Therefore the visual embedding dimension is $d_V = 512$. After flattening the 2D

---

[1]https://github.com/idansc/simple-avsd

Table 1: Results for the AVSD dataset for CIDEr, BLEU1, ...,
BLEU4, ROUGE-L, METEOR. We provide a comparison to the
baseline and a detailed ablation study separated into categories and
discussed in Sec. 4.5. We also report the number of parameters for
each baseline.

Model	C	B4	B3	B2	B1	R	M	P
baseline[25][2]	0.766	0.084	0.117	0.173	0.273	0.291	0.117	6.15M
basic baselines								
q	0.815	0.088	0.122	0.178	0.279	0.297	0.121	3.1M
q+h	0.843	0.089	0.123	0.178	0.277	0.296	0.122	4.51M
q+h+vgg-spatial	0.869	0.089	0.124	0.180	0.279	0.302	0.123	5.12M
q+h+vgg-spatial+audio	0.874	0.091	0.125	0.182	0.282	0.305	0.124	5.23M
basic baselines+attention								
q+att	0.849	0.090	0.124	0.179	0.278	0.298	0.121	3.35M
q+h+att	0.861	0.090	0.124	0.177	0.271	0.298	0.122	4.57M
q+h+vgg-spatial+att	0.908	0.093	0.129	0.185	0.283	0.307	0.125	7.4M
attention-model								
w/o-cross-data-evidence	0.896	0.095	0.131	0.190	0.292	0.309	**0.128**	7.5M
w/o-local-evidence	0.917	0.096	0.132	0.191	0.293	0.309	**0.128**	8.35M
w/o-question-prior	0.906	0.096	0.132	0.190	0.292	0.309	0.127	8.35M
sharing–weights	0.923	**0.097**	**0.133**	**0.191**	**0.293**	0.309	0.127	6.18M
video-fusion								
temporal-attention	0.877	0.091	0.126	0.182	0.281	0.302	0.124	8.4M
summation	0.890	0.093	0.128	0.183	0.283	0.303	0.124	7.35M
weighted-summation	0.876	0.094	0.130	0.187	0.289	0.304	0.126	7.85M
video-audio-lstm	0.865	0.076	0.101	0.141	0.210	0.286	0.108	8.35M
decoder-input								
q-first-state	0.704	0.078	0.110	0.163	0.257	0.279	0.112	8.35M
all-first-state	0.714	0.079	0.114	0.171	0.271	0.276	0.113	10.1M
all-concat-decoder-input	0.797	0.089	0.125	0.183	0.285	0.297	0.121	9.53M
q+h+a-concat-input	0.857	0.090	0.123	0.177	0.274	0.298	0.121	7.72M
i3d-features-&-spatial-temporal								
i3d-rgb-temporal	0.886	0.094	0.130	0.188	0.289	0.306	0.126	7.23M
i3d-rgb-flow-temporal	0.851	0.091	0.127	0.185	0.286	0.303	0.125	7.82M
i3d-rgb-spatial-10	0.928	**0.097**	**0.133**	0.190	0.290	0.310	0.127	6.58M
vgg-spatial-1	0.919	0.095	0.130	0.187	0.287	0.309	0.126	6.18M
vgg-spatial-16	0.903	0.093	0.128	0.186	0.287	0.307	0.127	28.88M
initialization								
default	0.877	0.090	0.123	0.178	0.274	0.300	0.121	8.35M
xavier	0.848	0.087	0.119	0.171	0.262	0.297	0.119	8.35M
he	0.913	0.095	0.131	0.189	0.290	0.308	0.127	8.35M
beam-search hyper-parameters								
w/o beam	0.924	0.082	0.109	0.152	0.226	0.298	0.114	8.35M
2-width	0.934	0.094	0.128	0.183	0.279	0.311	0.126	8.35M
4-width	0.931	0.096	0.131	0.188	0.287	0.310	0.127	8.35M
5-width	0.926	0.096	0.132	0.188	0.289	0.309	0.127	8.35M
Ours	**0.941**	0.096	0.131	0.187	0.285	**0.311**	**0.128**	8.35M

