Multimodal Few-Shot Learning with Frozen Language Models

Simple Overview

- In recent years, deep learning techniques have been applied and achieved great results in many domains such as computer vision and natural language processing. However, having large enough data to train on is essential for many deep learning applications to achieve high performance. For some applications, it is often expensive or hard to collect enough training samples. Thus, few-shot learning research has gained increasing attention over recent years trying to solve the problem. **Few-shot learning aims to make the model learn to perform a new task and generalize well with only a few training samples. And this will solve the problem of data.** Language models are few-shot learners but despite their powerful ability, it is limited to text only. In this paper, they present an approach to how to take advantage of the power of language models in the field of computer vision. It is called frozen which is a method for giving pre-trained language model access to visual information in a way that extend its few-shot learning capabilities to a multimodal setting, **without any updating its weights.** Frozen consists of an image encoder which trained to encode images into the word embedding and the language model generates captions for those images. Without any update in language model parameters, **but gradients are back-propagated through it to train the image encoder.**

- The produced multimodal has proven its ability to learn wide range of tasks such as visual question answering using zero shot learning with no training or gradient updates, by providing the model with image and question, it can answer. The ability of model will improve when providing it with more than one pair of image and question which is called few-shot learning

- Analyzing the extent to which Frozen can leverage the encyclopedic knowledge in the language model towards visual tasks. For example, we have a photo of airplane, and a question who invented it? Can the multimodel handle such a task? Yes, that based on the pretrained language model which trained on a billion of information and articles. The vision encoder can detect that the image contains airplane, the pretrained language model uses the information to retrieve that factual knowledge that airplanes were invented by wright brothers.

- In the multi-modal setting, fast-binding refers to a model's ability to associate a word with a visual category in a few shots.

Main Contribution

- Training vision encoder based on language models, the new model is a multimodal few-shot learner retains all of the features of large pretrained language model, but can also process text and image inputs.
- Verify that prompting them with both visual and language information can be strictly more effective than doing so with language information alone

Limitations

Despite the power of frozen as we saw in the paper, it still has drawbacks. multimodel achieves very good performance on specific tasks but it's still far from state-of-the-art performance on other tasks in which using the full training set is better than few-shot learning.