Zero-Shot Video Captioning with Topic-Aware Mixture of Experts

Xin Wang
University of California, Santa Barbara
xwang@cs.ucsb.edu

William Yang Wang
University of California, Santa Barbara
william@cs.ucsb.edu

1 Introduction

Video captioning aims at automatically describing the content of a video in natural language and has attracted increasing attention in recent years in both NLP [34] and computer vision communities [18]. Although existing video captioning methods have achieved promising results, they largely rely on paired videos and textual descriptions for supervision [40] and thus cannot generalize well to novel activities that have never been seen before. However, it is prohibitively expensive to collect paired training data for every possible activity. Therefore, we introduce a new task of *zero-shot video captioning*, where a model is required to accurately describe novel activities in videos without any explicit paired training data.

Unlike zero-shot activity recognition that predicts the category of an unseen activity, zero-shot video captioning focuses on the language generation part—learning to describe out-of-domain videos of a novel activity without paired captions. An example of zero-shot video captioning is shown in Figure 1, where an existing method fails to correctly caption a video about the novel activity "*sharpening knives*" because it has learned no knowledge about the activity in training. Moreover, videos of different activities usually require different captioning strategies in various aspects, *i.e.* word selection, semantic construction, style expression etc, which poses a great challenge in the open vocabulary scenario. Despite the difference, some activities share similar characteristics, *e.g.*, *playing baseball* and *playing football* are both sports activities and a few words can be used to describe both in common.

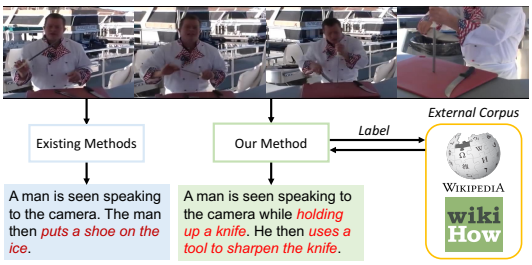
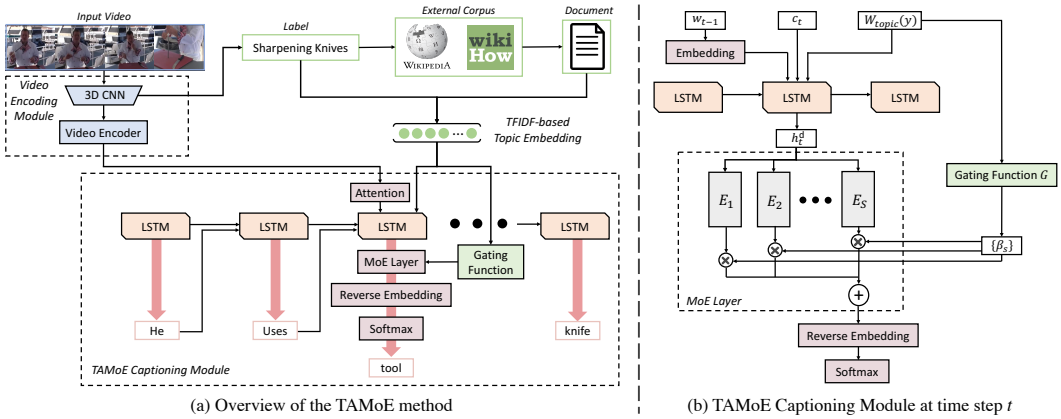


Figure 1: Example of zero-shot video captioning.

Therefore, we propose a novel Topic-Aware Mixture of Experts (TAMoE) approach to caption videos of unseen activities. First, we define a set of *primitive experts* that are sharable by all possible activities, each of which has their own parameters and learns a specialized mapping from latent features to the output vocabulary (the primitive captioning strategies). Then we introduce a *topic-aware gating function* that learns to decide the utilization of those primitive experts and compose a topic-specific captioning model based on a certain topic. Besides, in order to leverage world knowledge from external corpora, we derive a *topic embedding* for each activity from the pretrained semantic embeddings of the most relevant words. When captioning a novel activity, our TAMoE method is capable of inferring the composition of the primitive experts conditioned on the topic embedding, transferring the knowledge learned from seen activities to unseen ones. Empirical results show that our framework generalizes better to unseen activities and can produce more pertinent captions for novel activities compared with the state-of-the-art methods. [1]

[1] This paper is in submission to AAAI 2019.



