COMP 786 (Fall 2020)
Natural Language Processing

Week 11: Language+Vision (incl. several Guest Research Talks)

Mohit Bansal
Language+Vision
Example Major Language+Vision Tasks

- Image Captioning
- Referring Expressions
- Image/Visual Question Answering
- Visual Dialog
- Video Captioning
- Video QA/Dialogue
- Cross-Modal Pretraining Models & Text-to-Image Generation
Brief Task Definitions and Example Papers/Models
Image Captioning

CNN

LSTM

LSTM

LSTM

LSTM

LSTM

<BOS> a dog is jumping <EOS>
Example Early Methods

[Ordonez et al., 2011]
Example Early Methods

Input Image

1) Object(s)/Stuff
   a) dog
   b) person
   c) sofa

2) Attributes
   - dog: brown 0.01, striped 0.16, furry 0.26, wooden 0.2, feathered 0.06
   - person: brown 0.32, striped 0.02, furry 0.04, wooden 0.2, feathered 0.04
   - sofa: brown 0.04, striped 0.03, furry 0.10, wooden 0.8, feathered 0.08

3) Prepositions
   - near(a, b) 1, near(b, a) 1, against(a, b) 0.11, against(b, a) 0.04, beside(a, b) 0.24, beside(b, a) 0.17
   - near(c, a) 1, near(c, b) 1, against(a, c) 0.3, against(b, c) 0.05, beside(a, c) 0.45, beside(b, c) 0.45

4) Constructed CRF

5) Predicted Labeling
   - <<null, person_b>, against, <brown, sofa_1>>
   - <<null, dog_a>, near, <null, person_b>>
   - <<null, dog_a>, beside, <brown, sofa_1>>

6) Generated Sentences
   This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

[Kulkarni et al., 2013]
Show, Attend, and Tell

Attention: 

$$e_{ti} = f_{\text{att}}(a_i, h_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$
Show, Attend, and Tell

Figure 4. Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word).

5.2. Evaluation Procedures

A few challenges exist for comparison, which we explain here. The first challenge is a difference in choice of convolutional feature extractor. For identical decoder architectures, using a more recent architectures such as GoogLeNet (Szegedy et al., 2014) or Oxford VGG (Simonyan & Zisserman, 2014) can give a boost in performance over using the AlexNet (Krizhevsky et al., 2012).

In our evaluation, we compare directly only with results which use the comparable GoogLeNet/Oxford VGG features, but for METEOR comparison we include some results that use AlexNet.

The second challenge is a single model versus ensemble comparison. While other methods have reported performance boosts by using ensembling, in our results we report a single model performance.

Finally, there is a challenge due to differences between dataset splits. In our reported results, we use the predefined splits of Flickr8k. However, for the Flickr30k and COCO datasets is the lack of standardized splits for which results are reported. As a result, we report the results with the publicly available splits used in previous authors of Vinyals et al. (2014), Karpathy & Li (2014) and Kiros et al. (2014b). For fairness, we only compare against results for which we have verified that our BLEU evaluation code is the same.

4 BLEU-n is the geometric average of the n-gram precision. For instance, BLEU-1 is the unigram precision, and BLEU-2 is the geometric average of the unigram and bigram precision.

5.3. Quantitative Analysis

In Table 1, we provide a summary of the experiment validating the quantitative effectiveness of attention. We obtain state of the art performance on the Flickr8k, Flickr30k and MS COCO. In addition, we note that in our experiments we are able to significantly improve the state-of-the-art performance METEOR on MS COCO. We speculate that this is connected to some of the regularization techniques we used (see Sec. 4.2.1) and our lower-level representation.

5.4. Qualitative Analysis: Learning to attend

By visualizing the attention learned by the model, we are able to add an extra layer of interpretability to the output of the model (see Fig. 1). Other systems that have done this rely on object detection systems to produce candidate alignment targets (Karpathy & Li, 2014). Our approach is much more flexible, since the model can attend to "non-object" salient regions.

The 19-layer OxfordNet uses stacks of 3x3 filters meaning the only time the feature maps decrease in size are due to the max pooling layers. The input image is resized so that the shortest side is 256-dimensional with preserved aspect ratio. The input to the convolutional network is the center-cropped 224x224 image. Consequently, with four max pooling layers, we get an output dimension of the top convolutional layer of 14x14. Thus in order to visualize the attention weights for the soft model, we upsample the weights by a factor of $2^4 = 16$ and apply a Gaussian filter.

[Xu et al., 2015]
Visual Referring Expressions

RefCOCO TestB

RefCOCO+ TestA

bottom left banana
second banana from left
top right banana
red shirt
man in black
blue shirt
Joint Comprehension+Generation Model

Speaker

- CNN-LSTM structure.
- Referring expression generation.
- Features are composed of two parts: appearance similarity and location and size similarity.
- Improving performance on both the comprehension and generation tasks.
- Reinforcer module helps improve the speaker by sampling more discriminative (less ambiguous) expressions.

Listener

- Joint-embedding model learned to minimize the distance between paired object and expression representations.
- Encoder encodes the visual information from the target object and semantic information from the referring expression into a joint embedding space.
- Listener embeds the joint feature into the LSTM for referring expression comprehension task, given a referring expression representation, the listener embeds it into the joint space, then selects the closest object in the embedding expression representation.
- Listener receives a ranking loss on objects.

Reinforcer

- Sampling
- Reward Loss
- Generation loss

Model

- Man in the middle wearing yellow
- Sampling
- Reward Loss
- Generation loss
- Embedding Loss
- Normalization
- MLP
- LSTM
- Concat
- LSTM
- FC

References

[Yu et al., 2017]
Joint Comprehension+Generation Model

Figure 1: Joint generation examples using our full model with “+rerank” on three datasets. Each sentence shows the generated expression for one of the depicted objects (color coded to indicate correspondence).

Figure 2: Example comprehension results using our full model on three datasets. Green box shows the ground-truth region and blue box shows our correct comprehension based on the detected regions.
VQA: Visual Question Answering

What color are her eyes? What is the mustache made of?

How many slices of pizza are there? Is this a vegetarian pizza?

Is this person expecting company? What is just under the tree?

Does it appear to be rainy? Does this person have 20/20 vision?
Demo

- [http://vqa.cloudcv.org/](http://vqa.cloudcv.org/)

Predicted top-5 answers with confidence:

<table>
<thead>
<tr>
<th>Answer</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>67.267%</td>
</tr>
<tr>
<td>1</td>
<td>22.324%</td>
</tr>
<tr>
<td>3</td>
<td>9.115%</td>
</tr>
<tr>
<td>4</td>
<td>0.945%</td>
</tr>
<tr>
<td>0</td>
<td>0.242%</td>
</tr>
</tbody>
</table>

[Agrawal et al., 2015]
Simple VQA Baseline

“How many horses are in this image?”

[Agarwal et al., 2015]
Figure 1: Flowchart of our proposed hierarchical co-attention model. Given a question, we extract its word level, phrase level and question level embeddings. At each level, we apply co-attention on both the image and question. The final answer prediction is based on all the co-attended image and question features.
Hierarchical Co-Attention Model

Figure 2: (a) Parallel co-attention mechanism; (b) Alternating co-attention mechanism.

[Lu et al., 2016]
Hierarchical Co-Attention Model

Figure 4: Visualization of image and question co-attention maps on the COCO-QA dataset. From left to right: original image and question pairs, word level co-attention maps, phrase level co-attention maps and question level co-attention maps. For visualization, both image and question attentions are scaled (from red:high to blue:low). Best viewed in color.

[Lu et al., 2016]
Multimodal Compact Bilinear (MCB) Model

Modeling textual or visual information with vector representations trained from large language or visual datasets has been successfully explored in recent years. However, tasks such as visual question answering require combining these vector representations with each other. Approaches to multimodal pooling include element-wise product or sum, as well as concatenation of the visual and textual representations. We hypothesize that these methods are not as expressive as an outer product of the visual and textual vectors. As the outer product is typically infeasible due to its high dimensionality, we instead propose utilizing Multimodal Compact Bilinear pooling (MCB) to efficiently and expressively combine multimodal features. We extensively evaluate MCB on the visual question answering and grounding tasks. We consistently show the benefit of MCB over ablations without MCB. For visual question answering, we present an architecture which uses MCB twice, once for predicting attention over spatial features and again to combine the attended representation with the question representation. This model outperforms the state-of-the-art on the Visual7W dataset and the VQA challenge.

1 Introduction

Representation learning for text and images has been extensively studied in recent years. Recurrent neural networks (RNNs) are often used to represent sentences or phrases (Sutskever et al., 2014; Kiros et al., 2015), and convolutional neural networks (CNNs) have shown to work best to represent images (Donahue et al., 2013; He et al., 2015). For tasks such as visual question answering (VQA) and visual grounding, most approaches require joining the representation of both modalities. For combining the two vector representations (multimodal pooling), current approaches in VQA or grounding rely on concatenating vectors or applying element-wise sum or product. While this generates a joint representation, it might not be expressive enough to fully capture the complex associations between the two different modalities.

In this paper, we propose to rely on Multimodal Compact Bilinear pooling (MCB) to get a joint representation. Bilinear pooling computes the outer product between two vectors, which allows, in contrast to element-wise product, a multiplicative interaction between all elements of both vectors. Bilinear pooling models (Tenenbaum and Freeman, 2000) have recently been shown to be beneficial for fine-grained classification for vision only tasks (Lin et al., 2015). However, given their high dimensionality ($n^2$), bilinear pooling has so far not been widely used.

In [Fukui et al., 2016],
Figure 3: Our architecture for VQA: Multimodal Compact Bilinear (MCB) with Attention. Conv implies convolutional layers and FC implies fully connected layers. For details see Sec. 3.2.

where \( \cdot \) is the convolution operator. Additionally, the convolution theorem states that convolution in the time domain is equivalent to element-wise product in the frequency domain. The convolution \( x \ast q \) can be rewritten as

\[
\text{FFT}^{-1}(\text{FFT}(x) \cdot \text{FFT}(q))
\]

where \( \cdot \) refers to element-wise product. These ideas are summarized in Figure 2 and formalized in Algorithm 1, which is based on the Tensor Sketch algorithm of Pham and Pagh (2013). We invoke the algorithm with \( v_1 = x \) and \( v_2 = q \). We note that this easily extends and remains efficient for more than two multi-modal inputs as the combination happens as element-wise product.

3.2 Architectures for VQA

In VQA, the input to the model is an image and a question, and the goal is to answer the question. Our model extracts representations for the image and the question, pools the vectors using MCB, and arrives at the answer by treating the problem as a multi-class classification problem with 3,000 possible classes.

We extract image features using a 152-layer Residual Network (He et al., 2015) that is pretrained on ImageNet data (Deng et al., 2009). Images are resized to 448 x 448, and we use the output of the layer ("pool5") before the 1000-way classifier. We then perform L2 normalization on the 2048-D vector.

Input questions are first tokenized into words, and the words are one-hot encoded and passed through a learned embedding layer. The tanh nonlinearity is used after the embedding. The embedding layer is followed by a 2-layer LSTM with 1024 units in each layer. The outputs of each LSTM layer are concatenated to form a 2048-D vector.

The two vectors are then passed through MCB. The MCB is followed by an element-wise signed square-root and L2 normalization. After MCB pooling, a fully connected layer connects the resulting 16,000-D multimodal representation to the 3,000 top answers.

Attention.

To incorporate spatial information, we use soft attention on our MCB pooling method. Explored by (Xu et al., 2015) for image captioning and by (Xu and Saenko, 2016) and (Yang et al., 2015) for VQA, the soft attention mechanism can be easily integrated in our model.

