Exploring Episodic Memory through Cross-modal representations

Abhiraj Mohan
mohan056@umn.edu

Emily Mulhall
mulha024@umn.edu

Jayant Sharma
jayant@umn.edu

Abstract

Large scale egocentric datasets, such as Ego4D (Grauman et al., 2021), and EPIC-KITCHENS (Damen et al., 2021), have paved the way for solving several egocentric tasks. Egocentric data captures a first-person view of the world that cannot be captured by an environmental camera (Lee et al., 2012). We can begin to augment and improve an agent's understanding capabilities through activity recognition and episodic memory (Dodd and Gutierrez, 2005), (Hayes and Shah, 2017). Ego4D and EPIC-KITCHENS both contain thousands of videos of daily tasks paired with annotations, and thus, they allow for cross-modality learning. For our project, we tackle the episodic memory (Tulving, 1972) benchmark task described by the authors of Ego4D through both visual and natural language cues. The majority of current state-of-the-art baseline models are limited to learning from only the video features and input queries (Wang et al., 2019), (Anne Hendricks et al., 2017a). Through leveraging the textural annotations in addition to the visual cues, we were able to surpass the baseline. Our largest challenge was ensuring that the model took advantage of both cues and did not rely only on narration features. We combined these features in a couple different ways, and found that implementing an ensemble method led to the best performance.

1 Motivation

One of the ways in which artificial intelligence may prove fruitful in the future is the development of robotic assistants. Consider the visual sensory stream of such an agent - it would be constantly changing with the robot's movement, in contrast with the static feed of a surveillance camera. Such video is egocentric - captured from first person viewpoint, and it is long, fluid, uncurated. It naturally brings the agent's intentions and interactions with the environment to the forefront. Consequently, egocentric video datasets (Damen et al., 2021), (Lee et al., 2012), (Li et al., 2018), (Pirsiavash and Ramanan, 2012), (Singh et al., 2016) provide a testbed for agents of the future to perceive, learn and recall like humans do (or better).

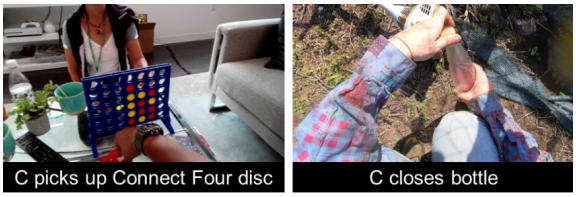

Figure 1: Ego4D Example images and narrations. Figure borrowed from (Grauman et al., 2021)

Given the importance of large datasets in building intelligent systems, Ego4D (Grauman et al., 2021) is a recent attempt by the computer vision community to build a large scale egocentric dataset and a suite of benchmark challenges to catalyze future embodied perception research. The first version of the dataset release comprises 3000+ hours of first person video that is narrated along with benchmark specific annotations for sizeable subsets. The different benchmarks address the past, present and future - search & recall for what has already transpired; identifying hand object interactions, state changes and socially relevant goings-on in the present and forecasting future actions or movement trajectories.

The Episodic Memory benchmark caters to understanding the past - its component tasks search the history of episodes as seen through the first person lens for queries that maybe visual (image), natural language (free form sentence) or category drawn from a taxonomy built from action verbs. Currently implemented baselines operate by correlating the query representation with the video representation and determining the most likely occurrence.

While helpful, these baselines ignore the rich information that narrations carry. Akin to how captions provide useful context for images or figures,

it stands to reason that perception systems would benefit from leveraging annotator descriptions. The narrations are dense at 13.2 sentences per minute of video; an example is show in Figure 1. One of the major challenges here is that the narrations are only weakly aligned. Unlike image frames that come in at a constant frame-per-second (fps), the narrations are sparse and vary according to the action density in the video as well as annotator bias. This is the technical gap that we address in our project - developing cross-modal representations for video and weakly aligned text descriptions to aid in egocentric video understanding.

2 Literature Survey

The neural learning revolution has allowed us to tackle tasks that need visual and language understanding far better than ever before. The main task of this project is that of episodic memory: given a video and a query, can we determine the part of the video required for the answer? There are many tasks related to this one, including visual query-answering, language cross-modality alignment, and span-based question answering.

