Unified Transformer for Multi-modal Few-shot Classification

Stanford CS229 Project Report

Songlin Li
Department of Computer Science
Stanford University
svli97@stanford.edu

Fang Qin
Department of Electrical Engineering
Stanford University
fangq@stanford.edu

1 Introduction

Language contains rich information which helps us understand the world. Human can easily identify "cars" by learning the descriptions of "cars" without going through thousands of pictures. By encoding abstract concepts into natural language, human can easily generalize to unseen samples even if the samples contains noises. In this project, we explore an under-explored multi-modal setting where ground truth natural language descriptions are available during training time but unavailable during test time. We choose few-shot classification [3] to test model's ability to quickly generalize to new samples. Previous work has demonstrated that architectures such as Convolutional Neural Network (CNN) with Language Models as bottleneck [1] or regularizer [2] can excel at this task. With the rise of Transformer [3], Transformer-based models were developed to tackle multi-modal challenges. In this project, we aimed to adapt LXMERT [4] to our multi-modal few-shot classification setting and analyze the performance of various training strategies.

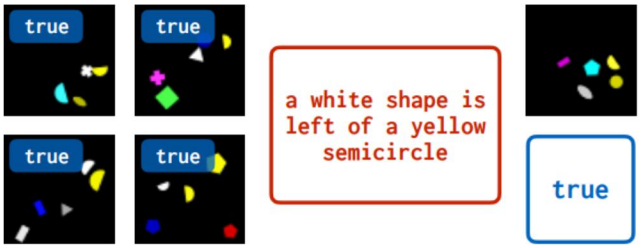


Figure 1: A illustration of our version of few-shot classification problem [1]. N = 1, K = 4.

1.1 Few-Shot Classification

In a few-shot learning problem, the dataset consists of thousands of tasks. Each task has a support set that contains N classes. Each class has K images. Each task also has a query set that contains images that may or may not from the same classes as the support set. For each query image, the model predicts which class the image belongs to. Captions are available in the training split only.

2 Related Work

A great amount of research work are focusing on bridging linguistic and visual information, such as visual reasoning [5], visual question answering [6] and language assisted visual classification, etc.

2.1 Language assisted visual classification

He and Peng [7] proposes two-stream model combining vision and language for fine-grained image classification but it doesn't have transfer. DeViSE [8] is a deep visual-semantic embedding model for zero-shot prediction. Xing etc [9] proposes a model to enhance metric-based few-shot learning

methods, but captions are provided for both model at test time. Learning with Latent Language model [1] and Language-shaped learning (LSL) model [2] for few-shot classification have no language at test time where the first model use language as a bottleneck and the second one uses visual information shaped by language to avoid the bottleneck.

2.2 Transformers based vision-and-language model

Most state-of-the-art transformer based multi-modal models follow a similar structure: each modality is assigned with one encoder, followed by another encoder that aggregates outputs from both encoders. Most architectures employ some kind of convolution layer as the initial feature extractor. Transformer layers will then attend to different tokens and aggregate the information by aligning visual tokens with language tokens.

3 Dataset and Features

We use the dataset ShapeWorld [10] as our training and testing dataset which contains 9000 training, 1000 validation, and 4000 test tasks (Figure 2). Each task is consisted of 4 images showing a visual concept with English language description. The visual concept describes how two objects are spatially related to each other with their color and/or shape information. There are also 2 to 3 shapes presented as the distraction for the task. Given a query image, our task is to predict whether it belongs to the virtual concept or not. The figure below shows an example of out tasks. The support data consists of images with the text on the right describing the content of the images. The model will tell whether the query image obtain the rule described in the text as the support images.

Figure 2: Example of tasks with image and query for ShapeWorld dataset.

4 Method

4.1 LXMERT

LXMERT has demonstrated state-of-the-art performance on many multi-modal tasks such as Visual Question Answering. LXMERT utilizes transformers [3] as its backbone architecture. One of the key improvement of LXMERT compared to generic Transformer-based models is that LXMERT is able to take in image features and text tokens together. The image features are packed in to a sequence and passed into LXMERT similar to sentences. The transformer layers allow LXMERT to attend to feature tokens across modalities. A special positional encoding was added to allow the model to identity whether the input is image or text.

In the few-shot learning setting, instead of using object-level feature as done in the original work, we experimented with two visual feature extraction method:

- using the output of a pre-trained 16-layer VGGNet (VGG16) as features [11]
- using raw image patches (we are breaking images into 7 x 7 patches).