For each spatial grid location in the visual representation (i.e. last convolutional layer of ResNet [res5c], last convolutional layer of VGG [conv5]), we use MCB pooling to merge the slice of the visual feature with the language representation. As depicted in Figure 3, after the pooling we use two convolutional layers to predict the attention weight for each grid location. We apply softmax to produce a normalized soft attention map. We then take a weighted sum of the spatial vectors using the attention map to create the attended visual representation. We also experiment with generating multiple attention maps to allow the model to make multiple "glimpses" which are concatenated before being merged with the language representation through another MCB pooling for prediction. Predicting attention maps with MCB pooling allows the model to effectively learn how to attend to salient locations based on both the visual and language representations.

Answer Encoding.

For VQA with multiple choices, we can additionally embed the answers. We [Fukui et al., 2016]
<table>
<thead>
<tr>
<th>Model</th>
<th>Test-dev Open Ended</th>
<th>MC</th>
<th>Test-standard Open Ended</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y/N</td>
<td>No.</td>
<td>Other</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>MCB</td>
<td>81.2</td>
<td>35.1</td>
<td>49.3</td>
<td>60.8</td>
</tr>
<tr>
<td>MCB + Genome</td>
<td>81.7</td>
<td>36.6</td>
<td>51.5</td>
<td>62.3</td>
</tr>
<tr>
<td>MCB + Att.</td>
<td>82.2</td>
<td>37.7</td>
<td>54.8</td>
<td>64.2</td>
</tr>
<tr>
<td>MCB + Att. + GloVe</td>
<td>82.5</td>
<td>37.6</td>
<td>55.6</td>
<td>64.7</td>
</tr>
<tr>
<td>MCB + Att. + Genome</td>
<td>81.7</td>
<td>38.2</td>
<td>57.0</td>
<td>65.1</td>
</tr>
<tr>
<td>MCB + Att. + GloVe + Genome</td>
<td>82.3</td>
<td>37.2</td>
<td>57.4</td>
<td>65.4</td>
</tr>
<tr>
<td>Ensemble of 7 Att. models</td>
<td>83.4</td>
<td>39.8</td>
<td>58.5</td>
<td>66.7</td>
</tr>
<tr>
<td>Naver Labs (challenge 2nd)</td>
<td>83.5</td>
<td>39.8</td>
<td>54.8</td>
<td>64.9</td>
</tr>
<tr>
<td>HieCoAtt (Lu et al., 2016)</td>
<td>79.7</td>
<td>38.7</td>
<td>51.7</td>
<td>61.8</td>
</tr>
<tr>
<td>DMN+ (Xiong et al., 2016)</td>
<td>80.5</td>
<td>36.8</td>
<td>48.3</td>
<td>60.3</td>
</tr>
<tr>
<td>FDA (Ilievski et al., 2016)</td>
<td>81.1</td>
<td>36.2</td>
<td>45.8</td>
<td>59.2</td>
</tr>
<tr>
<td>D-NMN (Andreas et al., 2016a)</td>
<td>81.1</td>
<td>38.6</td>
<td>45.5</td>
<td>59.4</td>
</tr>
<tr>
<td>AMA (Wu et al., 2016)</td>
<td>81.0</td>
<td>38.4</td>
<td>45.2</td>
<td>59.2</td>
</tr>
<tr>
<td>SAN (Yang et al., 2015)</td>
<td>79.3</td>
<td>36.6</td>
<td>46.1</td>
<td>58.7</td>
</tr>
<tr>
<td>NMN (Andreas et al., 2016b)</td>
<td>81.2</td>
<td>38.0</td>
<td>44.0</td>
<td>58.6</td>
</tr>
<tr>
<td>AYN (Malinowski et al., 2016)</td>
<td>78.4</td>
<td>36.4</td>
<td>46.3</td>
<td>58.4</td>
</tr>
<tr>
<td>SMem (Xu and Saenko, 2016)</td>
<td>80.9</td>
<td>37.3</td>
<td>43.1</td>
<td>58.0</td>
</tr>
<tr>
<td>VQA team (Antol et al., 2015)</td>
<td>80.5</td>
<td>36.8</td>
<td>43.1</td>
<td>57.8</td>
</tr>
<tr>
<td>DPPnet (Noh et al., 2015)</td>
<td>80.7</td>
<td>37.2</td>
<td>41.7</td>
<td>57.2</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al., 2015)</td>
<td>76.5</td>
<td>35.0</td>
<td>42.6</td>
<td>55.7</td>
</tr>
</tbody>
</table>

Results

Table 4: Open-ended and multiple-choice (MC) results on VQA test set (trained on train+val set) compared with state-of-the-art: accuracy in %. See Sec. 4.4.

Linear pooling has no impact on accuracy compared to full bilinear pooling. Section 3 in Table 1 demonstrates that the MCB brings improvements regardless of the image CNN used. We primarily use ResNet-152 in this paper, but MCB also improves performance if VGG-19 is used. Section 4 in Table 1 shows that our soft attention model works best with MCB pooling. In fact, attending to the Concatenation + FC layer has the same performance as not using attention at all, while attending to the MCB layer improves performance by 2.67 points.

Table 2 compares different values of $d$, the output dimensionality of the multimodal compact bilinear feature. Approximating the bilinear feature with a 16,000-D vector yields the highest accuracy. We also evaluated models with multiple attention maps or channels. One attention map achieves 64.67% accuracy, two 65.08%, and four 64.24% accuracy (trained on train+val). Visual inspection of the generated attention maps reveals that an ensembling or smoothing effect occurs when using multiple maps.

Table 3 presents results for the Visual7W multiple-choice QA task. The MCB with attention model outperforms the previous state-of-the-art by 7.9 points overall and performs better in almost every category.

4.4 Comparison to State-of-the-Art

Table 4 compares our approach with the state-of-the-art on VQA test set. Our best single model uses MCB pooling with two attention maps. Additionally, we augment our training data with images and QA pairs from the Visual Genome dataset. We also concatenate the learned word embedding with pretrained GloVe vectors (Pennington et al., 2014).

Each model in our ensemble of 7 models uses MCB with attention. Some of the models were trained with data from Visual Genome, and some were trained with two attention maps. This ensemble averages across these models to produce the best results, demonstrating an even greater improvement over state-of-the-art models.
Making the V in the VQA matter!

Who is wearing glasses?
- man
- woman

Where is the child sitting?
- fridge
- arms

Is the umbrella upside down?
- yes
- no

How many children are in the bed?
- 2
- 1

[Goyal et al., 2017]
Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>Ans Type</th>
<th>UU</th>
<th>UB</th>
<th>B_{halfB}</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCB [9]</td>
<td>Yes/No</td>
<td>81.20</td>
<td>70.40</td>
<td>74.89</td>
<td>77.37</td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>34.80</td>
<td>31.61</td>
<td>34.69</td>
<td>36.66</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>51.19</td>
<td>47.90</td>
<td>47.43</td>
<td>51.23</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>60.36</td>
<td>54.22</td>
<td>56.08</td>
<td>59.14</td>
</tr>
<tr>
<td>HieCoAtt [25]</td>
<td>Yes/No</td>
<td>79.99</td>
<td>67.62</td>
<td>70.93</td>
<td>71.80</td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>34.83</td>
<td>32.12</td>
<td>34.07</td>
<td>36.53</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>45.55</td>
<td>41.96</td>
<td>42.11</td>
<td>46.25</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>57.09</td>
<td>50.31</td>
<td>51.88</td>
<td>54.57</td>
</tr>
</tbody>
</table>

[Goyal et al., 2017]
Visual Dialog

We introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a question about the image, the agent has to ground the question in image, infer context from history, and answer the question accurately. Visual Dialog is disentangled enough from a specific downstream task so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress. We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). VisDial v0.9 has been released and contains 1 dialog with 10 question-answer pairs on \( \sim 120k \) images from COCO, with a total of \( \sim 1.2M \) dialog question-answer pairs.

We introduce a family of neural encoder-decoder models for Visual Dialog with 3 encoders – Late Fusion, Hierarchical Recurrent Encoder and Memory Network – and 2 decoders (generative and discriminative), which outperform a number of sophisticated baselines. We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a set of candidate answers and evaluated on metrics such as mean-reciprocal-rank of human response. We quantify gap between machine and human performance on the Visual Dialog task via human studies.

Putting it all together, we demonstrate the first 'visual chatbot'! Our dataset, code, trained models and visual chatbot are available on visualdialog.org.

1. Introduction

We are witnessing unprecedented advances in computer vision (CV) and artificial intelligence (AI) – from 'low-level' AI tasks such as image classification \([20]\), scene recognition \([63]\), object detection \([34]\] – to 'high-level' AI tasks such as learning to play Atari video games \([42]\) and Go \([55]\), answering reading comprehension questions by understanding short stories \([21]\, [65]\), and even answering questions about images \([6]\, [39]\, [49]\, [71]\] and videos \([57]\, [58]\]!

What lies next for AI? We believe that the next generation of visual intelligence systems will need to possess the ability to hold a meaningful dialog with humans in natural language about visual content. Applications include:

• Aiding visually impaired users in understanding their surroundings \([7]\) or social media content \([66]\] (AI: 'John just uploaded a picture from his vacation in Hawaii', Human: 'Great, is he at the beach?', AI: 'No, on a mountain').

• Aiding analysts in making decisions based on large quantities of surveillance data (Human: 'Did anyone enter this room last week?', AI: 'Yes, 27 instances logged on camera', Human: 'Were any of them carrying a black bag').

[Das et al., 2017]
Demo

- [http://visualchatbot.cloudcv.org/](http://visualchatbot.cloudcv.org/)

[Das et al., 2017]
**Visual Dialog vs VQA**

**Captioning**
Two people are in a wheelchair and one is holding a racket.

**Visual Dialog**
Q: How many people are on wheelchairs?
A: Two
Q: What are their genders?
A: One male and one female
Q: Which one is holding a racket?
A: The woman

**VQA**
Q: How many people on wheelchairs?
A: Two
Q: How many wheelchairs?
A: One
Video Captioning

Ground truth: A woman is slicing a red pepper.
In this work, we use deep recurrent nets (RNNs), motivated by the work of Donahue et al. (2014), who simultaneously proposed a multimodal MT (Koehn, 2010), RNNs naturally combine with MT (Koehn, 2010), Farhadi et al. (2010), Yao et al. (2010), and some of the earliest works by Aker and Gaizauskas (1996). Propelled by the success of deep learning, several groups released and showed that it works well for describing dynamic scenes and video. Donahue et al. (2014) and Vinyals et al. (2014) from videos. Our framework is based on deep image description models in Donahue et al. (2014); Vinyals et al. (2014) only showed results on a narrow domain where the corpus size is smaller. While Donahue et al. (2014) and transfer the knowledge to the video domain, we generate sentences for open-domains and video. Donahue et al. (2014) and Vinyals et al. (2014) simultaneously proposed a multimodal vector-based representations, such as those for images has received considerable attention, with others. We identify the most likely description for a given video by training a model to maximize the log likelihood of the sentence.

The structure of our video description network is shown in Figure 2. Input videos are encoded with convolutional nets, which use a visual convnet to encode a deep state vector, analog of this model, with an architecture which uses a visual convnet to encode a deep state vector, and an LSTM to decode the vector into a sentence. In contrast to traditional statistical code" the vector into a sentence (i.e. a sequence of words). In this work, we apply the same principle of code" the vector into a sentence (i.e. a sequence of words) and show that it works well for describing dynamic scenes and video.

