Grounded Situation Recognition with Pre-trained Models

Professor: Kai-Wei Chang
Author: Shanxiu He

December 19, 2020

1 Abstract

Recently, Vision and Language models that employ traditional vision backbone along with BERT are shown to be effective in many tasks, such as Visual Question Answering and Visual Commonsense Reasoning. In this study, we wish to apply such model to a new task called Grounded Situation Recognition (GSR). While the state of the art model employs LSTM as its language backbone, we use LXMERT model to process visual features and language features and examine if the pre-trained weight provides any benefit to the learning. The results indicate that, even without a joint training with object detector, our model is on par with previous CRF baseline on Situation Recognition in validation results.

2 Related Works

Grounded Situation Recognition GSR [2], comes with the newest dataset called SWiG, is a new task which requires a model to provide both structural role predictions and grounding boxes predictions. SWiG dataset builds on top of the original imSitu dataset [5] and adds 278,336 bounding-box groundings to entity class. This task takes inspirations from semantic role labeling. To be more specific, the task's language counterpart would test a model by predicting a realized frame with different semantic roles for each verb given. While imSitu [5] attempts to predict such frame based on additional image information, we also need the model to predict the groundings for each role, which requires more visual reasoning abilities for our model.For instance, in the picture shown, the model would predict values given the image as well as the possible bounding boxes that correspond to the objects as shown in Figure 1. We refer the audience to the original paper for more detailed task definition.

SWiG and imSitu Dataset imSitu is one dataset collected from FrameNet, a verb and role lexicon developed, and contains over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. SWiG uses a

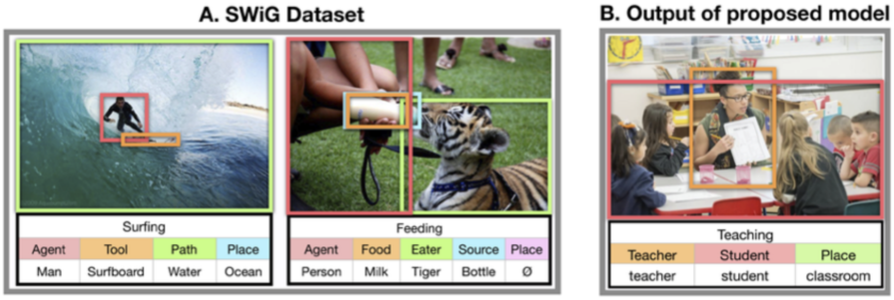

Figure 1: Two examples from the dataset [3]: semantic frames describe primary activities and relevant entities. Groundings are bounding-boxes colored to match roles. The second picture is the output of original model (dev set image)

different resolution of the original dataset, retains 55,000 images among the original images, and add 278,336 grounding annotations.

VisualBert and LXMERT

Two state-of-the-art vision and language models, built with a large-scale Transformer architecture. VisualBERT uses a single transformer structure that takes the concatenation of dense image and text embedding and predicts result on the last layer. LXMERT is designed to process different levels of inputs first and then combine. It consists of an object relation encoder, a language encoder, and a cross-modality encoder. We focus LXMERT's image representations extracted from its object relation encoder in the first second and imitates VisualBERT structure in the building of a new model of SWiG.

ISL and JSL

Some previous implementations are done on these tasks. ISL and JSL are both newest models proposed by [3]. They both use ResNet as visual backbone and then utilize LSTM to process mainly language information. The key difference is that ISL separates the prediction of noun frame and grounding while JSL trains LSTM to take in both ResNet features and language input. Currently, the state-of-the-art model is JSL and we plan to match the performance on both setting.

3 Methods

In this project, we build a LXMERT-based model suiting for GSR task. We first extract the features from new dataset via FRCNN [1], the same setting as LXMERT in its experiments with VQA and GQA. Then, we merge the word embedding as well as the extracted visual features into our models.

3.1 Extract FRCNN Features

Following the setting of LXMERT, we extract top 36 bounding boxes from each image in SWiG dataset to be utilized in training or testing time.
Sanity Check As our current model does not take object detector into training process, we verify the recall of currently extracted boxes and ground truth grounding annotations in SWiG. After the testing, we observe that FRCNN at least covers 90% of the ground truth boxes in each of the training, testing, and validation image sets. This ratio is relatively high which indicates that we could employ these FRCNN features directly in our model.

3.2 Frame Prediction LXMERT for SWiG

In this section, we detail current implementation our model. As shown in Figure 2, our model takes in a ground truth verb and an image. Given FrameNet, we pass in the ground truth verb as well as the corresponding frame, such as [PLACE], [AGENT] and [TOOl] in the language portion and the extracted top 36 features as the visual input. After fusion, we use the language output in corresponding position as our predictions for each noun in the frame.

In the future, we would add grounding supervision to the model. For the independent model, we would feed the vision output to a linear classifier, projecting into the max possible roles for each verb (maximum 6 in our setting). Then, for each role, we set up a threshold for passing, indicating whether the role is grounded or not, and select the top feature as our prediction. In the future, we also plan to further regress on the locations of the grounding boxes instead of directly using the extracted features or even include the extracting process in our joint model.