We use Bert Tokenizer to tokenize the descriptions. Under the first setting, each image is corresponding to one visual token. Consequently, each training task has one visual sequence which consists of visual tokens from both support set and query set. We use the output of LXMERT's last layer as the hidden representation of a token. We encode the support set by taking the mean of support images' representations. Similarly, we encode the query set by taking the query's image's corresponding representation. For caption, we follow the Bert style by taking the representation of the [CLS] token.

We experiment with taking the mean of caption sequence instead of using the [CLS] token. There is no significant difference. A classification head takes in visual representations of the support set and the query set, then predict whether they share the same visual concept. x_n denotes the features extracted by VGG from the nth image. w_m denotes the ith token of concept description. rep denotes LXMERT.

$$p = classifier(rep(x_{s1}...x_{sN}), rep(x_{q1}), rep(w_1...w_l))) \tag{1}$$

Under the second setting, we encode each image separately: each visual sequence consists of only one image's patches. We then predict by calculating the dot product between the representation of support and query set. x_k denotes the kth patch. In both setting, we take $p > 0.5$ as the threshold to predict positive.

$$p = \frac{1}{K}\sum_{s=1}^{K} rep(x_{s1}...x_{sk}, w_1...w_l) \cdot rep(x_{q1}...x_{qk}) \tag{2}$$

We optimize our model with binary cross entropy loss.

$$Loss = -\sum_{i=1}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{3}$$

In terms of optimizer, we use the Bert version of Adam. We observed that warm-up steps are necessary for LXMERT to learn. The reason might be that initial gradients are bias due to the bias of the mini-batches. Therefore, a large learning rate will cause to LXMERT model to stuck in an undesired local minima. We use linear warm-up and learning rate decay: learning rate starts with 0 before gradually going up then going down linearly.

4.2 Mixed Modality Training

Since we don't have access to captions during testing, we have to train our model to be capable of inference without the extra modality. Unlike LSL [2] or L3 [1] where language is used as extra supervision or bottleneck, our approach directly use language as input. We will illustrate some important properties about transformer strcuture in the later section.

4.2.1 Mixed Modality Training as Multi-Task Learning

One technique we experiment with is training mixed modality problem as multi-task learning. Instead of having one classification head only, we added one extra classification head that takes in caption's representation and query image's representation. We hypothesize that by allowing direct comparison between visual representation and language representation, the vision encoder can learn to attend to the spatial relationship that's encoded in the caption.

4.2.2 Random Sampling

One of the simplest but highly effective method is random sampling. During training, we randomly decide whether to train our model with ground truth caption or support image. We will use separate classification head for two cases as mentioned above. Intuitively, since we are not going to have captions during test time, by training without captions we hypothesize that model will rely less on the extra information thereby be able to make the correct prediction without language support. Our results show that this step is necessary for LXMERT to achieve a high performance.

5 Experiments/Results/Discussion

5.1 Experiments

Across all experiments, we use 5e-6 as our learning rate and a 0.05 warm-up ratio (5% of total epochs are used as warm-up steps). Our batch size is 100 for the first visual feature extraction setting, which is the same as the LSL and 24 for the second setting. This difference is largely due to memory constraint since LXMERT will produce hidden outputs for all visual tokens and the visual sequence under the second setting is much larger than the first one (6 vs 49). We train our models with 800 epochs. Based on our observation, all settings can converge before reaching the end. We use accuracy, precision, and recall as our main quantitative evaluation metrics.

5.1.1 Baseline

Our baseline model is LSL, an end-to-end model which use the visual representations shaped by language[2]. We also compared the performance across training strategies, specifically the availability of caption.

5.1.2 Oracle Language

Although previous works have demonstrated that utilizing language as bottleneck or regularizer can improve models' performances, it's unclear whether presenting language as input can improve model performance. To illustrate the effect of correct description, we feed the model with the ground truth concept description of the support set and the query images. This setting's performance provides an indicator of the potential upper bound the model can achieve due to the presence of perfect captions. The model's performance was boosted by 24 %. Additionally, model training is much more stable and can generalize better. Figure 5.2

5.2 Results

The performances of all settings under the first feature extraction method are recorded in Table [2]. As shown in the table we are able to match LSL's performance (67.29 % accuracy) even though our model doesn't use language as supervision. This illustrate an interesting property of transformer-based multi-modal architecture: using extra modality as input can improve performance on missing modality setting. In terms of success of training, random sampling is essential. This observation aligns with our hypothesis that it's necessary to to train under the same setting as testing. Unfortunately, multi-task is not as beneficial as expected. This phenomena indicates that the language encoder without additional gradient flow (provided by extra classification head) can be trained as well as one with. This is reasonable since the last few layers of LXMERT focusing on aggregating information across both modalities.