We extract fc features for each frame, mean pool the features across the entire video and input this at every time step to the LSTM network. The LSTM outputs one time step to the LSTM network. The LSTM outputs one feature at each time step, based on the video features (and typically a Recurrent Neural Network (RNN), to "decode" the representation into a sentence. In con-
Each input video frame is scaled to $256 \times 256$, and is cropped to a random $227 \times 227$.

We use CNNs that are pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.

We apply a convolutional neural network (CNN) to input images and provide the output of the top layer as input to the LSTM. The rest of the LSTM architecture remains unchanged for flow inputs.

In our combined model, we use a shallow fusion technique to integrate flow and RGB features. At each time step of the decoding phase, the model proposes a set of candidate words. We then rescore these hypotheses with the weighted sum of the scores by the flow and RGB networks, and use CNNs that are pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.

We note that, during the decoding phase, the visual frame is used to obtain the emitted word (as: $p_t = \text{argmax} \{ z_t \}$) and video-to-text approaches [3, 4]. We use CNNs that are pretrained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.

We apply a convolutional neural network (CNN) to input images and provide the output of the top layer as input to the LSTM. The rest of the LSTM architecture remains unchanged for flow inputs.

In our combined model, we use a shallow fusion technique to integrate flow and RGB features. At each time step of the decoding phase, the model proposes a set of candidate words. We then rescore these hypotheses with the weighted sum of the scores by the flow and RGB networks, and use CNNs that are pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.

We note that, during the decoding phase, the visual frame is used to obtain the emitted word ($p_t = \text{argmax} \{ z_t \}$) and video-to-text approaches [3, 4]. We use CNNs that are pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.

We apply a convolutional neural network (CNN) to input images and provide the output of the top layer as input to the LSTM. The rest of the LSTM architecture remains unchanged for flow inputs. In our combined model, we use a shallow fusion technique to integrate flow and RGB features. At each time step of the decoding phase, the model proposes a set of candidate words. We then rescore these hypotheses with the weighted sum of the scores by the flow and RGB networks, and use CNNs that are pre-trained on the 1.2M image ILSVRC-2012 object classification subset of the ImageNet dataset [1]. We train on the 16-layer VGG model [2] initialized with weights trained on the Caffe Reference Net (a variant of AlexNet) and also put of the fc7 layer (after applying the ReLU non-linearity) to the LSTM unit. In this work, we report results using the output of the fc6 layer as input.
Hierarchical Encoder

Hierarchical Encoder

(a) Stacked LSTM video encoder

(b) Hierarchical Recurrent Neural Encoder

[Pan et al., 2016]
M-to-M Multi-Task for Video Captioning

Video Encoder

Language Encoder

A man in red shorts is lifting weights.

Video Decoder

Language Decoder

[Pasunuru and Bansal, 2017a]
Guest Research Talk by Ramakanth Pasunuru:

1) Multi-Task Video Captioning with Video and Entailment Generation (ACL 2017)
2) Reinforced Video Captioning with Entailment Rewards (EMNLP 2017)
3) Game-Based Video-Context Dialogue (EMNLP 2018)
Guest Research Talk by Hyounghun Kim:

1) Improving Visual Question Answering by Referring to Generated Paragraph Captions (ACL 2019)
2) Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA (ACL 2020)
Guest Research Talk by Jie Lei:

1) TVQA: Localized, Compositional Video Question Answering (EMNLP 2018)
2) TVQA+: Spatio-Temporal Grounding for Video Question Answering (ACL 2020)
Guest Research Talk by Darryl Hannan (moved to next week):

1) ManyModalQA: Modality Disambiguation and QA over Diverse Inputs (AAAI 2020)
Guest Research Talk by Jaemin Cho (moved to next week):

1) LXMERT: Learning Cross-Modality Encoder Representations from Transformers (EMNLP 2019)
2) X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers (EMNLP 2020)
Topic:
Multi-Task and Reinforcement Learning for Video captioning
Game-Based Video-Context Dialogue

(presented by Ramakanth Pasunuru)
Image Captioning

A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Architecture of Image Captioning Model
Image Captioning with Attention

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

\[ e_{ti} = \text{att}(a_i, h_{t-1}) \]
\[ \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \]
Examples

A woman is throwing a frisbee in a park.
A dog is standing on a hardwood floor.
A stop sign is on a road with a mountain in the background.
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.
A giraffe standing in a forest with trees in the background.
Video Captioning

Applications:

• Assistance to visually impaired

• Improving online video search

• Grounded robotic instruction tasks

Ground truth: A woman is slicing a red pepper.

Ground truth: A group of boys are fighting.

[Kojima et al., 2002; Lee et al., 2008; Khan and Gotoh, 2012; Barbu et al., 2012; Das et al., 2013; Rohrbach et al., 2013; Yu and Siskind, 2013; Venugopalan et al., 2014, 2015, 2016]
Multi-Task for Video Captioning

Video Captioning Challenges:
• Lack of sufficient labeled data
• Spatial-visual modeling
• *Logical* storyline dynamics
• *Temporal* across-frame dynamics

We share knowledge w/ 2 related directed-generation tasks/datasets (textual+visual):

1. Premise-to-Entailment Generation
   • To help learn better caption decoder representations, since caption is also entailed by video.

2. Video-to-Video Generation (Unsupervised)
   • To help learn richer video encoder representations, aware of temporal action context.

Ground truth: A person is mixing powdered ingredients with water.
A woman is mixing flour and water in a bowl.

Our model: A woman is mixing ingredients in a bowl.
Multi-Task for Video Captioning

Video Encoder

Language Encoder

A man in red shorts is lifting weights.

ENTAILMENT GENERATION

Language Decoder

A man is exercising

Video Decoder

UNSUPERVISED VIDEO PREDICTION

(slides by Ramakanth Pasunuru) [Pasunuru & Bansal, ACL 2017]
Baseline Video Captioning Model

- Sequence-to-sequence encoder-decoder model ('f' denotes frames; 'w' denotes words)
- Attention-based (Bahdanau et al., 2015)
- State-of-the-art Inception-v4 image frame features
- Strong baseline (>= previous work)
Textual Entailment

- Directional, logical-implication relation between two sentences:
  - **Premise:** A girl is jumping on skateboard in the middle of a red bridge.
  - Entailment: The girl does a skateboarding trick.
  - Contradiction: The girl skates down the sidewalk.
  - Neutral: The girl is wearing safety equipment.

  - **Premise:** A blond woman is drinking from a public fountain.
  - Entailment: The woman is drinking water.
  - Contradiction: The woman is drinking coffee.
  - Neutral: The woman is very thirsty.

- Can we use entailment as linguistic inference to help related directed/conditioned generation tasks? (Yes, for e.g. video captioning or document summarization).

- Large-scale SNLI corpus allows training accurate classification and RNN-style generation models.
Entailment Generation

- Helps learn better video-entailing caption decoder representations.
- Since caption needs to be entailed by visual premise of video (i.e., describes subsets of objects/events logically implied by full video content), we teach it about entailment via MTL.

[Kolesnyk et al., 2016]

[Pasunuru & Bansal, ACL 2017]
Unsupervised Video Prediction

• Helps learn richer video encoder representations that are aware of temporal context and action sequence/completion.

• Robust to missing frames and varying frame lengths or motion speeds.

• 80:20% frame division between encoder and decoder.

• UCF-101 action videos dataset.
M-to-1 Multi-Task Model

Video Encoder

Language Encoder

A man in red shorts is lifting weights.

Language Decoder

A man is exercising.
1-to-M Multi-Task Model

Video Encoder

LSTM → LSTM → LSTM → LSTM

UNSUPERVISED VIDEO PREDICTION

Video Decoder

LSTM → LSTM → LSTM → LSTM

Language Decoder

A man is exercising
M-to-M Multi-Task for Video Captioning

Training in alternate mini-batches: mixing ratio = \[
\frac{\alpha_v}{\alpha_v + \alpha_f + \alpha_e} : \frac{\alpha_f}{\alpha_v + \alpha_f + \alpha_e} : \frac{\alpha_e}{\alpha_v + \alpha_f + \alpha_e}
\]
## Results (YouTube2Text)

<table>
<thead>
<tr>
<th>Models</th>
<th>METEOR</th>
<th>CIDEr-D</th>
<th>ROUGE-L</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Previous Work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM-YT (Venugopalan et al., 2015b)</td>
<td>26.9</td>
<td>-</td>
<td>-</td>
<td>31.2</td>
</tr>
<tr>
<td>S2VT (Venugopalan et al., 2015a)</td>
<td>29.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Temporal Attention (Yao et al., 2015)</td>
<td>29.6</td>
<td>51.7</td>
<td>-</td>
<td>41.9</td>
</tr>
<tr>
<td>LSTM-E (Pan et al., 2016b)</td>
<td>31.0</td>
<td>-</td>
<td>-</td>
<td>45.3</td>
</tr>
<tr>
<td>Glove + DeepFusion (Venugopalan et al., 2016)</td>
<td>31.4</td>
<td>-</td>
<td>-</td>
<td>42.1</td>
</tr>
<tr>
<td>p-RNN (Yu et al., 2016)</td>
<td>32.6</td>
<td>65.8</td>
<td>-</td>
<td>49.9</td>
</tr>
<tr>
<td>HINRE + Attention (Pan et al., 2016a)</td>
<td>33.9</td>
<td>-</td>
<td>-</td>
<td>46.7</td>
</tr>
<tr>
<td><strong>Our Baselines</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (V)</td>
<td>31.4</td>
<td>63.9</td>
<td>68.0</td>
<td>43.6</td>
</tr>
<tr>
<td>Baseline (G)</td>
<td>31.7</td>
<td>64.8</td>
<td>68.6</td>
<td>44.1</td>
</tr>
<tr>
<td>Baseline (I)</td>
<td>33.3</td>
<td>75.6</td>
<td>69.7</td>
<td>46.3</td>
</tr>
<tr>
<td>Baseline + Attention (V)</td>
<td>32.6</td>
<td>72.2</td>
<td>69.0</td>
<td>47.5</td>
</tr>
<tr>
<td>Baseline + Attention (G)</td>
<td>33.0</td>
<td>69.4</td>
<td>68.3</td>
<td>44.9</td>
</tr>
<tr>
<td>Baseline + Attention (I)</td>
<td>33.8</td>
<td>77.2</td>
<td>70.3</td>
<td>49.9</td>
</tr>
<tr>
<td>Baseline + Attention (I) (E) ⊗</td>
<td>35.0</td>
<td>84.4</td>
<td>71.5</td>
<td>52.6</td>
</tr>
</tbody>
</table>

**Dataset:** 1970 videos with 40 reference captions for each video clip.

**Metrics:** All the above metrics are automatic based on phrase matching between generated and reference caption. For example, BLEU is based on n-gram matching, ROUGE-L is based on longest common subsequence matching.
### Results (YouTube2Text)