(a) Overview of the TAMoE method

(b) TAMoE Captioning Module at time step t

Figure 2: (a) Overview of our TAMoE method; (b) The detailed version of the TAMoE caption module.

2 TAMoE Approach for Zero-Shot Video Captioning

We show in Figure 2(a) the overall framework of our Topic-Aware Mixture of Experts (TAMoE) approach, which mainly consists of the *video encoding module*, the *TFIDF-based topic embedding*, and the *TAMoE transfer learning module*. The *video encoding module* encodes video-level features and predicts the activity label. Then, the topic-related documents can be fetched from the external corpus and used to calculate the *TFIDF-based topic embedding*, which represents the semantic meaning of the activity. In the decoding stage, the *TAMoE captioning module* takes both the video features and the topic embedding as input and generates the caption by dynamically composing specialized experts. Below we discuss each module in details.

2.1 Video Encoding Module

Given the input video $\nu = \{v_1, v_2, \ldots, v_n\}$, we employ the pretrained 3D convolutional neural networks to predict the activity label y and extract the segment-level features $\{f_j\}$ where $j = 1, 2, \ldots, m \ll n$ (we use I3D features in our experiments[2]). The I3D features include short-range temporal dynamics while keeping advanced spatial representations. Then our model sends the segment-level features $\{f_j\}$ to the video encoder, which is a bidirectional LSTM, to model long-range temporal contexts. It outputs the hidden representations $\{h_j^e\}$ with e denoting the video encoder, which encodes the video-level features.

2.2 TFIDF-based Topic Embedding

To learn the knowledge of the activities without paired captions, we fetch topic-related documents from various data sources, e.g., Wikipedia and WikiHow. We also employ the pretrained *fasttext* embeddings [21] to calculate the representations of the topics. Given the predicted label y ($y \in Y$) and the related documents D_y, we need to compute the topic-specific knowledge representations.

Term Frequency-Inverse Document Frequency (TF-IDF) is an efficient statistical method to reflect the importance of a word to a document. Here we propose a topic-aware TF-IDF weighting $g_k(y)$ to calculate the relevance of each unigram x_k to the topic-related documents D_y instead of a single document:

$$g_k(y) = \frac{z_k(y)}{\sum_{x_l \in D_y} z_l(y)} \log\left(\frac{|Y|}{\sum_{y' \in Y} \min(1, z_k(y'))}\right) \quad (1)$$

where $z_k(y)$ is the number of times the unigram x_k occurs in the documents D_y related to label y. The first term is the term frequency of the unigram x_k, which places a higher weight on words that frequently occur in the topic-related documents D_y. The second term measures the rarity of x_k with inverse document frequency, reducing the weight if x_k commonly exists across all the topics. Then our TF-IDF embedding $W_{tfidf}(y) = \sum_{x_k \in D_y} g_k(y) W_{fasttext}(x_k)$, where $W_{fasttext}$ denotes the

[2]I3D [7] is the state-of-the-art 3D CNN model for video classification.

pretrained fasttext embeddings. Eventually, the topic embedding $W_{topic}(y)$ is a concatenation of the TF-IDF embedding and the average embedding of the activity label.

2.3 TAMoE Captioning Module

Attention-based Decoder LSTM The backbone of the captioning model is an attention-based LSTM. At each time step t in the decoding stage, the decoder LSTM produces its output h_t^d (d denoting the decoder) by considering the word at previous step w_{t-1}, the visual context vector c_t, the topic embedding $W_{topic}(y)$ and its internal hidden state h_{t-1}^d. In formula,

$$h_t^d = LSTM([w_{t-1}, c_t, W_{topic}(y)], h_{t-1}^d), \quad \text{where} \quad c_t = \sum \alpha_{t,j} h_j^e \tag{2}$$

The context vector c_t is a weighted sum of the encoded video features $\{h_j^e\}$, whose weights $\{\alpha_t^j\}$ are learned by the attention mechanism proposed in [4].