spatial dimension, we obtain the spatial dimension $n_V = 49$.
For **audio** features we use VGGish that operates on 0.96s
log-Mel spectrogram patches extracted from 16kHz audio,
and outputs a $d_A = 128$ dimensional vector. VGGish inputs
overlap by 50%, therefore an output is provided every 0.48s.
Dropout parameters before the last FC layer, and the LSTM
layers are set to 0.5. For the **question** representation we set
the word embedding dimension to 128. The questions are
embedded to $d_Q = 256$ dimensional vectors, extracted from
the last hidden state of their LSTM-net. The **history** con-
sists of $T = 10$ question-answer pairs, which we denote by
$H = (Q, A)_{t \in \{1,...,T\}}$. We use an LSTM-net with a hidden
state of $d_H = 128$ to encode the history.

4.3. Training

We use a cross-entropy loss on the probabilities,
$p(y_i|y_{<i}, x)$ to train the answer generator, the attention and
the embedding layers jointly end-to-end. The total amount
of trainable parameters are 8,359,107. We use the Adam
optimizer [34] with a learning rate of 0.001 and a batch size
of 64. During training after each epoch we evaluate our
performance on the validation set using a perplexity metric.
We stop our training after two consecutive epochs with no
improvement.

We use a standard machine with an Nvidia Tesla M40
GPU for all our experiments. Training our system takes 4
epochs to converge *vs*. 9 epochs for the baseline (see Fig. 5).
Each epoch takes 8 minutes *vs*. 13 minutes for the baseline.
In total, training our model takes approximately 30 minutes.

4.4. Performance Evaluation:

We evaluate the performance of our system using several
metrics. Our prime metric is CIDEr, the Consensus-based
Image Description Evaluation, which measures the similar-
ity of a sentence to the consensus [66]. We also evaluate
our performance on the ROUGE-L metric (Recall Oriented
Understudy of Gisting Evaluation). This is a recall-based
metric that measures the longest common subsequence of
tokens [38]. The METEOR metric is a unigram precision
and recall that allows for matchings between candidates and
references [7]. We also evaluate our performance using the
traditional BLEU score, which measures the effective over-
lap between a reference sentence and a candidate sentence.
We measure the geometric mean of the effective n-gram
precision scores, for $n = 1, \ldots, 4$ and refer to these as
BLEU1,..., BLEU4.

4.5. Quantitative Results and Insights for a Good Baseline

We compare to the baseline discussed in [25]. In the fol-
lowing we explore the various components of audio-visual
dialog systems and present our insights for constructing a
simple and effective baseline. These insights cover all as-
pects of our system: feature embedding, attention, fusion
and training techniques. We particularly emphasize the im-
portance of spatial features for AVSD, which we contrast
with the action recognition based I3D features.