<table>
<thead>
<tr>
<th>Models</th>
<th>METEOR</th>
<th>CIDEr-D</th>
<th>ROUGE-L</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Previous Work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM-YT (Venugopalan et al., 2015b)</td>
<td>26.9</td>
<td>-</td>
<td>-</td>
<td>31.2</td>
</tr>
<tr>
<td>S2VT (Venugopalan et al., 2015a)</td>
<td>29.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Temporal Attention (Yao et al., 2015)</td>
<td>29.6</td>
<td>51.7</td>
<td>-</td>
<td>41.9</td>
</tr>
<tr>
<td>LSTM-E (Pan et al., 2016b)</td>
<td>31.0</td>
<td>-</td>
<td>-</td>
<td>45.3</td>
</tr>
<tr>
<td>Glove + DeepFusion (Venugopalan et al., 2016)</td>
<td>31.4</td>
<td>-</td>
<td>-</td>
<td>42.1</td>
</tr>
<tr>
<td>p-RNN (Yu et al., 2016)</td>
<td>32.6</td>
<td>65.8</td>
<td>-</td>
<td>49.9</td>
</tr>
<tr>
<td>HINRE + Attention (Pan et al., 2016a)</td>
<td>33.9</td>
<td>-</td>
<td>-</td>
<td>46.7</td>
</tr>
<tr>
<td><strong>Our Baselines</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (V)</td>
<td>31.4</td>
<td>63.9</td>
<td>68.0</td>
<td>43.6</td>
</tr>
<tr>
<td>Baseline (G)</td>
<td>31.7</td>
<td>64.8</td>
<td>68.6</td>
<td>44.1</td>
</tr>
<tr>
<td>Baseline (I)</td>
<td>33.3</td>
<td>75.6</td>
<td>69.7</td>
<td>46.3</td>
</tr>
<tr>
<td>Baseline + Attention (V)</td>
<td>32.6</td>
<td>72.2</td>
<td>69.0</td>
<td>47.5</td>
</tr>
<tr>
<td>Baseline + Attention (G)</td>
<td>33.0</td>
<td>69.4</td>
<td>68.3</td>
<td>44.9</td>
</tr>
<tr>
<td>Baseline + Attention (I)</td>
<td>33.8</td>
<td>77.2</td>
<td>70.3</td>
<td>49.9</td>
</tr>
<tr>
<td>Baseline + Attention (I) (E) ⊗</td>
<td>35.0</td>
<td>84.4</td>
<td>71.5</td>
<td>52.6</td>
</tr>
<tr>
<td><strong>Our Multi-Task Learning Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⊗ + Video Prediction (1-to-M)</td>
<td>35.6</td>
<td>88.1</td>
<td>72.9</td>
<td>54.1</td>
</tr>
<tr>
<td>⊗ + Entailment Generation (M-to-1)</td>
<td>35.9</td>
<td>88.0</td>
<td>72.7</td>
<td>54.4</td>
</tr>
<tr>
<td>⊗ + Video Prediction + Entailment Gener (M-to-M)</td>
<td>36.0</td>
<td>92.4</td>
<td>72.8</td>
<td>54.5</td>
</tr>
</tbody>
</table>

**Human evaluation:** Multi-task model is better than baseline.

* All models (1-to-M, M-to-1 and M-to-M) stat. signif. better than strong SotA baseline.
Results (Entailment Generation)

• Video captioning mutually also helps improve the entailment-generation task in turn (w/ statistical significance).

<table>
<thead>
<tr>
<th>Models</th>
<th>M</th>
<th>C</th>
<th>R</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entailment Generation</td>
<td>29.6</td>
<td>117.8</td>
<td>62.4</td>
<td>40.6</td>
</tr>
<tr>
<td>+Video Caption (M-to-1)</td>
<td><strong>30.0</strong></td>
<td><strong>121.6</strong></td>
<td><strong>63.9</strong></td>
<td><strong>41.6</strong></td>
</tr>
</tbody>
</table>
**Analysis Example**

**Ground truth:** Two men are fighting.
A group of boys are fighting.

**Baseline model:** A group of men are dancing.

**Multi-task model:** Two men are fighting.

*Complex example where the multi-task model performs better than baseline.*
We introduce a novel entailment-enhanced reward (CIDEnt) that corrects phrase-matching based metrics (such as CIDEr) to only allow for logically-implied partial matches and avoid contradictions.
Our mixed loss is defined as:

$$\text{MixLoss} = \alpha \cdot \text{CE} + (1 - \alpha) \cdot \text{RL}$$

where CE is the cross-entropy loss and RL is the reinforcement loss. This formulation ensures that we only trust the CIDEr (as opposed to a contradiction or neutral) when the entailment score is high, and otherwise, we penalize the metric reward score by de-regularizing the metrics to ensure better read-ability and fluency of the generated caption.

### Entailment Corrected Reward

We address the above issue by using an entailment-based reward in cases when the entailment score is low. Thus, ensuring the generated caption logically implies (i.e., is paraphrase or directed partial match w/) ground-truth caption.

$$\text{CIDEr}_{\text{Ent}} = \begin{cases} 
\text{CIDEr} - \lambda, & \text{if } \text{Ent} < \beta \\
\text{CIDEr}, & \text{otherwise}
\end{cases}$$

**Table:** Examples of captions sampled during policy gradient and their CIDEr vs Entailment scores.

<table>
<thead>
<tr>
<th>Ground-truth caption</th>
<th>Generated (sampled) caption</th>
<th>CIDEr</th>
<th>Ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a man is spreading some butter in a pan</td>
<td>puppies is melting butter on the pan</td>
<td>140.5</td>
<td>0.07</td>
</tr>
<tr>
<td>a panda is eating some bamboo</td>
<td>a panda is eating some fried</td>
<td>256.8</td>
<td>0.14</td>
</tr>
<tr>
<td>a monkey pulls a dogs tail</td>
<td>a monkey pulls a woman</td>
<td>116.4</td>
<td>0.04</td>
</tr>
<tr>
<td>a man is cutting the meat</td>
<td>a man is cutting meat into potato</td>
<td>114.3</td>
<td>0.08</td>
</tr>
<tr>
<td>the dog is jumping in the snow</td>
<td>a dog is jumping in cucumbers</td>
<td>126.2</td>
<td>0.03</td>
</tr>
<tr>
<td>a man and a woman is swimming in the pool</td>
<td>a man and a whale are swimming in a pool</td>
<td>192.5</td>
<td>0.02</td>
</tr>
</tbody>
</table>

penalizing CIDEr reward when entailment score is low. Thus, ensuring the generated caption logically implies (i.e., is paraphrase or directed partial match w/) ground-truth caption.
# Results (MSR-VTT)

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr-D</th>
<th>CIDEnt</th>
<th>Human*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Previous Work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venugopalan (2015b)*</td>
<td>32.3</td>
<td>23.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yao et al. (2015)*</td>
<td>35.2</td>
<td>25.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xu et al. (2016)</td>
<td>36.6</td>
<td>25.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pasunuru and Bansal (2017)</td>
<td>40.8</td>
<td>28.8</td>
<td>60.2</td>
<td>47.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank1: v2t_navigator</td>
<td>40.8</td>
<td>28.2</td>
<td>60.9</td>
<td>44.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank2: Aalto</td>
<td>39.8</td>
<td>26.9</td>
<td>59.8</td>
<td>45.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank3: VideoLAB</td>
<td>39.1</td>
<td>27.7</td>
<td>60.6</td>
<td>44.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Our Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-Entropy (Baseline-XE)</td>
<td>38.6</td>
<td>27.7</td>
<td>59.5</td>
<td>44.6</td>
<td>34.4</td>
<td>-</td>
</tr>
<tr>
<td>CIDEr-RL</td>
<td>39.1</td>
<td>28.2</td>
<td>60.9</td>
<td>51.0</td>
<td>37.4</td>
<td>11.6</td>
</tr>
<tr>
<td>CIDEnt-RL (New Rank1)</td>
<td>40.5</td>
<td>28.4</td>
<td>61.4</td>
<td>51.7</td>
<td>44.0</td>
<td>18.4</td>
</tr>
</tbody>
</table>

**Table:** Our primary video captioning results on MSR-VTT (CIDEnt-RL is stat. significantly better than CIDEr-RL in all metrics, and CIDEr-RL is better than Baseline-XE).
Game-Based Video-Context Dialogue
Visual Context

Image-based Context

[Das et al., 2017]

[Mostafazadeh et al., 2017]

[De Vries et al., 2017]

[Celikyilmaz et al., 2014]
Our Twitch-FIFA Dataset
Our Twitch-FIFA Dataset

Video + Chat based Context

Multiple speakers
The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right).
Task

The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)

Applications of Video-Grounded Dialogue

- Personal Assistants
- Intelligent tutors
- Human-robot Collaboration
Twitch-FIFA Dataset Statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Videos</td>
<td>33</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Total Hours</td>
<td>58.4</td>
<td>11.9</td>
<td>15.4</td>
</tr>
<tr>
<td>Final Filtered #Instances</td>
<td>10,510</td>
<td>2,153</td>
<td>2,780</td>
</tr>
<tr>
<td>Avg. Chat Context Length</td>
<td>69.0</td>
<td>63.5</td>
<td>71.2</td>
</tr>
<tr>
<td>Avg. Response Length</td>
<td>6.5</td>
<td>6.5</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Twitch-FIFA dataset’s chat statistics (lengths are defined in terms of number of words)

- Anonymized user identities
Discriminative Model

Our **Triple Encoder** discriminative model with bidirectional LSTM-RNN encoders for video, chat context, and response

\[ f = \sigma([h^v_f; h^u_f; h^r_f]W h^r_f + b) \]
Our **Tri-Directional Attention Flow (TriDAF)** model with all pairwise modality attention modules, as well as self attention on video context, chat context, and response as inputs.
Results

<table>
<thead>
<tr>
<th>Models</th>
<th>r@1</th>
<th>r@2</th>
<th>r@5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BASELINES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most-Frequent-Response</td>
<td>10.0</td>
<td>16.0</td>
<td>20.9</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>9.6</td>
<td>20.9</td>
<td>51.5</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>10.8</td>
<td>21.8</td>
<td>52.5</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>11.4</td>
<td>22.6</td>
<td>53.2</td>
</tr>
<tr>
<td>Chat-Response-Cosine</td>
<td>11.4</td>
<td>22.0</td>
<td>53.2</td>
</tr>
<tr>
<td><strong>DISCRIMINATIVE MODEL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dual Encoder (C)</td>
<td>17.1</td>
<td>30.3</td>
<td>61.9</td>
</tr>
<tr>
<td>Dual Encoder (V)</td>
<td>16.3</td>
<td>30.5</td>
<td>61.1</td>
</tr>
<tr>
<td>Triple Encoder (C+V)</td>
<td>18.1</td>
<td>33.6</td>
<td>68.5</td>
</tr>
<tr>
<td>TriDAF+Self Attn (C+V)</td>
<td>20.7</td>
<td>35.3</td>
<td>69.4</td>
</tr>
</tbody>
</table>

Performance of our baselines and discriminative models for recall@$k$ metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.
Thank You
Visual QA
Improving Visual Question Answering by Referring to Generated Paragraph Captions

Hyounghun Kim and Mohit Bansal
University of North Carolina at Chapel Hill
Main Idea

1. Visual question answering is a task to answer diverse questions about images.

2. In order to answer all the questions successfully, the ability to understand different aspects of an image is required.
Main Idea

- Image captioning task is to describe contents or topics from images.
- Singe-sentence captions usually focus on obvious and the most salient part of an image, so tend to describe similar contents.
- On the other hand, paragraph captions contain diverse aspects of an image.