Mixture-of-Expert Layer and Topic-Aware Gating Function Following Equation 2, the output of the decoder LSTM h_t^d is then fed into the Mixture-of-Experts (MoE) layer (see Figure 2(b)). Here each expert is an underlying mapping function from the latent representation h_t^d to the vocabulary, which learns the captioning primitives that are shareable to all topics. All the experts in the same MoE layer have the same architecture, which is parameterized by a fully-connected layer and a nonlinear ReLU activation. Let S denote the number of experts and E_s be the s-th expert, then output of the MoE layer is

$$o_t = \sum_{s=1}^{S} \beta_s E_s(h_t^d), \quad \text{where} \quad \beta_s = \frac{\exp(G(W_{topic}(y))_s/\tau)}{\sum_{i=1}^{S} \exp(G(W_{topic}(y))_i/\tau)} \tag{3}$$

β_s is the gating weight of the expert E_s, representing the utilization of the expert E_s. And it is determined by the topic-aware gating function G, which is a multilayer perceptron in our model. The temperature τ determines the diversity of the gating weights. Lower temperatures encourage sparser utilization and thus more specialized expert learning. The topic-aware gating function G is conditioned on the topic embedding $W_{topic}(y)$ and learns to combine the expertise of those primitive experts for a certain topic. Intuitively, G learns topic-aware language dynamics and composes different expert utilization for different topics based on the topic embeddings, which can implicitly transfer the utilization across topics.

Embedding and Reverse Embedding Layers In addition, we also employ semantic word embeddings in our captioning model to help generate descriptions of novel activities. Incorporating pretrained embeddings assigns semantic meanings to those out-of-domain words and thus can facilitate the open vocabulary learning [32]. Particularly, we load the fasttext embeddings into both the embedding layer and the reverse embedding layer (see Figure 2(b)), and freeze their weights during training. So the embedding layer represents the input word (one-hot vector) into semantically meaningful dense vectors, while the reverse embedding layer is placed before the softmax layer to reverse the mapping from the feature vectors into the vocabulary space.

2.4 Learning

Cross Entropy Loss We adopt the cross entropy loss to train our modes. Let θ denote the model parameters and $w_{1:T}^*$ be the ground-truth word sequence, then the training loss is defined as

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p(w_t^* | w_{1:t-1}^*, \theta) \tag{4}$$

where $p(w_t | w_{1:t-1}, \theta)$ is the probability distribution of the next word.

Variational Dropout In order to regularize our MoE layer, we adopt the variational dropout [12, 20] when training the TAMoE module. Different from the standard dropout, variational dropout samples a binary dropout mask only once upon the first call and then repeatedly uses that locked dropout mask within samples. In addition, variational dropout helps stabilize the training of the topic-aware gating mechanism by making the expert behaviors consistent within samples.

		Seen Test Set						Unseen Test Set					
Model	Embedding	CIDEr	B-1	B-2	B-3	M	R	CIDEr	B-1	B-2	B-3	M	R
Base	task-specific	29.67	23.57	12.06	7.02	9.77	21.45	21.59	22.34	10.57	5.76	9.01	20.06
Base	fasttext	31.48	23.88	12.20	7.11	10.16	21.69	22.51	22.50	11.01	6.02	9.43	20.70
Topic	task-specific	33.06	24.48	12.64	7.27	10.49	22.24	23.06	22.06	10.34	6.05	9.40	20.60
Topic	fasttext	33.72	24.53	12.56	7.20	10.24	22.11	24.06	22.97	11.09	5.98	9.70	20.98
TAMoE	task-specific	34.38	**25.79**	13.29	**7.44**	10.69	**23.03**	24.39	23.36	**11.19**	6.05	9.28	**21.46**
TAMoE	fasttext	**35.53**	25.51	**13.93**	7.39	**10.83**	22.51	**28.23**	**24.34**	11.18	**6.14**	**9.96**	21.17

Table 1: Comparison with the baselines on the held-out ActivityNet-Captions dataset. We report the results in terms of CIDEr, BLEU (B), METEOR (M), and ROUGE-L (R) scores. For each model, we test the impact of pretrained word embeddings by comparing two word embedding initialization strategies: (1) *task-specific*, that randomly initializes the embeddings and learns them during training, and (2) *fasttext*, that uses pretrained fasttext embeddings (fixed in training).