Question Bias and Basic Baselines: We revisit the scores
published by [25] and assess a basic seq2seq-type baseline,
with no attention [64]. In this variant, which we call **q** in
Tab. 1, we encode the question using a word embedding
(with embedding dimension of 128) and a 1-layer LSTM-net
(with hidden state dimension of 256 compared to a dimen-
sion of 128 in the baseline), without any video or history
related features. For decoding, another 1-layer LSTM-net
(with hidden state dimension of 256 compared to a dimen-
sion of 128 in the baseline) is used. Surprisingly, this model
alone was able to surpass the current baseline of [25]. Sim-
ilar results are also reported in [55]. This indicates that
there might be bias-problem within the AVSD dataset, no
visual information is needed. For instance a common ques-
tion is "How many people are in the video?", but videos
in many cases feature only one person. Another example
are questions of the form "is it indoor?" which are mean-
ingless since the CHARADES dataset focuses on indoor
activities. Another possible explanation for this good result
is the encoding of the answer in the question. For instance,
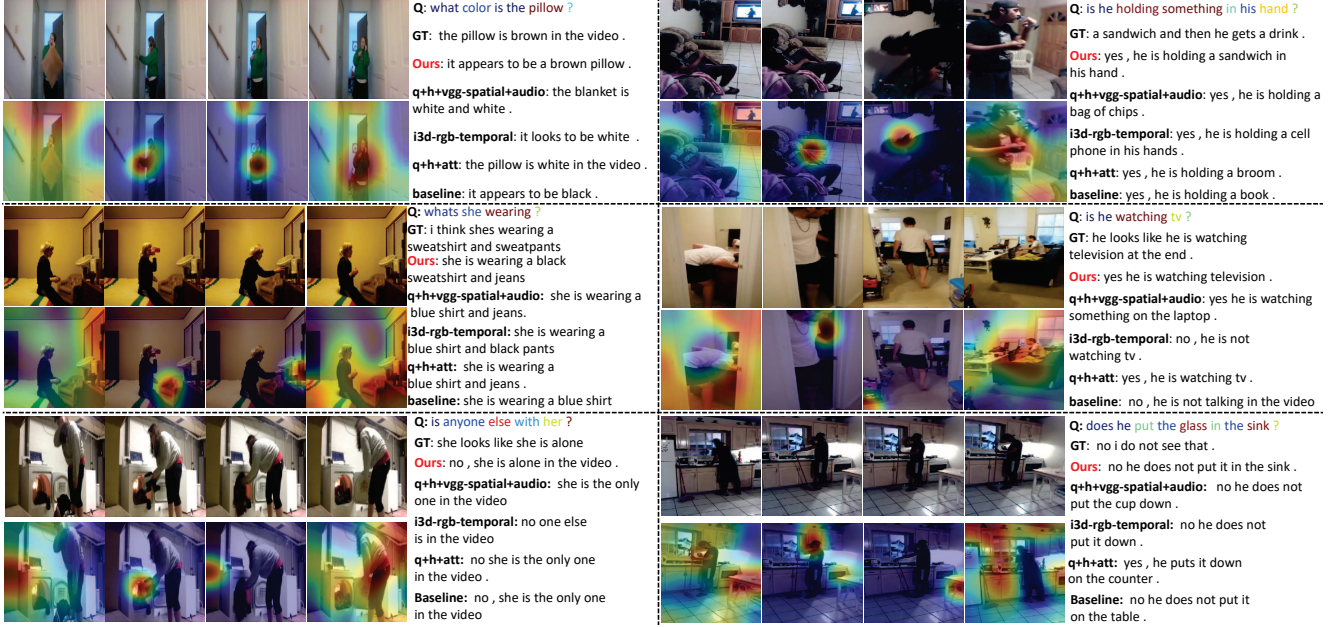a question "this person is standing in a kitchen correct?" is

Figure 6: An illustration of out 4-framed samples from a video along with the relevant attention variables. Our attention treats any frame as different component. This allows the attention module to learn different attention behaviors for different temporal locations. We observe the first and fourth samples are noisier, while the second and third attend to specific interesting locations. Our multimodal attention also generates attention for questions, illustrated over the question via a word heat map. We provide generated answers for different baseline models: q+h+att, is a model with only history and question input; i3d-rgb-temporal is a model with temporal features instead of spatial; q+h+vgg-spatial+audio is a model without attention. We also compare to the generated answer by [25]. the ground-truth is denoted by GT, and our final model denoted by Ours.

answered with "yes he is in the kitchen." Moreover, generative evaluation is also more prone to biases, as the evaluation emphasizes correct sentence structure rather than correctness of the answer. Very recently, a discriminative approach was proposed [1]. The bias problem is not unique to AVSD, and was also discussed for Visual Question Answering [22].

To further improve the most basic baseline **q**, we add more modalities. We use the fusion and embedding techniques of the proposed model but omit attention. Instead of attention, we use a mean over the representation for visual and auditory data sources, and the last hidden state of the LSTM-net is used to represent the question data source. We found that our model can utilize any modality supplement, even without attention. In the 'basic baselines+attention' section of Tab. 1 we assess versions with attention, which brings us closer to our full model.