<table>
<thead>
<tr>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) A girl is eating donuts with a boy in a restaurant</td>
</tr>
<tr>
<td>2) A boy and girl sitting at a table with doughnuts.</td>
</tr>
<tr>
<td>3) Two kids sitting a coffee shop eating some frosted donuts</td>
</tr>
<tr>
<td>4) Two children sitting at a table eating donuts.</td>
</tr>
<tr>
<td>5) Two children eat doughnuts at a restaurant table.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.</td>
</tr>
</tbody>
</table>
Visual Questions vs. Paragraph Captioning (Ground-Truth)

Q. How many planes are in the sky? / A. One
Q. What color are the trees? / A. Green
Q. What color is the plane? / A. White and blue
Q. What color is the sky? / A. Cream and gray
Q. What is in the sky? / A. The plane
Q. What color are the tires? / A. Black
Q. Where was the picture taken? / A. At an airport

GT Paragraph Caption: “The image is of a plane taking off on a runway. There are two planes in the background on the tarmac and one in the sky that has just taken off and is at a very low altitude. The plane that has just taken off is white with blue and gray stripes on it and white writing on the tail. There are trees on the outside of the airport and it is sunset.”
Q. How many planes are in the sky? / A. One
Q. What color are the trees? / A. Green
Q. What color is the plane? / A. White and blue
Q. What color is the sky? / A. Cream and gray
Q. What is in the sky? / A. The plane
Q. What color are the tires? / A. Black
Q. Where was the picture taken? / A. At an airport

GT Paragraph Caption: “The image is of a plane taking off on a runway. There are two planes in the background on the tarmac and one in the sky that has just taken off and is at a very low altitude. The plane that has just taken off is white with blue and gray stripes on it and white writing on the tail. There are trees on the outside of the airport and it is sunset.”
Q. How many planes are in the sky? / A. One
Q. What color are the trees? / A. Green
Q. What color is the plane? / A. White and blue
Q. What color is the sky? / A. Cream and gray
Q. What is in the sky? / A. The plane
Q. What color are the tires? / A. Black
Q. Where was the picture taken? / A. At an airport

Generated Paragraph Caption: “A plane is on the runway. The plane is white. The plane is a plane. The airplane is white. The tail of the plane is red. The sky is very cloudy. The clouds are white. There are trees on the ground. The planes are white and blue. The sky is blue.”
Manual Investigation (Human Evaluation)

<table>
<thead>
<tr>
<th>Text</th>
<th>% of Answerable Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth Caption</td>
<td>55.00</td>
</tr>
<tr>
<td>Generated Paragraph Caption</td>
<td>42.67</td>
</tr>
</tbody>
</table>

- Choose random 300 questions.
- Count the questions that can be answered only with text material.

- These results are evidence that a paragraph caption can help VQA task if it can be integrated into VQA model in appropriate ways.
- Paragraph captions provide intermediate textual symbolic evidence for clues.
VTQA Model

VTQA (VQA + TextQA) with Early, Late, and Later Fusion

(slides by Hyoung Hun Kim)
A man is sitting on the snow. The man is wearing a black jacket. The man is wearing black pants.

Snow is white. Sky is blue. Tree is green.

"Where is the man sitting?"

"On the snow"

(slides by Hyounghun Kim)
A man is sitting on the snow. The man is wearing a black jacket. The man is wearing black pants.

Paragraph Caption

Early Fusion

Visual Feature

Faster R-CNN

Object Properties

Paragraph Caption

Paragraph Caption with RL

Question Feature

Late Fusion

Answer Recommendation

Object Properties

Visual Feature

Faster R-CNN

Early Fusion

"Where is the man sitting?"

"On the snow"

ACL 2019

(slides by Hyounghun Kim)
Paragraph Captioning

- Paragraph Captioning Model (Melas-Kyriazi et al., 2018).¹
- Trained with RL using CIDEr-D metric as a reward.
- Repetition penalty applied.
- We will discuss more rewards we tried (saliency, #objects, VQA accuracy) later.

VTQA Model

A man is sitting on the snow.

The man is wearing a black jacket.

The man is wearing black pants.

Snow is white.

Sky is blue.

Tree is green.

Paragraph Caption

Visual Feature

Object Properties

Early Fusion

Question Feature

Attention

Late Fusion

Para-Capt. Model with RL

"Where is the man sitting?"

"On the snow"

Answer Recommendation

man, snowboard, snow, trees, sky, white...

Softmax
Faster R-CNN
- Detects objects in an image.
- Extracts visual features from each object.

Cross-Attention
- Creates a similarity matrix between visual features and paragraph caption features.
- According to the matrix, relevant features are selected with weights.

Object Property
- Encoded with GRU and concatenated to visual features.
A man is sitting on the snow. The man is wearing a black jacket. The man is wearing black pants.

"Where is the man sitting?"

"On the snow"
"Where is the man sitting?"

- **Attention**
  - Question is encoded with GRU.
  - Attention is applied over features from early fusion module w.r.t. a question feature.

- **Consensus**
  - Each module plays as a voter.
  - The answers that get high scores from multiple voters have a high chance to be selected as the final answer.
VTQA Model

A man is sitting on the snow. The man is wearing a black jacket. The man is wearing black pants.

Paragraph Caption

Question Feature

Attention

Late Fusion

Object Properties

Early Fusion

Para-Capt. Model with RL

Para-Capt. Model with RL

"Where is the man sitting?"

"On the snow"

ACL 2019
Later Fusion (Answer Recommendation)

- **Object Property**
  - Properties from detected objects can be considered as recommended answers.
  - man, snowboard, white…

- **Extra Score**
  - An extra score is added to those recommended answers.
  - Extra score = $c \times$ standard deviation over all original scores. $c$ is tuned to 1 using validation dataset.

<table>
<thead>
<tr>
<th>Extra Score</th>
<th>0.05</th>
<th>0</th>
<th>0.05</th>
<th>0</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Score</td>
<td>0.15</td>
<td>0.01</td>
<td>0.21</td>
<td>0.23</td>
<td>…</td>
</tr>
</tbody>
</table>

(slides by Hyoughun Kim)
Dataset

- Question-answer pairs from Visual Genome\(^1\)
- Paragraph caption annotations from Krause et al. (2017)\(^2\)
- We follow the image splits of Krause et al. (2017) and exclude those who do not have question-answer pair
- So, the final question-answer pairs split:
  
  171,648 / 29,759 / 29,490 (train / validation / test)

---

\(^1\)Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations  Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia-Li, David Ayman Shamma, Michael Bernstein, Li Fei-Fei

### Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA baseline</td>
<td>44.68</td>
</tr>
<tr>
<td>VQA + MFB baseline</td>
<td>44.94</td>
</tr>
<tr>
<td>VTQA (full model)</td>
<td><strong>46.86</strong></td>
</tr>
</tbody>
</table>

- Run each model 5 times and average them.
- MFB: Multimodal Factorized Bilinear pooling

- Our VTQA model stat. significantly outperforms the baseline VQA model ($p < 0.001$).
- Applied MFB (which is employed in near state-of-the-art models) for comparing with stronger baseline.

---

Ablation Study

<table>
<thead>
<tr>
<th>Model</th>
<th>Val Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. VTQA + EF (base model)</td>
<td>45.41</td>
</tr>
<tr>
<td>2. VTQA + EF + LF</td>
<td>46.36</td>
</tr>
<tr>
<td>3. VTQA + EF + AR</td>
<td>46.95</td>
</tr>
<tr>
<td>4. VTQA + EF + LF + AR</td>
<td>47.60</td>
</tr>
</tbody>
</table>

• EF: Early Fusion, LF: Late Fusion, AR: Answer Recommendation.

- Our LF improves the accuracy by 0.95% (from 1 to 2).
- Our AR improves the accuracy by 1.54% (from 1 to 3) and 1.24% (from 2 to 4).
TextQA Model

<table>
<thead>
<tr>
<th>TextQA</th>
<th>Val Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>43.96</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.07</td>
</tr>
</tbody>
</table>

- GT: Ground-Truth

(slides by Hyoung hun Kim)
# TextQA Model vs. Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Val Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>43.96</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>55.00</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.67</td>
</tr>
</tbody>
</table>

- GT: Ground-Truth

(slides by Hyounghun Kim)
TextQA Model vs. Human Evaluation

<table>
<thead>
<tr>
<th>TextQA</th>
<th>Val Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>43.96</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.07</td>
</tr>
<tr>
<td>Human Eval.</td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>GT Para-Capt.</td>
<td>55.00</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.67</td>
</tr>
</tbody>
</table>

- GT: Ground-Truth

- But, our model does not seem to fully extract information from GT paragraph caption now (43.96 vs. 55.00).
TextQA Model vs. Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Val Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>43.96</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Para-Capt.</td>
<td>55.00</td>
</tr>
<tr>
<td>Generated Para-Capt.</td>
<td>42.67</td>
</tr>
</tbody>
</table>

- GT: Ground-Truth

- But, our model does not seem to fully extract information from GT paragraph caption now (43.96 vs. 55.00).
- Also, generated paragraph captions are not good enough to give useful information compared to GT paragraph captions.
Attention Visualization

Q: how many glasses are in the picture
A: 2

The paragraph contains a direct clue for the question:
→ “there are two glasses on the table”

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer
Q: where was the photo taken  
A: in kitchen

The paragraph contains a direct clue for the question → “a young girl is standing in the kitchen”

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer
Attention Visualization

Q: what is being cooked
A: hot dogs

The paragraph contains a clue that help infer the answer
→ “there is a **hot dog on the grill**”

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer
Q: what is the crowd watching  
A: tennis match

The paragraph has a couple of sentences that give indirect clues:

→ “a man is standing on a tennis court playing tennis”
→ “the tennis court is blue and white”
→ “the spectators are sitting in the stands watching the game”
Q: what is the girl holding
Ground-truth A: bag
Model's A: suitcase

The paragraph misleads the model to the wrong answer
→ “the woman is holding a suitcase”
Q: what is the man riding
Ground-truth A: bike
Model's A: frisbee

The paragraph misleads the model to the nearest wrong answer
→ “the man is holding a white frisbee”
Video QA
Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA

(slides by Hyounghun Kim)
Task / Dataset

We explore the TVQA dataset in this paper.

The TVQA dataset consists of question and answer pairs, video frames, and corresponding subtitles.

The task is to choose the correct answer among 5 candidates.

Motivation

• We present 3 contributions to improve video+dialogue QA:

1. Dense-captions have diverse visual clues in the symbolic textual form. Thus, help provide cues for answering questions by matching/aligning keyword/phrase.

2. Frame gates pass relevant frames which has useful information for answering questions.

3. Localization/frame selection task can be cast as multi-label classification task and allow applying new customized losses.

(slides by Hyounghun Kim)
A: Petting a dog

Q: What is Castle doing when Kate pulls up in her car?

Beckett: That's too bad. You two make a cute couple.

Beckett: What's up, Castle? You proposing?

Oh, no. Just waiting for you.

Beckett: What's up, Castle? You proposing?

A: Upstairs.

House downstairs?

Unkname: Carol!

Esposito: Upstairs. Go.
• The question and each of the 5 answer candidates are concatenated to create 5 QA pairs.

• Subtitles and dense captions are aligned with each frame which is extracted at 0.5 fps.

• All the input features are encoded with the convolutional layer.
First, the QAs are aligned with subtitle and video in the word/object level.

Next, the fused features are aligned again in the frame level.

The same dual-level attention is done with dense captions in the place of the video feature.
- After the dual-level attention, we obtain two fused features: QA-SUB-VID and QA-SUB-DENSE.
- They are complementary, so need to be integrated.
- We use multi-head self-attention to combine them.
The integrated feature still needs to be filtered to get more relevant information and select useful spatial-temporal information.

Thus, frame-selection gates are applied.