2.1 Visual question answering

Visual question answering was one of the earliest areas of application for language+vision tasks. The field of VQA modeling is fairly dominated by neural learning techniques that combine image embeddings with query embeddings (Wu et al., 2015). Neural-QA (Malinowski et al., 2015) combined CNN-learned image embeddings with language embeddings from the query using LSTM cells. (Ren et al., 2015) treated the VQA problem as a classification problem and predicted the final answer as one of several possibilities. More recently, attention mechanisms and transformer architectures have been heavily employed in VQA. (Anderson et al., 2018) utilized a top-down attention mechanism for combining the query features as context with visual features that are obtained using a bottom-up attention CNN architecture.

2.2 Video and language cross-modality alignment

The task of video and language cross-modality alignment is as follows: given a video and text description or query, determine the specific part(s) of the video that aligns with the text. There are many different approaches to this task, but most fall into the categories of feature-based, attention-based, or sequential-based. A feature-based method is (Anne Hendricks et al., 2017b), which uses temporal context features to capture both local and global context in the video. Similarly, (Zhang et al., 2019a) uses a tree attention network to extract features from the input query. Attention-based methods like (Xu et al., 2019) and (Liu et al., 2018a), uses attention to focus on relevant temporal locations in the video. Additional methods like (Liu et al., 2018b), (Jiang et al., 2019), and (Yuan et al., 2019) use attention to capture the interaction between the natural language query and the video. On the other hand, we have sequence-based methods like (Zhang et al., 2019b), which use recurrent neural network (RNN) structures to capture the temporal relationship between text and video. Finally, (Zhang et al., 2020b) uses a temporal adjacent network, and is discussed further in our baseline section below.

2.3 Span-based Question Answering

Highly related to the task of this project is span-based question answering, or the challenge of answering questions about a span of text or video. (Wang and Jiang, 2016) combines two pre-existing models, match-LSTM (Wang and Jiang, 2015) and Pointer-net (Vinyals et al., 2015) to correctly determine the answer to the query within the given context passage. Many methods utilize attention, including (Seo et al., 2016), which uses bi-directional attention, (Xiong et al., 2017), which uses co-attention, and (Yu et al., 2018), which uses self-attention. In addition, (Huang et al., 2017) uses multi-level attention to fuse local and global semantic information. More recently, pre-trained models are rising to the forefront. This includes transformer-based models like BERT (Devlin et al., 2019) and its variations such as RoBERTa (Liu et al., 2019) and ViLBERT (Lu et al., 2019). Other transformer-based methods for this task include (Yang et al., 2019), (Sun et al., 2019), (Yu and Jiang, 2019), and (Tan and Bansal, 2019). Lastly, (Zhang et al., 2020a) uses a span-based framework and is discussed further in our baseline section.

2.4 Egocentric video datasets

Egocentric datasets aim to capture first-person videos depicting diverse actions performed on a large scale. These datasets provide the basis for solving several downstream tasks. Datasets such as the UT-Egocentric (Lee et al., 2012) and the Activi-

ties of Daily Life (ADL) (Pirsiavash and Ramanan, 2012) are compiled using a diverse set of actions in different settings. On the other hand, the EPIC-KITCHENS dataset (Damen et al., 2021) is a collection of activities that were all recorded in kitchen settings, annotated by a narrator and a transcriber. The Disney dataset (Fathi et al., 2012) is collected through agents at theme parks and covers social interactions. By pairing first-person and third-person views at the same timestamp, the Charades-Ego (Sigurdsson et al., 2018) dataset allows for joint modeling of the actor and observer perspectives.

3 Problem Formulation

The Ego4D Episodic Memory benchmark is comprised of three related tasks about querying the past as seen in a video. In this project we address the natural language queries (NLQ) task, as defined below. One key aspect of this task is that the model does not need an external knowledge base to provide an answer, and all queries should be answerable from the contents of the video itself. For example, "What is capital of France?" is not a fair query, whereas "What did I do in France just before my flight home?" is.

3.1 Task Definition

The task is defined as, given an egocentric video $\mathcal{V}$ and a natural language query $\mathcal{Q}$, identify the contiguous track of frames $r = \{r_s, r_{s+1} \dots r_e\}$ from which the answer to the query may be deduced. Note that the task is coarser than VQA - it is only sought to temporally localize the query rather than to answer it itself. Our hypothesis is that this task definition naturally lends itself to modular pipelines where the first phase is to localize the query, and the second is to delve into the short localized segment to retrieve the answer.