3 Experiments

Dataset We set up the zero-shot learning scenario based on the ActivityNet-Captions dataset [18]. We re-split the videos of the total 200 activities into the *training set* (170 activities), the *validation set* (15 activities), and the *unseen test set* (15 activities). Each activity is unique and only exists in one split above. In order to compare with the model's performance on the supervised split, we then further split an additional *seen test set* that shares the same activities with the training set but has different video samples. More details are provided in the supplementary material.

Evaluation Metrics We use four popular and diverse metrics for language generation, CIDEr, BLEU, METEOR, and ROUGE-L, using the evaluation code provided by [18]. Among these metrics, only CIDEr weighs the topic relevance of n-grams and thus can better reflect a model's capability on captioning novel activities. Therefore, we use CIDEr as the major metric.

Baselines We compare three models on the Held-out ActivityNet-Captions dataset. (1) **Base**: we first implement the state-of-the-art attention-based sequence-to-sequence model used in [24, 36] as our baseline. Simply put, the Base model is an attention-based encoder-decoder model without topic embeddings and the topic-aware gating function. (2) **Topic**: the Topic model has a very similar architecture with the Base model, except that its decoder takes the proposed topic embedding as an additional input. (3) **TAMoE**: the proposed TAMoE model is illustrated in Figure 2, which consists of the video encoding module, the topic embedding, the topic-ware gating function, and the Mixture-of-Experts layer.

Results on Seen and Unseen Test Sets Table 1 shows the results on both seen and unseen test sets. It can be noted that incorporating pretrained fasttext embeddings brings a consistent improvement across models on both test sets, especially for the zero-shot learning scenario on the unseen test set. Besides, solely adding the proposed topic embedding can bring some improvement. These validate the hypothesis that the pretrained embeddings can bring useful prior knowledge to assist caption generation and facilitate the generation of out-of-domain words that do not appear during training. More importantly, our TAMoE model significantly improves the scores over the baseline models. For instance, our full TAMoE model outperforms the Base model on both the seen and the unseen test sets with respectively 19.75% and 30.75% relative improvement on CIDEr. The remarkable improvement on the unseen test set clearly demonstrates the superior capability of the proposed model on captioning novel activities. More experiments can be found in the supplementary material.

4 Conclusion

In this paper, we formally define the task of zero-shot video captioning and set up a common setting for evaluation. In order to accurately describe videos of novel activities, we seek solutions based on what and how to utilize and transfer. Thus we incorporate the topic embedding mined from the external corpus, as well as propose a topic-ware mixture of experts framework to effectively utilize the external knowledge and learn to compose different utilization of the experts for different activities.

References

[1] K. Ahmed, M. H. Baig, and L. Torresani. Network of experts for large-scale image categorization. In *ECCV*, 2016.

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.

[3] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell, J. Mao, J. Huang, A. Toshev, O. Camburu, et al. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[5] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *CVPR*, 2017.

[6] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

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

[8] S. Chen, J. Chen, Q. Jin, and A. Hauptmann. Video captioning with guidance of multimodal latent topics. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1838–1846. ACM, 2017.

[9] R. Collobert, S. Bengio, and Y. Bengio. A parallel mixture of svms for very large scale problems. In *NIPS*, 2002.

[10] R. Collobert, Y. Bengio, and S. Bengio. Scaling large learning problems with hard parallel mixtures. *International Journal of pattern recognition and artificial intelligence*, 17(03):349–365, 2003.

[11] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[12] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *NIPS*, 2016.

[13] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.

[14] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[16] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

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

[18] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[19] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018.

[20] S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing lstm language models. In *ICLR*, 2018.

[21] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *LREC*, 2018.

[22] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, 2016.

[23] R. Pasunuru and M. Bansal. Multi-task video captioning with video and entailment generation. In *ACL*, 2017.

[24] R. Pasunuru and M. Bansal. Reinforced video captioning with entailment rewards. In *EMNLP*, 2017.