We use max-pooling, global gate, and local gate which is supervised by human importance annotations.
**Model**

**Features**
- Video
- Q-A
- Subtitle

**Dual-Level Attention**
- Word/Object Level Att.
- Frame-Level Att.

**Video-DenseCapt. Integration**
- Multi-Heads Self-Cross Att.

**Frame-Selection Gates**
- Max-Pool
- Global Gate
- Local Gate
- Frame Score Margin

**New Loss Supervision**
- [IOFSM/BBCE]

### Loss Functions

**IOFSM:**
\[
\text{loss}_{io} = 1 + \text{Avg}(\text{OFS}) - \text{Avg}(\text{IFS})
\]

**BBCE:**
\[
\text{loss}_{bbce} = - \left( \sum_i \log(s_i^f) / T_{Fin} \right) + \sum_j \log(1 - s_j^f) / T_{Fout}
\]

- To give higher weights to the relevant frames (in-frame) and lower weights to the less important frames (out-frame), we introduce new loss, In-and-out frame score margin (IOFSM).
- Also, for more balanced training signal, we introduce Balanced Binary Cross-Entropy (BBCE).
### Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test-Public (%)</th>
<th>Val (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>bbt</td>
</tr>
<tr>
<td>1 jacobssy (anonymous)</td>
<td>66.01</td>
<td>68.75</td>
</tr>
<tr>
<td>2 multi-stream (Lei et al., 2018)</td>
<td>66.46</td>
<td>70.25</td>
</tr>
<tr>
<td>3 PAMN (Kim et al., 2019b)</td>
<td>66.77</td>
<td>-</td>
</tr>
<tr>
<td>4 Multi-task (Kim et al., 2019a)</td>
<td>67.05</td>
<td>-</td>
</tr>
<tr>
<td>5 ZGF (anonymous)</td>
<td>68.77</td>
<td>-</td>
</tr>
<tr>
<td>6 STAGE (Lei et al., 2020)</td>
<td>70.23</td>
<td>-</td>
</tr>
<tr>
<td>7 akalsdnr (anonymous)</td>
<td>70.52</td>
<td>71.49</td>
</tr>
<tr>
<td>8 Ours (hstar)</td>
<td>74.09</td>
<td>74.04</td>
</tr>
</tbody>
</table>

- Our model outperforms the state-of-the-art models by a large margin on both validation and test-public splits.
- Also, our model’s scores across the TV shows are more balanced than any other models.
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test-Public (%)</th>
<th>Val (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>bbt</td>
</tr>
<tr>
<td>1 jacobssy (anonymous)</td>
<td>66.01</td>
<td>68.75</td>
</tr>
<tr>
<td>2 multi-stream (Lei et al., 2018)</td>
<td>66.46</td>
<td>70.25</td>
</tr>
<tr>
<td>3 PAMN (Kim et al., 2019b)</td>
<td>66.77</td>
<td>-</td>
</tr>
<tr>
<td>4 Multi-task (Kim et al., 2019a)</td>
<td>67.05</td>
<td>-</td>
</tr>
<tr>
<td>5 ZGF (anonymous)</td>
<td>68.77</td>
<td>-</td>
</tr>
<tr>
<td>6 STAGE (Lei et al., 2020)</td>
<td>70.23</td>
<td>-</td>
</tr>
<tr>
<td>7 akalsdnr (anonymous)</td>
<td>70.52</td>
<td>71.49</td>
</tr>
<tr>
<td>8 Ours (hstar)</td>
<td><strong>74.09</strong></td>
<td><strong>74.04</strong></td>
</tr>
</tbody>
</table>

- Our model outperforms the state-of-the-art models by a large margin on both validation and test-public splits.
- Also, our model’s scores across the TV shows are more balanced than any other models.
### Ablation

<table>
<thead>
<tr>
<th>Model</th>
<th>Val Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Single-Att + Frame-Span</td>
<td>69.86</td>
</tr>
<tr>
<td>2 Single-Att + Frame-Selection Gates</td>
<td>70.08</td>
</tr>
<tr>
<td>3 Dual-Att + Frame-Span</td>
<td>70.20</td>
</tr>
<tr>
<td>4 Dual-Att + Frame-Selection Gates (w/o NewLoss)</td>
<td>71.26</td>
</tr>
<tr>
<td>5 Dual-Att + Frame-Selection Gates</td>
<td>72.51</td>
</tr>
<tr>
<td>6 Dual-Att + Frame-Selection Gates (w/o NewLoss) + RoBERTa</td>
<td>72.53</td>
</tr>
<tr>
<td>7 Dual-Att + Frame-Selection Gates + RoBERTa</td>
<td>73.34</td>
</tr>
<tr>
<td>8 Dual-Att + Frame-Selection Gates + RoBERTa + DenseCapts</td>
<td>74.20</td>
</tr>
</tbody>
</table>

- Each component of our model helps increase performance.

- Especially, our new losses, OFSM and BBCE (row 5 vs 4: $p < 0.0001$, row 7 vs 6: $p < 0.005$), and dense captions (row 8 vs 7: $p < 0.005$) improve performance significantly.

- Using IOFSM alone decreases the score by increasing OFS std.

- Using IOFSM+BBCE increases the score by increasing avg. IFS.

<table>
<thead>
<tr>
<th>Loss</th>
<th>Val Score (%)</th>
<th>IFS avg</th>
<th>IFS std</th>
<th>OFS avg</th>
<th>OFS std</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 BCE</td>
<td>71.26</td>
<td>0.468</td>
<td>0.108</td>
<td>0.103</td>
<td>0.120</td>
</tr>
<tr>
<td>2 IOFSM</td>
<td>70.75</td>
<td>0.739</td>
<td>0.127</td>
<td>0.143</td>
<td>0.298</td>
</tr>
<tr>
<td>3 BCE+IOFSM</td>
<td>72.22</td>
<td>0.593</td>
<td>0.128</td>
<td>0.111</td>
<td>0.159</td>
</tr>
<tr>
<td>4 BBCE</td>
<td>72.27</td>
<td>0.759</td>
<td>0.089</td>
<td>0.230</td>
<td>0.231</td>
</tr>
<tr>
<td>5 BBCE+IOFSM</td>
<td>72.51</td>
<td>0.764</td>
<td>0.098</td>
<td>0.182</td>
<td>0.246</td>
</tr>
</tbody>
</table>
Visualization (word/object level att.)

- Dense captions help localize the relevant frame by matching keyword/phrase (e.g., “a woman sitting”, “holding a glass”).

- Subtitles also help answer the question by providing a nearly exact clue for the answer (i.e., “… anything about acting.”).

- Object level attention helps by aligning the word in the QA and the object feature in a video frame (i.e., the woman’s hand and ‘sign’ in the QA).
Frame-level attention can align relevant frames from different features.

In the example, to answer the question, the model needs to find a frame in which ‘he (Esposito) searched Carol’s house downstairs’, then find a frame which has a clue for ‘where did Esposito search’. Our frame-level attention can properly align the information fragments from different features (Frame 20 and 25) to help answer questions.
Visualization (new losses)

- Our new losses (IOFSM+BBCE) changes the score distribution over all frames.

- Before applying our losses (left side), overall scores are relatively low. After using the losses (right side), overall scores are increased, and especially, scores around in-frames get much higher.
Visual Dialog
Model Bias for VisDial: We show that models have different behaviors on VisDial (Das et al. 2017) when being evaluated on different metrics (NDCG, MMR, recall@k, etc.).

Image-Only Model: Our image-only model gives higher NDCG score which could imply that image-only model is better at generalization.

Image-History Joint Model: Our image-history joint model has higher non-NDCG (MMR, recall@k, etc.) which could imply that image-history model is good at keyword matching / memorizing to give accurate answers.

Combined Fusion Model: Explicitly maintain two models; complementary abilities for a more balanced multimodal model.

<table>
<thead>
<tr>
<th>Models</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULL</td>
<td>57.81</td>
<td>64.47</td>
<td>50.87</td>
<td>81.38</td>
<td>90.03</td>
<td>4.10</td>
</tr>
<tr>
<td>H-5</td>
<td>58.24</td>
<td>64.29</td>
<td>50.61</td>
<td>81.35</td>
<td>90.22</td>
<td>4.10</td>
</tr>
<tr>
<td>H-1</td>
<td>59.29</td>
<td>62.86</td>
<td>49.07</td>
<td>79.76</td>
<td>89.08</td>
<td>4.35</td>
</tr>
<tr>
<td>Img-only</td>
<td>61.04</td>
<td>61.25</td>
<td>47.18</td>
<td>78.43</td>
<td>88.17</td>
<td>4.61</td>
</tr>
</tbody>
</table>

- Performance of models with different amounts of history.
- The more history the less NDCG (vice versa).

Consensus Dropout Fusion (CDF):
- Dropout the final features from the joint model randomly.
- Modulates the influence of the joint model.

<table>
<thead>
<tr>
<th>Models</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Img-Only</td>
<td>61.04</td>
<td>61.25</td>
<td>47.18</td>
<td>78.43</td>
<td>88.17</td>
<td>4.61</td>
</tr>
<tr>
<td>Joint</td>
<td>58.97</td>
<td>64.57</td>
<td>50.87</td>
<td>81.58</td>
<td>90.30</td>
<td>4.05</td>
</tr>
<tr>
<td>CDF</td>
<td>59.93</td>
<td>64.52</td>
<td>50.92</td>
<td>81.31</td>
<td>90.00</td>
<td>4.10</td>
</tr>
<tr>
<td>Ensemble</td>
<td>61.20</td>
<td>64.67</td>
<td>51.00</td>
<td>81.60</td>
<td>90.37</td>
<td>4.03</td>
</tr>
</tbody>
</table>

- CDF model shows more balanced results while the ensemble model takes the best from both models.

Hyounghun Kim, Hao Tan, Mohit Bansal

University of North Carolina at Chapel Hill

Table 5: Performance comparison between our models and other models on the test-standard dataset of VisDial v1.0. We run

<table>
<thead>
<tr>
<th>Models</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF (Das et al. 2017)</td>
<td>45.31</td>
<td>55.42</td>
<td>40.95</td>
<td>72.45</td>
<td>82.83</td>
<td>5.95</td>
</tr>
<tr>
<td>HRE (Das et al. 2017)</td>
<td>45.46</td>
<td>54.16</td>
<td>39.93</td>
<td>70.45</td>
<td>81.50</td>
<td>6.41</td>
</tr>
<tr>
<td>MN (Das et al. 2017)</td>
<td>47.50</td>
<td>55.49</td>
<td>40.98</td>
<td>72.30</td>
<td>83.30</td>
<td>5.92</td>
</tr>
<tr>
<td>MN-att (Das et al. 2017)</td>
<td>49.58</td>
<td>56.90</td>
<td>42.43</td>
<td>74.00</td>
<td>84.35</td>
<td>5.59</td>
</tr>
<tr>
<td>LF-att (Das et al. 2017)</td>
<td>49.76</td>
<td>57.07</td>
<td>42.08</td>
<td>74.83</td>
<td>85.05</td>
<td>5.41</td>
</tr>
<tr>
<td>CoreDMN (Kottur et al. 2018)</td>
<td>54.7</td>
<td>61.5</td>
<td>47.55</td>
<td>78.10</td>
<td>88.80</td>
<td>4.40</td>
</tr>
</tbody>
</table>

Table 6: The effect of round dropout: applying round dropout improves model's performance on NDCG by around 1.2 while also improving other metrics. (CA: cross-attention model (base model), RD: round dropout).

<table>
<thead>
<tr>
<th>Models</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Img+Img</td>
<td>61.97</td>
<td>62.24</td>
<td>48.20</td>
<td>79.49</td>
<td>88.83</td>
<td>4.41</td>
</tr>
<tr>
<td>Joint+Joint</td>
<td>59.84</td>
<td>65.60</td>
<td>52.06</td>
<td>82.46</td>
<td>90.87</td>
<td>3.88</td>
</tr>
<tr>
<td>Img+Joint</td>
<td>61.50</td>
<td>65.04</td>
<td>51.38</td>
<td>81.93</td>
<td>90.45</td>
<td>3.96</td>
</tr>
</tbody>
</table>

• Performance of ensemble models with different combinations. Img+Img model has the highest value of NDCG while Joint+Joint model has the highest values for other metrics. Img+Joint model has more balanced results.