3.2 Evaluation Metrics

The prescribed metric for evaluation, in keeping with prior work (Zhang et al., 2020c) is **recall@k, IoU=m**. It computes the percentage of times that at least one of the top k prediction windows has an intersection-over-union overlap of at least m. Methods are evaluated at $k = \{1, 3, 5\}$ and $m = \{0.3, 0.5\}$.

3.3 Baselines

Two recent methods from natural language grounding literature - 2D-TAN (Zhang et al., 2020c) and VSLNet (Zhang et al., 2020a) are adapted to the task and benchmarked. The results are provided in 1. Notably, both methods use precomputed SlowFast (Feichtenhofer et al., 2019) video features and language features from a pretrained BERT (Devlin et al., 2019) model for training. We propose to start from these precomputed features as well and only train the cross-modality representation and prediction parts of the model.

Baseline	r@1, IoU=0.3	r@5, IoU=0.3
2D-TAN	5.04	**12.89**
VSLNet	**5.45**	**33.45**

Table 1: Performance of the NLQ baselines

4 Method

We took a deliberately stepped approach to developing our model by first designing simple experiments to test our hypothesis, that introducing textural features from narrations would lead to improved performance. Once that was validated, we developed deep learning models to exploit the useful signals discovered. Our code can be found at https://github.com/emulhall/episodic-memory.

4.1 Terminology

For a query $\mathcal{Q}$ associated with egocentric video $\mathcal{V}$, we denote by $\mathcal{Q}_{text}$, $\mathcal{Q}^s_{time}$ and $\mathcal{Q}^e_{time}$ the natural language text, the starting time and the ending time, respectively. These three components together define query $\mathcal{Q}$. The start and end times delimit the part of the video where the answer to the query may be found.

The set of narrations associated with video $\mathcal{V}$ are denoted by $[\mathcal{N}_k]$ where k is the index into the chronologically sorted array. Each narration $\mathcal{N}_k$ is comprised of the natural language description and a video timestamp for the point in time to which the narration corresponds - $\mathcal{N}_{k,text}$ and $\mathcal{N}_{k,time}$ respectively.

4.2 Proof of Concept

First, we viewed some video data comprising the benchmark task by creating visualizations in which the videos were overlaid with queries and narrations according to their time intervals or timestamps. Figure 2 shows an example.

We observed that narrations highlighted interactions of the camera wearer with their environment,

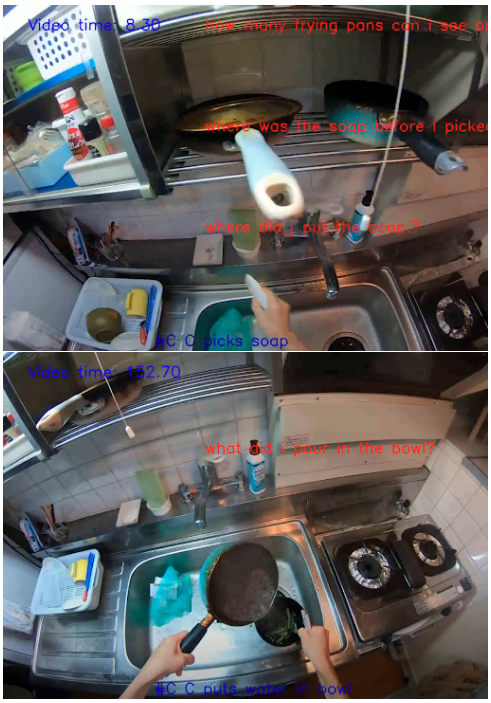

Figure 2: Visualization examples. Queries are in red, and the narrations are the blue text at the bottom of the frame.

and such interactions were also the subject of numerous queries. For example, in the top frame of Figure 2, the query, “Where was the soap before I picked it up?" can likely be answered around the frame in which the narration is, “#CC picks up soap," where #CC is the camera wearer. Likewise, in the bottom frame, the query, "What did I pour in the bowl?" is directly answered by the narration at that frame, “#CC puts water in bowl."

Having observed a few instances, such as these, supporting our hypothesis, we next sought to garner quantitative validation over the full dataset with a simple experiment.

4.2.1 Narrations

Following machine learning tradition, we developed our own version of a nearest neighbor (NN) baseline. There are two major components here - first is the feature space and distance metric, and second is the metric to baseline. We used a large language model (LLM) feature extractor to get sentence features, and the distance metric was cosine distance in normalized feature space.