	Oracle Captions	Image only	Random Sampling	Multi-Task Only	Random Sampling + Multi-Task
Accuracy	77.94	62.89	66.94	51.70	66.43
Precision	79.55	61.65	67.27	52.73	64.50
Recall	74.67	67.01	65.23	46.21	73.49

??

Table 1: Test accuracies comparison between different settings using VGG features

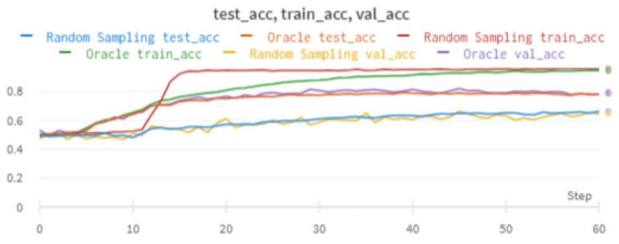


Figure 3: Training behaviors of selected methods

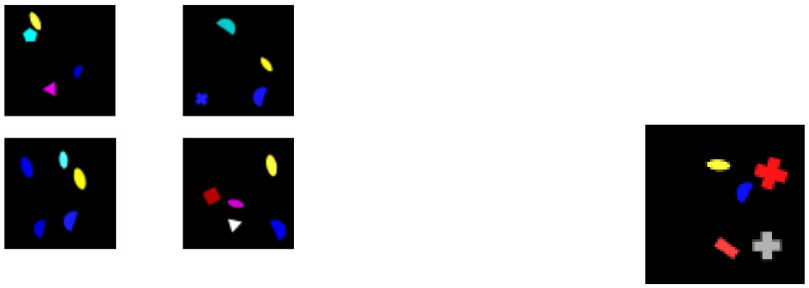
With the second feature extraction method, we experimented with oracle caption and image only setting. Unlike the first feature extraction method, oracle captions' presence doesn't improve the performance of the model. This is very likely caused by the dominance of visual modality since vision encoder receive much more training signals than language encoder due to a much longer sequence length. We conclude that language has limited effect when the second feature extraction method is employed. This result plus the performance of the oracle setting of the first feature extraction method indicated that one way to achieve high accuracy is potentially let one modality dominate the other. In the first oracle caption setting, we are only feeding query image and captions into LXMER: the language sequences are longer than visual sequence. However, language dominance is able to achieve a much higher accuracy compared to visual dominance, demonstrating the value of captions.

	Oracle Captions	Image only
Accuracy	67.06	68.75
Precision	62.72	64.16
Recall	84.59	85.72

Table 2: Test accuracies comparison between different settings using image patches

5.3 Discussion

We selected few samples with wrong prediction in the test dataset for analysis from model trained with random sampling. One of the tasks is shown in figure 4 and 5 below. The support images' concept are aligned with the caption. In the query image, there is a shape above rectangle but its green. It's hard for model to learn the color different so it results in error. For task in figure 5, concept in support images and query image are all aligned with the caption but it is still predicted wrongly.



(a) support (b) query

Figure 4: Task with caption: A yellow ellipse is above a blue semi circle.

(a) support (b) query

Figure 5: Task with caption: A shape is above a blue rectangle.

6 Conclusion/Future Work

In this report, we demonstrate a new method on how training with multiple modalities can improve performance on a missing modality setting. We also illustrate an important property of transformer-based multi-modal model: performance improvement can be achieved even without additional supervision from other modality. Additionally, sampling is essential for the model to generalize during testing when one of the modality is missing. With additional resource, we would like to experience a different visual feature extraction method that has been applied in many other tasks: using object detection model to extract objects from the image and use the hidden outputs are feature sequences.

7 Contributions

- Contribution of team members:
  Fang Qin: Evaluation and training of the LXMERT model; visualization of the results
  Songlin Li: Implementation and training of the LXMERT model; setting up LXMERT for ShapeWorld Dataset
- TA : Jeff Z. HaoChen

References

[1] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June 2018.

[2] Jesse Mu, Percy Liang, and Noah Goodman. Shaping visual representations with language for few-shot classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4823–4830, Online, July 2020. Association for Computational Linguistics.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

[5] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.

[7] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[9] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[10] Alexander Kuhnle and Ann A. Copestake. Shapeworld - A new test methodology for multimodal language understanding. *CoRR*, abs/1704.04517, 2017.

[11] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks For Large-Scale Image Recognition. Technical report, 2015.