- Evaluation on test-standard dataset.

- Consensus dropout fusion and different dropout rates. With different dropout rates, consensus dropout fusion model yields different scores of all metrics.

• Notes on Table 6:
  - As shown in Table 6, our round dropout model (p=0.25) outperforms the I+I model. This might imply that the balanced scores (NDCG: 59.90, MRR: 64.05, R@1: 50.28, R@5: 80.95, R@10: 90.60, Mean: 4.00) are in between our consensus dropout fusion model (see Sec.3.4) with different model combinations. Img+Img model (3 Img models) has highest values for other metrics. Img+Joint model has more balanced results.

- Figures:
  - Figure 5: Coreference and memorization examples of the image-history joint model (a darker blue square indicates a higher attention score). Note that attention scores for captions are always high since they have more general information than others. On the right, we show two examples for answer prediction of the image-only model.
  - Figure 6: Ablation Study of each model (image-only and joint models) over the complementary relation between image-only and joint models to build ensemble models. This result shows that our image-only, joint, and ensemble models (I+I and J+J) are comparable to same-kind ensemble models (I+J ensemble model are comparable to same-kind ensemble models (I+I and J+J). To be specific, for the NDCG metric, I+I model has highest values for other metrics. Img+Joint model has more balanced results.

(slides by Hyounghun Kim)

AAAI 2020
Thank you
Outline

• Video Question Answering
  • TVQA: Localized, Compositional Video Question Answering, EMNLP 2018

• Language-driven Moment Localization
  • TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, ECCV 2020

• Future Event Prediction
  • What is More Likely to Happen Next? Video-and-Language Future Event Prediction, EMNLP 2020

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)
TVQA: Localized, Compositional Video Question Answering

tvqa.cs.unc.edu

Jie Lei, Licheng Yu, Mohit Bansal, Tamara L. Berg

EMNLP 2018, Brussels, Belgium

(Slides by Jie Lei)
What is Sheldon holding when he is talking to Howard about the sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink
What is Sheldon holding when he is talking to Howard about the sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink
Write a question:

[What/Why/... ] ___ [when/before/after] ___

Question + Localization

What is Sheldon holding when he is talking to Howard about sword?
What is Sheldon holding when he is talking to Howard about sword?

A computer

Mark the START and END timestamps:

2s | 10s

START | END

1) A comic book
2) A sword
3) A toy train
4) A drink
6 TV shows, 3 genres:

- **Sitcom**: *The Big Bang Theory, Friends, How I Met Your Mother*
- **Medical**: *Grey’s anatomy, House M.D.*
- **Crime**: *Castle*

<table>
<thead>
<tr>
<th>Show</th>
<th>Genre</th>
<th>#Sea.</th>
<th>#Epi.</th>
<th>#Clip</th>
<th>#QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBT</td>
<td>sitcom</td>
<td>10</td>
<td>220</td>
<td>4,198</td>
<td>29,384</td>
</tr>
<tr>
<td>Friends</td>
<td>sitcom</td>
<td>10</td>
<td>226</td>
<td>5,337</td>
<td>37,357</td>
</tr>
<tr>
<td>HIMYM</td>
<td>sitcom</td>
<td>5</td>
<td>72</td>
<td>1,512</td>
<td>10,584</td>
</tr>
<tr>
<td>Grey</td>
<td>medical</td>
<td>3</td>
<td>58</td>
<td>1,427</td>
<td>9,989</td>
</tr>
<tr>
<td>House</td>
<td>medical</td>
<td>8</td>
<td>176</td>
<td>4,621</td>
<td>32,345</td>
</tr>
<tr>
<td>Castle</td>
<td>crime</td>
<td>8</td>
<td>173</td>
<td>4,698</td>
<td>32,886</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>44</td>
<td>925</td>
<td>21,793</td>
<td>152,545</td>
</tr>
</tbody>
</table>


(Slides by Jie Lei)
### Different show comes with different words

<table>
<thead>
<tr>
<th>Show</th>
<th>Top unique nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBT</td>
<td>game, mom, laptop, water, store, dinner, book, stair, computer, food, wine, glass, couch, date</td>
</tr>
<tr>
<td>Friends</td>
<td>shop, kiss, hair, sofa, jacket, counter, coffee, everyone, coat, chair, kitchen, baby, apartment</td>
</tr>
<tr>
<td>HIMYM</td>
<td>bar, beer, drink, job, dad, sex, restaurant, wedding, party, booth, dog, story, bottle, club, painting</td>
</tr>
<tr>
<td>Grey</td>
<td>nurse, side, father, hallway, scrub, chart, wife, window, life, family, chief, locker, head, surgery</td>
</tr>
<tr>
<td>House</td>
<td>cane, team, blood, test, brain, pill, office, pain, symptom, diagnosis, hospital, coffee, cancer, drug</td>
</tr>
<tr>
<td>Castle</td>
<td>gun, victim, picture, case, photo, body, murder, suspect, scene, crime, money, interrogation</td>
</tr>
</tbody>
</table>

---

(Slides by Jie Lei)
• Task 1: Question answering without timestamp annotation

• Task 2: Question answering with timestamp annotation

(Slides by Jie Lei)
What is Sheldon holding when he is talking to Howard about sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink

(Howard:) Sheldon, he’s got Raj. Use your sleep spell. Sheldon. Sheldon.

(Sheldon:) I've got the Sword of Azeroth.

Subtitle sentences

Video frames

(Slides by Jie Lei)
ResNet101 → ImageNet feature (img)

Object Detector → Regional visual feature (reg)

Faster R-CNN trained on Visual Genome

Visual concepts feature (cpt)

Faster R-CNN detection example

The regional visual feature (image embeddings inside the bounding boxes) and visual concepts feature (shown in the caption) can be used to answer the question:

‘What is Sheldon holding when everyone is at the door?’ (basket).

brown door, gold sign, red sign, woman, white shorts, green sweater, man, blue shirt, white basket, woman, gray pants, gray door, standing man, gray shirt, black pants

(Slides by Jie Lei)

Anderson, Peter, et al. Bottom-up and top-down attention for image captioning and VQA. CVPR 2018
Multiple streams, each stream deals with different contextual input

(Slides by Jie Lei)
Multiple streams, each stream deals with different contextual input

- Upper stream -- Video

---

(Slides by Jie Lei)
Multiple streams, each stream deals with different contextual input

- Bottom stream -- Subtitle

(Slides by Jie Lei)
<table>
<thead>
<tr>
<th>Method</th>
<th>Video Feature</th>
<th>Test Accuracy</th>
<th>w/o ts</th>
<th>w/ ts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Random</td>
<td>-</td>
<td>20.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>1 Longest Answer</td>
<td>-</td>
<td>30.41</td>
<td>30.41</td>
<td></td>
</tr>
<tr>
<td>2 Retrieval-Glove</td>
<td>-</td>
<td>22.48</td>
<td>22.48</td>
<td></td>
</tr>
<tr>
<td>3 Retrieval-SkipThought</td>
<td>-</td>
<td>24.24</td>
<td>24.24</td>
<td></td>
</tr>
<tr>
<td>4 Retrieval-TFIDF</td>
<td>-</td>
<td>20.88</td>
<td>20.88</td>
<td></td>
</tr>
<tr>
<td>5 NNS-Glove Q</td>
<td>-</td>
<td>22.40</td>
<td>22.40</td>
<td></td>
</tr>
<tr>
<td>6 NNS-SkipThought Q</td>
<td>-</td>
<td>23.79</td>
<td>23.79</td>
<td></td>
</tr>
<tr>
<td>7 NNS-TFIDF Q</td>
<td>-</td>
<td>20.33</td>
<td>20.33</td>
<td></td>
</tr>
<tr>
<td>8 NNS-Glove S</td>
<td>-</td>
<td>23.73</td>
<td>29.66</td>
<td></td>
</tr>
<tr>
<td>9 NNS-SkipThought S</td>
<td>-</td>
<td>26.81</td>
<td>37.87</td>
<td></td>
</tr>
<tr>
<td>10 NNS-TFIDF S</td>
<td>-</td>
<td>49.94</td>
<td>51.23</td>
<td></td>
</tr>
<tr>
<td>11 Our Q</td>
<td>-</td>
<td>43.34</td>
<td>43.34</td>
<td></td>
</tr>
<tr>
<td>12 Our V+Q</td>
<td>img</td>
<td>42.67</td>
<td>43.69</td>
<td></td>
</tr>
<tr>
<td>13 Our V+Q</td>
<td>reg</td>
<td>42.75</td>
<td>44.85</td>
<td></td>
</tr>
<tr>
<td>14 Our V+Q</td>
<td>cpt</td>
<td>43.38</td>
<td>45.41</td>
<td></td>
</tr>
<tr>
<td>15 Our S+Q</td>
<td>-</td>
<td>63.14</td>
<td>66.23</td>
<td></td>
</tr>
<tr>
<td>16 Our S+V+Q</td>
<td>img</td>
<td>63.57</td>
<td>66.97</td>
<td></td>
</tr>
<tr>
<td>17 Our S+V+Q</td>
<td>reg</td>
<td>63.19</td>
<td>67.82</td>
<td></td>
</tr>
<tr>
<td>18 Our S+V+Q</td>
<td>cpt</td>
<td><strong>65.46</strong></td>
<td><strong>68.60</strong></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Question only

Add Video

Add Subtitle

Add Video, Subtitle

Both visual and textual information are important!

(Slides by Jie Lei)
### Accuracy for different methods on TVQA test set

<table>
<thead>
<tr>
<th>Method</th>
<th>Video Feature</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>w/o ts</td>
</tr>
<tr>
<td>Random</td>
<td>-</td>
<td>20.00</td>
</tr>
<tr>
<td>Longest Answer</td>
<td>-</td>
<td>30.41</td>
</tr>
<tr>
<td>Retrieval-Glove</td>
<td>-</td>
<td>22.48</td>
</tr>
<tr>
<td>Retrieval-SkipThought</td>
<td>-</td>
<td>24.24</td>
</tr>
<tr>
<td>Retrieval-TFIDF</td>
<td>-</td>
<td>20.88</td>
</tr>
<tr>
<td>NNS-Glove Q</td>
<td>-</td>
<td>22.40</td>
</tr>
<tr>
<td>NNS-SkipThought Q</td>
<td>-</td>
<td>23.79</td>
</tr>
<tr>
<td>NNS-TFIDF Q</td>
<td>-</td>
<td>20.33</td>
</tr>
<tr>
<td>NNS-Glove S</td>
<td>-</td>
<td>23.73</td>
</tr>
<tr>
<td>NNS-SkipThought S</td>
<td>-</td>
<td>26.81</td>
</tr>
<tr>
<td>NNS-TFIDF S</td>
<td>-</td>
<td>49.94</td>
</tr>
<tr>
<td>Our Q</td>
<td>-</td>
<td>43.34</td>
</tr>
<tr>
<td>Our V+Q</td>
<td>img</td>
<td>42.67</td>
</tr>
<tr>
<td>Our V+Q</td>
<td>reg</td>
<td>42.75</td>
</tr>
<tr>
<td>Our V+Q</td>
<td>cpt</td>
<td>43.38</td>
</tr>
<tr>
<td>Our S+Q</td>
<td>-</td>
<td>63.14</td>
</tr>
<tr>
<td>Our S+V+Q</td>
<td>img</td>
<td>63.57</td>
</tr>
<tr>
<td>Our S+V+Q</td>
<td>reg</td>
<td>63.19</td>
</tr>
<tr>
<td>Our S+V+Q</td>
<td>cpt</td>
<td>65.46</td>
</tr>
</tbody>
</table>

Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Timestamp information is helpful

(Slides by Jie Lei)
TVQA dataset and code:

https://tvqa.cs.unc.edu/
https://github.com/jayleicn/TVQA
TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)
TVR Moment Retrieval Example

Query: Rachel explains to her dad on the phone why she can't marry her fiancé. 
Query Type: video + subtitle
TVR Moment Retrieval Example

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.
Query Type: video + subtitle
TVR Data Analysis

Percentage of queries that have multiple actions or involve multiple people. We also show query examples, with unique person mentions underlined and actions in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#actions ≥2 (%)</th>
<th>#people ≥2 (%)</th>
<th>Query examples (query type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TACoS [28]</td>
<td>20</td>
<td>0</td>
<td>She rinses the peeled carrots off in the sink. (video)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>The person removes roots and outer leaves and rewash the leek. (video)</td>
</tr>
<tr>
<td>CharadesSTA  [8]</td>
<td>6</td>
<td>12</td>
<td>A person is eating food slowly. (video)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A person is opening the door to a bedroom. (video)</td>
</tr>
<tr>
<td>ActivityNet  Caption [21]</td>
<td>44</td>
<td>44</td>
<td>He then grabs a metal mask and positions himself correctly on the floor. (video)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>The same man comes back and lifts the weight over his head again. (video)</td>
</tr>
<tr>
<td>DiDeMo [13]</td>
<td>6</td>
<td>10</td>
<td>A dog shakes its body. (video)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A lady in a cowboy hat claps and jumps excitedly. (video)</td>
</tr>
<tr>
<td>TVR</td>
<td>67</td>
<td>66</td>
<td>Bert leans down and gives Amy a hug who is standing next to Penny. (video)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Taub argues with the patient that fighting in Hockey undermines the sport. (sub)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chandler points at Joey while describing a woman who wants to date him. (video+sub)</td>
</tr>
</tbody>
</table>

- Compared to existing datasets, TVR queries typically have more people and actions and require both video and sub (subtitle) context.
TVR Task

Video Corpus Moment Retrieval (VCMR)
- A query + A video corpus ➔ Retrieve the matched moment from the corpus.
  - Retrieve the GT video. (Video Retrieval)
  - Localize the moment from the retrieved video. (Single Video Moment Retrieval)

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.
Query Type: video + subtitle

(Slides by Jie Lei)
Our Model: Cross-modal Moment Localization (XML)

XML Backbone

Sheldon and Leonard go downstairs side by side.

Leonard: Sounds like a breakthrough

(Slides by Jie Lei)
Our Model: Cross-modal Moment Localization (XML)

Single Video Moment Retrieval
- We first compute query-clip similarity scores $S_{\text{query-clip}} \in \mathbb{R}^l$.
- We then apply Convolutional Start-End (ConvSE) detector:

$$S_{st} = \text{Conv1D}_{st}(S_{\text{query-clip}}), \quad S_{ed} = \text{Conv1D}_{ed}(S_{\text{query-clip}})$$
- The scores are normalized with softmax to output the probabilities $P_{st}, P_{ed} \in \mathbb{R}^l$. (Slides by Jie Lei)
### Baseline comparison on TVR test-public set, VCMR task.

<table>
<thead>
<tr>
<th>Model</th>
<th>w/ video</th>
<th>w/ sub.</th>
<th>IoU=0.5 R@1</th>
<th>IoU=0.5 R@5</th>
<th>IoU=0.5 R@10</th>
<th>IoU=0.5 R@100</th>
<th>IoU=0.7 R@1</th>
<th>IoU=0.7 R@5</th>
<th>IoU=0.7 R@10</th>
<th>IoU=0.7 R@100</th>
<th>Runtime ↓ (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>-</td>
</tr>
<tr>
<td><strong>Proposal based Methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCN</td>
<td>✓</td>
<td>✓</td>
<td>0.02</td>
<td>0.15</td>
<td>0.24</td>
<td>2.20</td>
<td>0.00</td>
<td>0.07</td>
<td>0.09</td>
<td>1.03</td>
<td>-</td>
</tr>
<tr>
<td>CAL</td>
<td>✓</td>
<td>✓</td>
<td>0.09</td>
<td>0.31</td>
<td>0.57</td>
<td>3.42</td>
<td>0.04</td>
<td>0.15</td>
<td>0.26</td>
<td>1.89</td>
<td>-</td>
</tr>
<tr>
<td><strong>Retrieval + Re-ranking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEE+MCN</td>
<td>✓</td>
<td>✓</td>
<td>0.92</td>
<td>3.69</td>
<td>5.58</td>
<td>17.91</td>
<td>0.42</td>
<td>1.89</td>
<td>2.98</td>
<td>10.84</td>
<td>66.8</td>
</tr>
<tr>
<td>MEE+CAL</td>
<td>✓</td>
<td>✓</td>
<td>0.97</td>
<td>3.75</td>
<td>5.80</td>
<td>18.66</td>
<td>0.39</td>
<td>1.69</td>
<td>2.98</td>
<td>11.52</td>
<td>161.5</td>
</tr>
<tr>
<td>MEE+ExCL</td>
<td>✓</td>
<td>✓</td>
<td>0.92</td>
<td>2.53</td>
<td>3.60</td>
<td>6.01</td>
<td>0.33</td>
<td>1.19</td>
<td>1.73</td>
<td>2.87</td>
<td>1307.2</td>
</tr>
<tr>
<td><strong>XML</strong></td>
<td>✓</td>
<td>✓</td>
<td><strong>7.25</strong></td>
<td><strong>16.24</strong></td>
<td><strong>21.65</strong></td>
<td><strong>44.44</strong></td>
<td><strong>3.25</strong></td>
<td><strong>8.71</strong></td>
<td><strong>12.49</strong></td>
<td><strong>29.51</strong></td>
<td><strong>25.5</strong></td>
</tr>
</tbody>
</table>

(Slides by Jie Lei)
Experiments

Performance of XML models that use only video, subtitle, or both as inputs.

- Use both video and subtitle performs the best.

(Slides by Jie Lei)
Experiments

Performance comparison of moment generation methods, under the same XML backbone.

- TAG and SlidingWindow rely on handcrafted rules, while ConvSE learns from data.
- ConvSE performs consistently better across different IoU thresholds.

(Slides by Jie Lei)
Data & Code Release

TV show Retrieval (TVR):

https://tvr.cs.unc.edu/
https://github.com/jayleicn/TVRetrieval

TV show Captions (TVC):

-- We collected additional descriptions for each TVR moment.

https://tvr.cs.unc.edu/tvc.html
https://github.com/jayleicn/TVCaption

(Slides by Jie Lei)
What is More Likely to Happen Next? Video-and-Language Future Event Prediction

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)
VLEP Task & Example

- **Task**: Given a video (with dialogue) as premise, predict what is most likely to happen next by selecting from two provided future events. This task requires using event schema knowledge, which is quite challenging for modern AI systems.

**Premise Event**

00:21,320 --> 00:23,381
[Mark] Oh yeah! Maybe a shake.

*(Premise Summary: A woman with a white shirt with black buttons grinds fruit slush in a blender.)*

**Future Events** (Which event is more likely to happen right after the premise?)

A. The woman in the white shirt pours the slush into a cup.
   
   *(Rationale: Slushy drinks are more commonly served in a cup, but there are hollowed out watermelon rinds sitting around the blender.)*

B. The woman in the white shirt pours the slush into a watermelon rind and passes it to Mark.
   
   *(Rationale: There are hollowed out watermelon rinds sitting around the blender.)*

A VLEP example with a YouTube Vlog video.

(Slides by Jie Lei)
VLEP Dataset Collection


- Adversarial Matching.
  - We sample negatives from existing human positives that is close to the given premise but not overly similar to the true positive.

Adversarial collection procedure.

(Slides by Jie Lei)
VLEP Dataset

- We collected 28.7K examples with 10K TV show and YouTube Vlog video clips from different genres. We also show top unique verbs in each genre.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Genre</th>
<th>#Shows (#Channels)</th>
<th>#Videos</th>
<th>#Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV show</td>
<td>Sitcom</td>
<td>3</td>
<td>4,117</td>
<td>12,248</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>2</td>
<td>1,558</td>
<td>5,198</td>
</tr>
<tr>
<td></td>
<td>Crime</td>
<td>1</td>
<td>1,072</td>
<td>4,306</td>
</tr>
<tr>
<td>YouTube Vlogs</td>
<td>Travel, Food</td>
<td>6</td>
<td>2,406</td>
<td>4,812</td>
</tr>
<tr>
<td></td>
<td>Family, Daily</td>
<td>3</td>
<td>1,081</td>
<td>2,162</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>-</td>
<td>15</td>
<td>10,234</td>
</tr>
</tbody>
</table>

Data statistics by genre.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Top Unique Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitcom</td>
<td>change, offer, hear, should, accept, yell, hang, join, apologize, shut, shout, realize</td>
</tr>
<tr>
<td>Medical</td>
<td>die, treat, cry, yell, smile, proceed, examine, approach, argue, save, admit, rush</td>
</tr>
<tr>
<td>Crime</td>
<td>kill, shoot, point, question, toss, hang, remove, catch, lie, deny, investigate,</td>
</tr>
<tr>
<td>Travel, Food</td>
<td>taste, add, pour, dip, cook, describe, cut, order, serve, stir, prepare, enjoy, buy</td>
</tr>
<tr>
<td>Family, Daily</td>
<td>drive, jump, wear, point, smile, touch, climb, dress, set, swim, hide, lay, blow</td>
</tr>
</tbody>
</table>

Top unique verbs in each genre.

(Slides by Jie Lei)
Method

• A transformer-based method.
  • Video feature: 2D appearance feature + 3D action feature.
  • Text feature: from a RoBERTa model fine-tuned on event schemas from ATOMIC knowledge base.
  • A multimodal transformer encoder for both video and text.

An ATOMIC example. Example sentence: X repels Y’s attack, then X want to file a police report.

Model overview

(Slides by Jie Lei)
Experiments

- We split data into 70% training, 15% development, 15% testing splits.
- Left: Video, dialogue are both useful for the task, when combined, we obtain the best performance of 67.46%, but is still far below human performance of 90.5%.
- Right: event schema knowledge is useful for the task, without ATOMIC sentences for fine-tuning, we see a lower performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chance</td>
<td>50.00</td>
</tr>
<tr>
<td>future only</td>
<td>58.09</td>
</tr>
<tr>
<td>video + future</td>
<td>59.03</td>
</tr>
<tr>
<td>dialogue + future</td>
<td>66.63</td>
</tr>
<tr>
<td>video + dialogue + future</td>
<td>67.46</td>
</tr>
<tr>
<td>human (dialogue + future)</td>
<td>76.25</td>
</tr>
<tr>
<td>human (video + dialogue + future)</td>
<td>90.50</td>
</tr>
</tbody>
</table>

Results on VLEP test splits.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>video + dialogue + future</td>
<td>67.46</td>
</tr>
<tr>
<td>- ATOMIC fine-tuning</td>
<td>66.96</td>
</tr>
</tbody>
</table>

Effect of ATOMIC fine-tuning.

(Slides by Jie Lei)
Data Release

https://github.com/jayleicn/VideoLanguageFuturePred
Thanks!

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)