While the answer to each query is a contiguous set of frames, observe that the narrations only correspond to a point in time. Even if the nearest neighbor retrieved narration be appropriate to the query, it is non-trivial to go from point in time to the time span expected as the correct answer to the query.

We, therefore, construct a simpler accuracy metric instead of the full intersection over union (IOU). We make the above discussion mathematically concrete below,

Given query $\mathcal{Q}$ for video $\mathcal{V}$ with narrations $[\mathcal{N}_k]$,

$$f(x_{text}) = \frac{g(x_{text})}{||g(x_{text})||}$$

$$\mathcal{N}_i = \underset{\mathcal{N}_k}{\text{argmax}}\, f(\mathcal{N}_{k,\,text})^T f(\mathcal{Q}_{text})$$

$$acc = \begin{cases} 1 & \mathcal{Q}^s_{time} \leq \mathcal{N}_{i,time} \leq \mathcal{Q}^e_{time} \\ 0 & \text{otherwise} \end{cases}$$

where g is a pretrained LLM feature extractor.

Intuitively, the nearest neighbor narration should correspond to a frame that falls within the start and end times of the query. We run top-k NN experiments with k = 1,3,5,10 and compare with random retrieval.

4.2.2 Image Captions

In addition, we explore the use of auto generated image captions as a complementary source of information to narrations. We observe that the narrations describe salient interactions of the camera wearer with the environment but fail to mention other objects that may be of interest. For e.g.: a pan on a kitchen shelf above the sink won’t be mentioned in the narrations if the camera wearer is only washing dishes in the observed video, however, it is important from an episodic memory standpoint since it might come in handy later. To this end, we explore if image captioning networks can provide useful scene descriptions that may be leveraged for the NLQ prediction task.

Image captioning networks have come a long way at describing the foreground and background of an image using natural language in recent years (Hossain et al., 2019). We visualize some sample egocentric images from the dataset and captions generated by OpenAI’s CLIP model (Contrastive Language-Image Pre-Training) (Mokady et al., 2021) in Figure 3. While the model performs well on non-egocentric data, it clearly struggles with egocentric images. The model appears to be quite sensitive to the domain gap. We leave the exploration of dense captioning methods to the future.

In spite of our skepticism, we tested the utility of image captions through a setup identical to the one for narrations. We captioned videos at a constant frame rate and these were used as an input to the nearest neighbor pipeline.

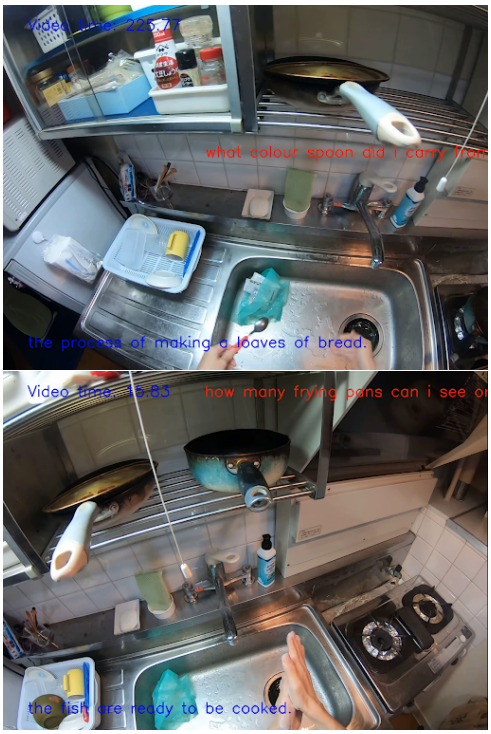

Figure 3: Captioning visualization examples. Queries are in red, and the captions are the blue text at the bottom of the frame.

4.3 Full Model

4.3.1 Representation

The input to the baseline model, VSLNet, is a feature representation of the video $\mathcal{V}^f$ with shape (N, K) where the first dimension indexes video location, e.g., index 0 is video start, $N-1$ is end, $\lfloor N/2 \rfloor$ is middle, etc. The second dimension is the feature dimension. Analogous to how word embeddings in recent language models embody context, the frame-wise video features also encode the temporal context found in the video.

One method to encode narrations could be to map the per narration features extracted by a language model onto the frame indices corresponding to their timestamp. While logical, this would result in a sparse and uneven feature distribution since the majority of frames do not have a corresponding narration.