[25] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition (GCPR)*, September 2014. Oral.

[26] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

[27] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue. Weakly supervised dense video captioning. In *CVPR*, 2017.

[28] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. T. Shen. Hierarchical lstm with adjusted temporal attention for video captioning. In *IJCAI*, 2017.

[29] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

[30] V. Tresp. Mixtures of gaussian processes. In *NIPS*, 2001.

[31] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. In *EMNLP*, 2016.

[32] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

[33] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *ICCV*, 2015.

[34] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*, 2015.
[35] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
[36] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, 2018.
[37] X. Wang, Y.-F. Wang, and W. Y. Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *NAACL HLT*, 2018.
[38] X. Wang, F. Yu, R. Wang, Y.-A. Ma, A. Mirhoseini, T. Darrell, and J. E. Gonzalez. Deep mixture of experts via shallow embedding. *arXiv preprint arXiv:1806.01531*, 2018.
[39] Y. Wu, L. Zhu, L. Jiang, and Y. Yang. Decoupled novel object captioner. In *ACM MM*, 2018.
[40] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.
[41] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *ICLR*, 2018.
[42] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.
[43] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016.
[44] Y. Yu, H. Ko, J. Choi, and G. Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, 2017.
[45] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Supplementary Material

A Held-out ActivityNet-Captions Dataset

ActivityNet [11] is a well-known benchmark for video classification and detection, which covers 200 classes of activities. Recently, [18] have collected the corresponding natural language description for the videos in the ActivityNet dataset, leading to the ActivityNet-Captions dataset. We set up the zero-shot learning scenario based on the ActivityNet-Captions dataset. We re-split the videos of the 200 activities into the the *training set* (170 activities), the *validation set* (15 activities), and the *unseen test set* (15 activities). Each activity is unique and only exists in one split above. We hold out the novel 15 activities for testing that appear during neither training nor validation. In order to compare with the model's performance on the supervised split, we then further split an additional *seen test set* that shares the same activities with the training set but has different video samples. The external text corpus is crawled from Wikipedia, WikiHow, and some related documents in the first Google Search page. On average there are 2.72 related documents per activity (the max is 10).

B Implementation Details

To preprocess the videos, we sample each video at $20fps$ and extract the I3D features [7] from these sampled frames. Note that the I3D model is pre-trained on the Kinects dataset [17] and used here without fine-tuning. The activity labels feeding to our model are predicted by a pretrained 3D CNN model [35] for activity classification. The vocabulary is built based on the training corpus and the unpaired external corpus. We use 300-dimensional pretrained fasttext embedding for words. All the hyper-parameters are tuned on the validation set. The maximum number of video features is 200 and the maximum caption length is 32. The video encoder is a biLSTM of size 512, and the decoder LSTM is of size 1024. We initialize all the parameters from a uniform distribution on $[-0.1, 0.1]$. Adadelta optimizer [45] is used with batch size 64. Learning rate starts at 1 and is then halved when the current CIDEr score does not surpass the previous best in 4 epochs. The maximum number of epochs is 100, and we shuffle the training data at each epoch. Schedule sampling [6] is also employed to train the models. Beam search of size 5 is used at test time. It takes around 6 hours to fully train a model on a TITAN X. The inference time is about 2.9s, including data and model loading time (2.6s).

C Ablation Study

C.1 Evaluation on Different N-grams

In order to take a closer look at the transfer influence of our TAMoE model on individual n-grams, we calculate the CIDEr score of unigrams, bigrams, trigrams, and fourgrams on the unseen test set separately. As seen in Table 2, our TAMoE model performs the best on all n-grams, but the CIDEr score of 4-grams is still not very satisfactory. A general limitation of current captioning systems is that the focus is still on learning word-level embeddings and generating a caption word by word. Incorporating phrase-level embeddings may alleviate this issue. We leave it for future study.

Model	Embedding	C-1	C-2	C-3	C-4
Base	task-specific	52.13	20.41	8.92	4.18
Base	fasttext	55.17	21.40	8.81	4.64
Topic	task-specific	55.79	20.94	9.47	4.68
Topic	fasttext	58.84	23.33	9.32	4.75
TAMoE	task-specific	58.81	22.98	10.42	6.00
TAMoE	fasttext	**67.48**	**25.89**	**12.09**	**7.47**

Table 2: Individual CIDEr scores of unigrams (C-1), bigrams (C-2), trigrams (C-3), and fourgrams (C-4) on the unseen test set, which are all novel activities.

