Paper review for "Gupta et al: Contrastive Learning for Weakly Supervised Phrase Grounding"

EECS 598 Paper Review – Week 11 - Changyuan Qiu

Humans can associate words with image regions quickly while watching movie with subtitles or making a diagnosis from MRI scans, and this association task is referred to as "phrase grounding" in vision-language tasks. Here the authors propose a novel mutual information (MI) based contrastive training framework for weakly supervised phrase grounding. The key idea is: first use an object detector and a language model to extract image region and word features respectively, then using contrastive training to trains a word-region attention mechanism to maximize a InfoNCE lower bound on word-region mutual information with a combination of 2 losses: (1) true image vs. negative image with the same caption (2) true caption vs. negative caption with the same image. For the second loss the authors also proposed a method to construct context-preserving negative captions by only substituting a noun word in the caption with contextually plausible (such contextually plausible captions are extracted according to the word probability rankings for the [MASK] token in BERT with near-synonyms & hypernyms excluded) but untrue words, e.g., replace "donut" with "cookie" in the caption "Chocolate 'donut' in front of a computer".

Prior works typically formulated phrase grounding as a multiple instance learning problem, where images are split into bag of region proposals, that caption scores are aggregated across all regions to make a prediction for the image and then the problem becomes a fully-supervised image-caption classification problem. The key difference between this work and prior works lies in the introduction of contrastive learning objective, which is very intuitive in that maximizing caption-region MI forces the model to downweight incorrect caption-region pairs and attend to the correct pairs and thus strengthen the model's ability to understand the phrase grounding task better and gain better performance. Their model trained on COCO achieve 76.74% pointing accuracy on Flickr30K Entities dataset, shows a 5.7% absolute gain compared with the previous SOTA Align2Ground. Such contrastive objective also shows data efficiency that their model trained on Flickr30K Entities which has only about 1/3 image-caption pairs as COCO achieves 74.94% pointing accuracy, which is quite close to the performance of the model trained on COCO. Their ablation studies also showed that the use of contextual plausible negative captions extracted with BERT gives a ~8% performance gain compared with the model using randomly sampled negative captions from training data, and excluding near-synonyms & hypernyms give a further ~2% improvement.

Regarding limitations of this paper, we could find that the usage of contextual plausible negative captions plays a key role in the performance gain (~10 %), which are extracted with the advanced pre-trained language model BERT, and all caption representations are also extracted with BERT, while prior work does not make use of this powerful model. It remains uncertain whether the performance gain really comes from the contrastive learning objective or from the powerful language model. But it is true that such formulation of the problem opens a way to make use of the great advance in language models these days. Also, the author simply add the two losses together and assign equal weights to them, and maybe he could add a trainable parameter to assign different weights to the 2 InfoNCE loss since it is intuitive that one pairing could be more important than the other and should get more focus.