Researchers working on 2D Pose Detection (Wei et al., 2016) discovered that converting a delta function into a continuous function using Gaussian filtering was conducive to neural network learning. We take inspiration from this body of literature as well as observe that the underlying action being described by video narrations must have been performed over a neighboring time span. We process narration features by filtering them using a radial basis kernel with bandwidth equal to a few seconds before pooling the features over the video. Concretely, for narration $\mathcal{N}_k$, we construct feature $\mathcal{N}_k^f \in \mathbb{R}^{N \times L}$ s.t.

$$\mathcal{N}_k^f[i] = \begin{cases} i = h(\mathcal{N}_{k,time}) & g(\mathcal{N}_{k,text}) \\ 0 & \text{otherwise} \end{cases}$$

where $h(.)$ is a function mapping timestamps to frame indices and g is an LLM feature extractor. We then smooth and consolidate all narration features for video $\mathcal{V}$:

$$\mathcal{N}^f = \sum_k \mathcal{N}_k^f * W_\sigma$$

where $*$ is the cross-correlation operator and W_σ a 1D Gaussian filter with bandwidth σ and width $\sim 6\sigma$.

4.3.2 Query Prediction

We attempted two different forms of combining video and narration features for final prediction. First, we used simple concatenation of video and narration embeddings. The rest of the model remained unchanged. We also try an ensembling approach. Here, we trained independent models with video only as well as narration only features with the full objective as described in the original paper.

We then train new start and end time prediction heads, the input to which are the concatenated features from the respective heads of the two original models. The models trained in the previous stage are now frozen and the prediction heads are the only component trained with a prediction loss; the query highlighting loss is ignored in this stage.

5 Experiments

5.1 Proof of Concept

We first discuss implementation details and results of our Proof of Concept experiments. To extract narration features, we used BERT (Devlin et al., 2018) as the LLM feature extractor.

kNN Retrieval	Accuracy(%)			
	r@1	r@3	r@5	r@10
Random	1	3	4	8
Image Captioning	2	4	6	9
Narration	6	11	15	20

Table 2: Proof of Concept

Method	IoU = 0.3(%)		IoU = 0.5(%)		mIoU
	r@1	r@5	r@1	r@5	
Video only	4.57	9.03	2.50	6.12	3.55
Narration only	6.97	13.58	3.41	8.26	5.12
Concat	6.56	13.58	3.41	8.26	5.12
MLP	4.96	10.33	2.45	5.91	3.78
Ensemble (Full Model)	**8.29**	**15.31**	**4.85**	**9.94**	**6.08**

Table 3: Model Performance

For image captioning, we used OpenAI's CLIP (Mokady et al., 2021) and captioned videos to roughly match the frequency of narrations - once every 5 seconds, or 0.2 fps. Even though only clips or sections of videos are part of the data subset that forms the benchmark, we nevertheless narrated the entirety of videos resulting in just over 300 hours of total video captioned. Once captioned, features were again extracted using BERT (Devlin et al., 2018).

Since kNN is an unsupervised learning method, we used the large train split. The train split comprises 754 videos and 11293 total queries with an average of 14.97 queries per video. The results are presented alongside a random retrieval baseline in Table 2.

We observe that the network generated image captions seem to provide only a marginal lift compared to random retrieval while the human annotated narrations clearly contain a useful signal. We build on the insight generated by this simple experiment, and incorporate the narrations into the VSLNet baseline.

5.2 Model Performance

Since our representation for the narrations is modeled after the video representation and aligned with it, we try a variety of experiments to ablate and combine the two representations. We tried the following different methods:

1. Video only - The model input is SlowFast video features (Feichtenhofer et al., 2019) only; this is the baseline released as part of the benchmark.

2. Narration only - We substitute the video features with our consolidated narration feature as described in Section 4.3.1. The kernel bandwidth used was ~ 1 second.

3. Concat - We concatenate the video and narration features along the feature dimension and use it as the input to the model.

4. MLP - We process the video and narration streams using three successive Conv1D operations with width=1 before concatenating along the feature dimension. We use ReLU non-linearity along with Batch Normalization (Ioffe and Szegedy, 2015) at every layer.

5. Ensemble (Full Model) - Our Full Model corresponding to a late feature fusion strategy as describe in Section 4.3.2.

The models were trained on the train split and tested on the validation split; labels for the test split are hidden since an active challenge on the benchmark is ongoing. The results are presented in Table 3.