Spatial vs. Temporal Information: Current methods focus on temporal models and often naïvely reduce the spatial dimension [25, 70, 80]. In contrast, for closely related visual reasoning tasks, such as visual dialog and visual question answering, it is broadly accepted that spatial attention is necessary. Therefore, it is unlikely that video reasoning is effective when simply reducing the spatial dimension. Indeed, we find better results when reducing the temporal dimension with sampling techniques and employing attention to reduce the spatial dimension. In Fig. 6 we observe that a small subset of frames (*e.g.*, 4) is usually enough for an

almost complete understanding of the video. In the 'i3d-features-&-spatial-temporal' section of Tab. 1, we compare spatial-based features to temporal-based ones. The temporal features are computed on a stack of 16 video frames, and are treated as an input modality to our attention mechanism. Attention choses the relevant temporal locations. The temporal attended representation was fed to the Aud-Vis LSTM-net along with the audio attended-features. For the i3d-rgb-flow version we also use the I3D model based on optical flow features as an additional data component. This resulted in a drop in performance compared to the spatial-based i3d-features reported in the i3d-rgb-spatial-10 line of Tab. 1. We also test different number of sampled frames. Interestingly, only one frame is already very useful for AVSD, and too many VGG-frames harm performance. Note that each frame is coupled to an attention-score and treated as a modality, which explains why too many frames can add noise to the inferred multimodal probability.

I3D Features *vs*. VGG: I3D features are widely used as video-based feature extractor (*cf*. [9]), discarding the classical image-based features, *e.g*., VGG. They are extracted from a model trained on the Kinetics Dataset, a dataset for action recognition, and have been shown to improve many video tasks. We find that while I3D features have repeatedly been shown to improve on action-recognition tasks, they are not as useful in the answer generation task of AVSD. Equipped with VGG features we were able to achieve comparable re-

sults to the i3d-rgb-spatial-20 version. The i3d-rgb-spatial features are 4 times bigger (7x7x512 *vs.* 2x7x7x1024), as well as more complicated to extract. Seeking simplicity, we report scores with the VGG-based features subsequently. This may also indicate a weakness in the dataset, as this solution seems to be sub-optimal for action-related questions (*e.g.*, classifying sequences of actions). Not only do we naïvely sample temporal frames, but also do we not use I3D features that were extracted from a network trained for action-recognition, yet we achieve good results.

Attention Model: We assess different components of the attention model. See Sec. 3.2 for details about local evidence and cross data evidence. We found that every component contributes to the model, especially the cross-data component. The cross-data component determines the attention score of an element by considering interactions with other modalities. For instance, a region in the second frame can affect a region in the third frame, or perhaps a word in the question.

To find the simplest attention module, we also explored the option of grouping together the parameters for all video frames, *i.e.*, $V_{V_1} = \ldots = V_{V_F}$, $L_{V_1} = \ldots = L_{V_F}$, and $R_{V_1} = \ldots = R_{V_F}$, which yields good results despite 2 million fewer parameters. This version allows to increase the number of processed frames, with no additional memory cost. Those results are reported in the 'sharing-weights' line of Tab. 1.

Multimodal Decoding Fusion: We experimented with several variants that reduce $a_A, a_{V_1}, \ldots, a_{V_F}$. In Tab. 1, section 'decoder-input,' we show a version that uses an additional multimodal attention step over the video-related attended vector, called temporal-attention. Another attempt is summation polling of the vectors, and weighted summation with scalers. Instead, we note the sequential information of $a_{V_1}, \ldots, a_{V_F}$ that naturally calls for the use of an additional LSTM unit, which we call Aud-Vis (see Fig. 4). We think audio is a more general cue while frames have more specific information. Ordering is guided by the intuition that LSTM-based encoding commonly starts with more general information. To verify this intuition, in video-audio-lstm, we performed additional experiments with ordering of $a_{V_1}, \ldots, a_{V_F}, a_A$.

Next we find a good way to input elements into the answer generation LSTM-net. We first analyze the basic **q** model. A classic decoder, where encoded **q** are fed as first hidden state to the LSTM-net is reported in the 'q-first-state' row in Tab. 1 (decoder-input section). This suggest that textual data should be concatenated to the decoder inputs. Concatenating all modalities to the input, which is reported in the 'all-concat-input' line in Tab. 1 drops the performance, suggesting that a dichotomy of video-related and textual-related features is useful. To incorporate the audio signal, we find it's best to use it as a first state in the Aud-Vis LSTM-net. A version where we concatenated the audio attended vector to a_T is referred to as 'q+h+a-concat-input+s-first-state.' The model behaves the best when the fused video related features were used as the initial state h_0 of the Ans-Generation LSTM-net. Our state-of-the-art model further improves the fusion technique by using the Aud-Vis LSTM-net to generate h_0 which captures the temporal information of audio attention a_A and the visual attention $a_{V_1}, \ldots, a_{V_F}$.