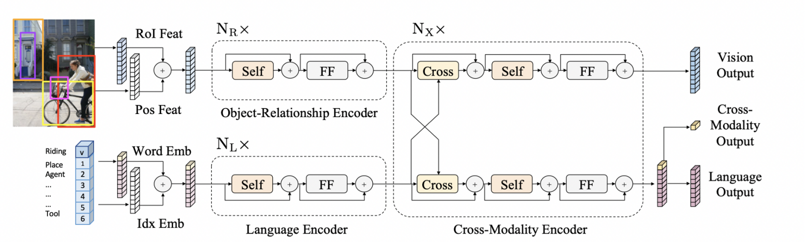

Figure 2: Frame Prediction Model for SWiG task, where BERT takes in a realized frame and predicts all of them at the same time. The image is taken from the original LXMERT [4] paper and noted with our own modification.

3.3 Plan: Recurrent LXMERT for SWiG

The recurrent model is majorly the same with the independent model, except for that we only input one role into LXMERT each time. One other possible implementation is to also input the grounding and noun prediction in previous timestamp to boost the structural prediction performance, imitating JSL's looping prediction.

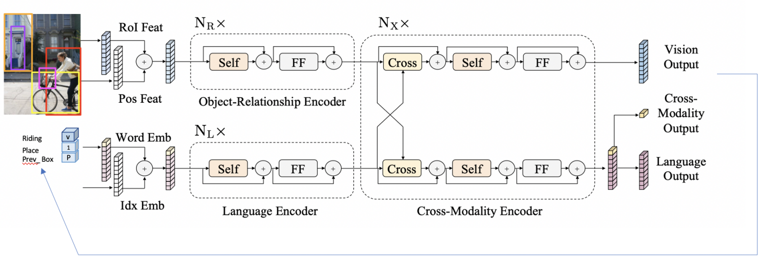

Figure 3: Recurrent LXMERT Model for SWiG task, where BERT takes only one role and predicts one at each time.The image is taken from the original LXMERT paper [4] and noted with our own modification.

4 Experiments

We conduct experiments on Frame Prediction LXMERT Model with only noun prediction and fixed object detector. We use a learning rate of 1e-5, batch size of 16, and the training takes approximately 24 hours on two GTX 1080 Ti cards. We conduct experiments on LXMERT with pre-trained LXMERT weights and $Bert_{base}$ weights. We observe that learning rate is crucial and could easily results in local optimal if not chosen carefully and surprisingly $Bert_{base}$ weights almost always perform better than LXMERT weights even if previous studies suggest the oppose.

5 Results and Observations

We conduct experiments on Frame Prediction LXMERT model where its grounding part and recurrent counterpart are still in progress.

In this figure, one notable observation is that even without joint training with the detector, our performance is on par with the previous baseline with only noun prediction, which is a fair comparison except we did not include the joint training. For ISL and JSL, they would also benefit from the grounding supervision which neither do our current models or CRF has access to.

From this figure, we could suggest that the pre-trained models benefit from the corpus training (at least from the language side) and make themselves comparable to the joint model. However, this result could also be resulted by the increased model size of LXMERT, compared to CRF which simply regresses on image features. For future studies, we should examine more in depth for the cause of the better performance and produce comaprable results with other baselines.

Model	Value-any	Value-all
Bert*	63.33	24.21
LXMERT*	62.92	24.19
CRF	65.7	29
JSL	73.53	38.32
ISL	72.77	37.49

Table 1: Frame Prediction Model with $Bert_{base}$ weights and LXMERT pre-trained weights. These six models are not directly comparable for different training settings.

6 Future Works

The first step we have in mind is to complete the current model and match with the baselines. In the future, we plan to extend this project to more directions and impose more motivations to our current task.

Image Captioning We wish to use Image Captioning data as our supervision. Since our model is structural prediction, it could potentially benefit more from these datasets inside of current dataset that LXMERT pre-trained on, such as Visual Genome.

Commonsense Extractor In future, we wish this model could turn into a working prototype that extracts a Situation based on the image and/or texts given. Also, it is possible to extract commonsense based on image captioning data as we mentioned before. A working demo would be helpful for the clients to see the inference and reasoning process though much higher accuracy and robustness would be required for current model.

Explainable Models We also plan to design probes for this task and see what the model is actually learning in the process. Possible directions in this route is to use few-shot learning and see if the model could extend to other conditions, exhibiting the ability to truly reason.

7 Acknowledgement

We wish to express graditude to Professor Chang as well Harold Liunian Li. Without their guidance, this project could not move this far.

References

[1] Peter Anderson et al. *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. 2018. arXiv: `1707.07998 [cs.CV]`.

[2] Sarah Pratt et al. *Grounded Situation Recognition*. 2020. arXiv: `2003.12058 [cs.CV]`.

[3] Sarah Pratt et al. "Grounded Situation Recognition". In: *arXiv preprint arXiv:2003.12058* (2020).

[4] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *arXiv preprint arXiv:1908.07490* (2019).

[5] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. "Situation Recognition: Visual Semantic Role Labeling for Image Understanding". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.