C.2 Impact of Different Features

In Table 3, we test the influence of the I3D video features and various versions of the topic embedding. Evidently, it performs the best to use the concatenation of the average label embedding and the TFIDF embedding from external corpus as the topic embedding. Besides, without videos features, the model is unable to generate diverse captions for different videos that also match the video content (the corresponding CIDEr score is as low as 15.77).

I3D Video Features	✓	✓	✓		✓
Average Label Embedding		✓		✓	✓
TFIDF Embedding			✓	✓	✓
CIDEr	22.51	25.96	26.61	15.77	28.23

Table 3: Impact of different features on the TAMoE model. *I3D Video Features* are the extracted video features using the pretrained I3D model; *Average Label Embedding* is the average embedding of the words in the predicted activity label; *TFIDF Embedding* is the weighted embedding of the external topic-related documents (see **??**).

C.3 Impact of The Number of Experts

An important hyper-parameter in our TAMoE model is the number of experts in the Mixture-of-Experts layer. We compare models with different numbers of experts. For a fair comparison, we adjust the dimensionality of each expert to ensure that different models have the same capacity (number of parameters). Note that we set the minimum expert dimensionality as 128 to ensure a lower bound of each expert's capacity. Their learning curves on the validation set are shown in Figure 3. As can be observed, the model with 8 experts of dimension 256 (*n8_d256*) works the best, and the single-expert model, which is indeed the Topic model, performs the worst. Besides, simply increasing the number of experts does not imply a gain in performance. For example, the performance of the model *n256_d128* (∼27.2M parameters) is worse than the best-performing model *n8_d256* (∼17.9M parameters).

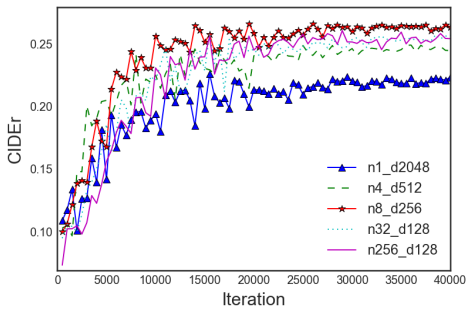


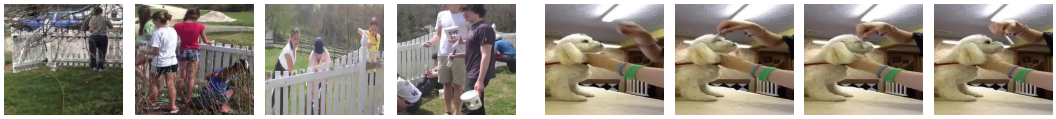
Figure 3: Learning curves of the TAMoE models with different numbers of experts (n) and different expert dimension (d). For example, *n4_d512* denotes the TAMoE model with 4 experts, each of dimension 512. Note the validation scores are calculated by greedy decoding, which are lower than than testing scores by beam search of size 5.

C.4 Topic-wise Result Comparison

To examine the performance of our method on each novel activity, we report the topic-wise comparison with the Base model in Table 4. The TAMoE model outperforms the Base model on most of the activities (12 out of 15), of which some activities are improved by a remarkable margin, e.g., *arm wresting*, *braiding hair*, *gargling mouthwash*, and *sharpening knives*. Meanwhile, we showcase the top-6 related words from the external corpus for each topic according to their TF-IDF weights to provide a better illustration of our topic embeddings.