Weight Initialization: An important aspect is the initialization of the deep net parameters. We observed a significant improvement using Kaiming normal initialization or Xavier initialization for all LSTM models [23, 21].

Beam Search Width: In an attempt to improve the overall evaluation time, we experimented with different beam width. We found that although beam search is useful for generation, a width of 2 achieves almost as good results. Our version use 3-width beam search.

4.6. Qualitative Results

In Fig. 6, we show several examples of generated answers of five models, our final model, a version without any attention (q+h+vgg-spatial+audio), a version with temporal I3D features (i3d-rgb-temporal), a version with only textual modalities (q+h+att), and the baseline [25]. The ground-truth is referred to via GT. Additionally, we take advantage of the interpretability of attention modules to also illustrate the attention probabilities of our final modal on 5 different modalities, *i.e.*, our 4-frames, and the question. First, we observe an interesting behavior of our attention model: each sampled frame is attended a differently, which captures different features from different frames. The first and fourth frames are noisier and extract general concepts, while the second and third capture unique aspects of the video, *e.g.*, a person, a couch. This behavior can be associated with the temporal aspect of the frames. Meaning it is more important to capture general aspects at the end and at the beginning, but in the middle we reveal the important specific concepts. Additionally, the question attention attends to the informative words. Our generated answers are usually more aware of the scene, and less prone to bias. For instance, in the first row, the question is "what color is the pillow?." We observe our model to be able to answer the correct color, while all other model variants answer with white, the most-common color of a pillow. In another question "whats she is wearing," our model was the only one to relate to her black sweatshirt.

5. Conclusion

We propose a simple baseline for Audio-Visual Scene-Aware Dialog that surpasses current techniques by 20% on the CIDEr metric. Pioneering on this task, we carefully evaluated our approach. We hope our analysis can bridge the gap between video-reasoning and image-reasoning.

References

[1] H. Alamri, V. Cartillier, A. Das, J. Wang, S. Lee, P. Anderson, I. Essa, D. Parikh, D. Batra, A. Cherian, T. K. Marks, and C. Hori. Audio-visual scene-aware dialog. *arXiv preprint arXiv:1901.09107*, 2019. 7

[2] H. Alamri, V. Cartillier, R. G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, and C. Hori. Audio Visual Scene-Aware Dialog (AVSD) Challenge at DSTC7. In *https://arxiv.org/abs/1806.00525*, 2018. 1, 5

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. In *Proc. CVPR*, 2016. 2

[4] J. Aneja, A. Deshpande, and A. G. Schwing. Convolutional Image Captioning. In *Proc. CVPR*, 2018. 1

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *Proc. ICCV*, 2015. 1, 2

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

[7] S. Banerjee and A. 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. 6

[8] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. ICCV*, 2017. 2

[9] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*, 2017. 1, 2, 7

[10] M. Chatterjee and A. G. Schwing. Diverse and Coherent Paragraph Generation from Images. In *Proc. ECCV*, 2018. 1

[11] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proc. CVPR*, 2015. 1

[12] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proc. EMNLP*, 2016. 2

[13] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. Visual Dialog. In *Proc. CVPR*, 2017. 1, 2

[14] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proc. ICCV*, 2017. 1, 2

[15] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. A. Forsyth. Diverse and Controllable Image Captioning with Part-of-Speech Guidance. In *Proc. CVPR*, 2019. 1, 2

[16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2015. 1

[17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. In *Proc. SIGGRAPH*, 2018. 3

[18] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proc. EMNLP*, 2016. 2, 3

[19] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and Methods for Multilingual Image Question Answering. In *Proc. NIPS*, 2015. 1, 2

[20] R. Gao, R. S. Feris, and K. Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video. In *Proc. CVPR Workshop*, 2018. 3

[21] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. AISTATS*, 2010. 8

[22] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. CVPR*, 2017. 7

[23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. ICCV*, 2015. 8

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

[25] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features. In *https://arxiv.org/abs/1806.08409*, 2018. 1, 3, 5, 6, 7, 8