The narrations are surprisingly effective at the prediction task, beating the video representation by a comfortable margin. We observed that combining the video and narration representations was far from straightforward. While concatenation gave results comparable to when only narrations were used, preprocessing the features using Conv1D operations fared suprisingly worse. We analyzed the

output of the Narration only and Concat models, but found no evidence to indicate that the model had collapsed to learning only from narrations; the output was equally correlated with the output from Video and Narration only models.

Finally, the ensembling strategy works well and converges quickly to deliver better results than if either one of the representations were used alone.

6 Analysis

6.1 Limitations

Throughout this project, we experienced a few limitations. First, the benchmark task is brand new, having been released only this month. Thus, besides the original baselines presented in (Grauman et al., 2021), there are no other methods to compare against. Second, the benchmark task does not provide the narrations for the test set. This is likely to prevent competitors from manually matching narrations to queries. However, this prevents us from being able to evaluate our method on an unseen test set. Lastly, any update to the features takes place in the data preparation module of the model. This phase generates the features and organizes them for input to the model in the training process, and can take quite a long time to complete. Thus, experimenting with features is a timely process, because each little change needs to re-run the entire data preparation script in addition to then running the training code.

6.2 Ethical Concerns

A few ethical concerns arose during this project. While Ego4D (Grauman et al., 2021) was collected by nearly 1,000 subjects over 9 different countries, it is not necessarily balanced by demographic. Thus, there is a possibility that the data is racially and/or gender biased. In addition, narrations were not necessarily annotated in the same counties as the data was originally collected. Thus, narrators could miss cultural nuances present in the video. Lastly, tasks performed during data collection are not necessarily balanced, which could lead to the model performing far better on certain tasks than others.

7 Conclusion

In this project, we implemented VSLNet (Zhang et al., 2020a) for the task of natural language episodic memory, in which the input is a video and natural language query, and the output is a start and end time within the video in which the answer to the query can be found. We tested our hypothesis, that introducing natural language features in addition to the visual features can lead to stronger performance, through using video narrations and image captions.

To validate our hypothesis, we first performed proof-of-concept experiments using a nearest neighbor method to match narration and caption embeddings to query embeddings. We found that, while captions did not introduce significant helpful information, the narrations did. Using this knowledge, we used the narration features as input to the VSLNet (Zhang et al., 2020a) model. We found the use of these narrations so beneficial to the task that using only narration features led to a better performance than the original video only model. We combined the visual and natural language features a couple of different ways, with varying success. Concatenating the features led to a performance similar to the narration only method, processing the video and narration streams using three successive Conv1D operations before concatenating along the feature dimension which led to a performance close to the original video only method, and finally, introducing an ensemble method, leading to the best performance by far. Thus, we proved our hypothesis correct, that natural language cues in addition to visual cues lead to the best performance on the task of natural language episodic memory.

As mentioned in the introduction, this work is a step towards machines perceiving and understanding scenes through visual and natural language cues. This perception is necessary for future projects, such as the development of robotic assistants. Given an understanding of a scene, we can then introduce instructions for robots to assist in the scene, train them in making useful suggestions to a user, and more. While we don't expect that performances on this task will immediately lead to the production of such assistants, the prospect of training models on in-the-wild video that is cheaply obtained along with expensive, albeit sparse expert annotations is very exciting.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 6077–6086.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017a. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017b. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2021. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Will Dodd and Ridelto Gutierrez. 2005. The role of episodic memory and emotion in a cognitive robot. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 692–697. IEEE.

Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Bradley Hayes and Julie A Shah. 2017. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6586–6593. IEEE.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pages 217–225.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE.

Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635.

Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018a. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568.

Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018b. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Gunnar A. Sigurdsson, Abhinav Kumar Gupta, Cordelia Schmid, Ali Farhadi, and Alahari Karteek. 2018. Actor and observer: Joint modeling of first and third-person videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7404.

Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. 2016. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Endel Tulving. 1972. 12. episodic and semantic memory. *Organization of memory/Eds E. Tulving, W. Donaldson, NY: Academic Press*, pages 381–403.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. 2015. Ask me anything: Free-form visual question answering based on knowledge from external sources. *CoRR*, abs/1511.06973.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020c. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI*.

Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019a. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1230–1238.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664.