Novel Activity	Base	TAMoE	Top-6 related words
making a lemonnade	28.63	31.66	lemonade, sugar, lemon, juice, lemons, pitcher
arm wrestling	23.72	35.96	wrestling, arm, opponent, strength, wrist, kindt
longboarding	20.51	28.79	longboard, board, foot, riding, longboarding, goofy
playing badminton	20.18	22.00	shuttle, racket, shuttlecock, court, backhand, serve
shuffleboard	14.95	20.85	shuffleboard, disks, discs, puck, pucks, scoring
slacklining	24.43	21.33	slackline, slacklining, line, balance, walking, balancing
hula hoop	17.50	26.29	hoop, hula, hoops, waist, hooping, pulse
playing drums	31.70	39.44	drum, snare, metronome, hat, hi, drums
braiding hair	21.30	36.80	braid, hair, section, strands, braids, braiding
gargling mouthwash	11.09	52.03	mouthwash, mouth, gargling, fluoride, swish, liquid
installing carpet	22.40	17.85	carpet, strips, tackless, wall, kicker, install
sharpening knives	24.77	43.63	stone, knife, sharpening, blade, sharpen, knives
grooming dog	18.33	26.61	dog, clippers, shampoo, fur, hair, grooming
assembling bicycle	22.17	28.74	handlebar, bike, stem, seat, locate, fork
painting fence	23.56	23.15	fence, paint, painting, sprayer, primer, wood

Table 4: Topic-wise comparison. We compare the CIDEr scores of the Base model and our TAMoE model within each activity. In the right-most column, we list the top words based on their TF-IDF weights in the external topic-related documents.



Base: the man continues to speak to the camera and leads into several shots of people
Ours: a group of people are seen ***standing around a fence*** and speaking to one another.

Base: a woman is seen sitting on a couch with a dog in a bathtub.
Ours: the lady ***grooms the dog*** with a brush.

Figure 4: Qualitative comparison between our TAMoE model and the Base model on describing novel activities.

C.5 Qualitative Comparison

Figure 4 showcases two qualitative examples on the unseen test set. In the first video about "*painting fence*", the Base model has no linguistic knowledge of the concept "*fence*", while our TAMoE model successfully recognizes it and produces a more pertinent description. In the second example about "*grooming dog*", the Base model fails to recognize the actual action though already knowing the objects, while our model generates a more accurate description of the video.

D More Related Work

Video Captioning Since S2VT [33]'s first sequence-to-sequence model for video captioning, numerous improvements have been introduced, such as attention [42, 44], hierarchical recurrent neural network [43, 22, 5, 28], multi-modal fusion[13, 27, 37], multi-task learning [23], etc. Meanwhile, a few large-scale datasets are introduced for video captioning, either for single-sentence generation [14, 40] or paragraph generation [25]. Recently, [18] propose the dense video captioning task, which aims at detecting multiple events that occur in a video and describing each of them. However, existing methods mainly focus on learning from paired training data and testing on similar videos. Though some work has attempted to utilize linguistic knowledge to assist video captioning [29, 31, 8], none of them has formally considered zero-shot video captioning to describe videos of novel activities, which is the focus of this study.

Novel Object Captioning in Images Recent studies on novel object captioning [3, 32] attempt to describe novel objects not appearing during training. Zero-shot video captioning shares a similar spirit in the sense that it also generates captions without paired data. But zero-shot video captioning is a more challenging task: images are static scenes, and methods based on noun word replacement can perform well on novel object captioning [2, 39, 19]; while describing novel activities in videos requires

both temporal understanding of videos and deeper understanding of the social or human knowledge of activities beyond the object level. Different activities need different captioning strategies, as well as share some common characteristics. Motivated by this, our method learns the underlying mapping experts from the latent representations to the vocabulary, with a topic-aware gating mechanism implicitly transferring the utilization, which is orthogonal to these methods for novel object captioning in images.

Mixture of Experts The mixture of Experts (MoE) is originally formulated by Jacob et al. [15], which learns to compose multiple expert networks with each to handle a subset of the training cases. Then MoE has been applied to various machine learning algorithms [16, 10], such as SVMs [9], Gaussian Processes [30], and deep networks [1, 38]. Recently, Shazeer et al. [26] proposes a sparsely-gated mixture-of-experts layer for language modeling, which benefits from the conditional computation. Yang et al. [41] extends it to a mixture of Softmax to break the softmax bottleneck and thus increase the capacity of the language model. In this work, we exploit the nature of MoE for transfer learning by training a topic-aware gating function to compose primitive experts and adapt to various topics.