[26] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting Visual Question Answering Baselines. In *Proc. ECCV*, 2016. 2

[27] U. Jain, S. Lazebnik, and A. G. Schwing. Two can play this Game: Visual Dialog with Discriminative Question Generation and Answering. In *Proc. CVPR*, 2018. 1, 2

[28] U. Jain*, Z. Zhang*, and A. G. Schwing. Creativity: Generating Diverse Questions using Variational Autoencoders. In *Proc. CVPR*, 2017. * equal contribution. 1, 2

[29] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. CVPR*, 2017. 2

[30] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, 2016. 1

[31] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 2015. 1, 2

[32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *Proc. CVPR*, 2014. 2

[33] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 3

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint 2014*. 6

[35] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proc. ECCV*, 2018. 2

[36] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, and X. Wang. Visual Question Generation as Dual Task of Visual Question Answering. In *https://arxiv.org/abs/1709.07192*, 2017. 1, 2

[37] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek. Videolstm convolves, attends and flows for action recognition. *CVIU*, 2018. 3

[38] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 6

[39] 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 *Proc. ECCV*, 2014. 1

[40] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Proc. NIPS*, 2017. 3

[41] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. NIPS*, 2016. 2, 3

[42] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Proc. AAAI*, 2016. 2

[43] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Proc. NIPS*, 2014. 2

[44] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proc. ICCV*, 2015. 1, 2

[45] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-rnn). In *Proc. ICLR*, 2015. 1, 2

[46] D. Massiceti, N. Siddharth, P. K. Dokania, and P. H. S. Torr. FlipDial: A Generative Model for Two-Way Visual Dialogue. In *https://arxiv.org/abs/1802.03803*, 2018. 1, 2

[47] N. Mostafazadeh, D. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017. 1

[48] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *Proc. ACL*, 2016. 1, 2

[49] M. Narasimhan, S. Lazebnik, and A. G. Schwing. Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *Proc. NIPS*, 2018. 1

[50] M. Narasimhan and A. G. Schwing. Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering. In *Proc. ECCV*, 2018. 1

[51] A. Owens and A. A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *Proc. ECCV*, 2018. 3

[52] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Learning Sight from Sound: Ambient Sound Provides Supervision for Visual Learning. *IJCV*, 2018. 3

[53] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. CVPR*, 2016. 2

[54] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Proc. NIPS*, 2015. 1, 2

[55] R. Sanabria, S. Palaskar, and F. Metze. Cmu sinbadas submission for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*, 2019. 6

[56] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Proc. NIPS*, 2017. 3

[57] I. Schwartz, A. G. Schwing, and T. Hazan. High-Order Attention Models for Visual Question Answering. In *Proc. NIPS*, 2017. 1, 2, 3, 4

[58] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing. Factor Graph Attention. In *Proc. CVPR*, 2019. 1, 2, 3, 4

[59] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proc. CVPR*, 2016. 1, 2

[60] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to Body Dynamics. In *Proc. CVPR*, 2018. 3

[61] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. ECCV*, 2016. 5

[62] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proc. NIPS*, 2014. 2

[63] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 2

[64] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proc. NIPS*, 2014. 6

[65] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*, 2015. 2

[66] R. Vedantam, L.-C. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, 2015. 6

[67] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. In *Proc. AAAI*, 2018. 2

[68] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, 2015. 1

[69] L. Wang, A. G. Schwing, and S. Lazebnik. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In *Proc. NIPS*, 2017. 1, 2

[70] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*. Springer, 2016. 2, 7

[71] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. *arXiv preprint arXiv:1711.07613*, 2017. 3

[72] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proc. CVPR*, 2018. 2

[73] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proc. ICML*, 2016. 1, 2

[74] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proc. ECCV*, 2016. 1, 2, 3

[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. ICML*, 2015. 1, 2, 3

[76] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proc. CVPR*, 2016. 1, 2, 3

[77] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proc. CVPR*, 2016. 1

[78] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proc. CVPR*, 2016. 2

[79] D. Zhang and M. Shah. Human Pose Estimation in Videos. In *Proc. ICCV*, 2015. 2

[80] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017. 2, 7

[81] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proc. CVPR*, 2016. 2