SUTS for CVLMs

David Krug

University of Washington

krug@uw.edu

Introduction

In this paper I introduce a Spatial Understanding (SU) Test Suite (TS) for Contrastive Vision-language Models (CVLMs) for my CLMS degree from the University of Washington. I begin by performing a literature review in §2, including recent developments related to transformers in the NLP and computer vision fields, existing VLMs, understandings about prompt sensitivity and prompt engineering, and analysis of the spatial understanding of VLMs. In §3 I describe my projects, SUTS in detail. In §4 I draw connections between the skills I used to develop SUTS and the skills I learned while completing my CLMS.

Literature Review

Transformers

Much of the recent advancements in language modeling owes its success to the Transformers (Vaswani et al., 2017). Models which use a Transformer architecture have been shown to have state-of-the-art performance on a variety of natural language processing (NLP) tasks, such as machine learning (Ott et al., 2018), text classification, and question answering.

Transformers achieve their high performance by relying on a self-attention mechanism in order to learn the relationships between elements in a sequence. This allows Transformers to model long distances dependencies between input elements. This also means Transformers can process a sequence in parallel, unlike recurrent networks which must process elements in a sequence one by one. Additionally, Other benefits of Transformer architecture are that it requires minimal inductive biases, it can be applied to multiple domains (as explored in §2.2) and can be scaled up to massive numbers of parameters, such as 1.6 trillion in the case of the Switch Transformer (Fedus et al., 2022).

The self-attention mechanism and massive scalability means that Transformer models are usually trained in two stages, first with supervised or self-supervised pre-training on a large training set, and then task specific fine-tuning on a relatively smaller training set. A typical pretraining task is next word prediction, although models like BERT (Devin et al., 2019) attend to the context of a given word in an a-directional fashion, so masked language modeling and next sentence prediction are used instead.

Vision Transformers

The obvious successful application of Transformers to the field of NLP has encourage their application to the field of computer vision. n vision models. Vision models which use Transformer architecture have found success at tasks such as image classification (Dosovitskiy et al., 2020), object detection (Carion et al., 2020), low-level image manipulation like resolution enhancement (Yang et al., 2020) and colorization (Kumar et al., 2021), image generation (Ramesh et al., 2021), 3D analysis?, and vision-question answering.

Earlier work in computer vision used CNNs to produce image encodings (LeCun et al., 1989). CNNs can be enhanced with non-local self-attention (Wang et al., 2017) or criss-cross self-attention (Huang et al,. 2019). Unlike these approaches though, Vision Transformers (ViT) (Dosovitskiy et al., 2020) applies self-attention with a Transformer architecture based on (Vaswani et al., 2017) with minimal changes. The input embeddings for ViT are flattened image patches paired with positional embeddings.

Various adaptation over the original ViT architecture have been made, such as introducing a teacher-student strategy and distillation tokens (Touvron et al., 2020), using all image patches to contribute to the loss calculation instead of only the classification token (Jiang et al., 2021), subdividing image patches and computing attention at both levels (Han et al., 2021), and recursively aggregating neighboring tokens into one token (Yuan et al., 2021).

Vision-language Models

A vision-language model (VLM) is any model which jointly processes information from the vision modality and the language modality. Although this covers a wide range of models spanning many different architectures, most state-of-the-art VLMs use Transformer architecture. The huge leap forward in performance at many NLP and computer vision tasks that Transformers enabled thanks to their pre-training paradigm has propelled Transformers to be the dominant architecture in the field, and indeed all the spatial understanding analyses in this literature review use Transformer architecture. That being said, some VLMs that do not use Transformers instead use RNNs (Donahue et al., 2016), MLPs (Wang et al. 2019), and bag-of-words text model optimized with canonical relation analysis (Gong et al., 2012).

As for Transformer based VLMs, they can be categorized into two categories: multi-stream and single-stream. Multi-stream VLMs feed each modality into a sperate Transformer and then learn cross-modal representations with a third Transformer, while single-stream VLMs use multi-modal inputs to feed a single Transformer.

The first extension of BERT (Devin et al., 2019) into a multi-modal domain was ViLBERT (Lu et al. 2019), a two-stream architecture in which each stream is dedicated either to language or vision inputs. The two-streams consists of a series of Transformer blocks, and cross-modal representation are learned by passing the keys-value pairs from one modality to the attention head of the other. ViLBERT is pre-trained on tasks like predicting if an image an text are related, and then fine tuned on down-stream tasks like visual question answering and caption based image retrieval.

Some similar models to ViLBERT are LXMERT (Tan and Bansal, 2019), which uses an object-relation encoder and additional pre-training tasks, and PEMT (Lee et al., 2020) which aims to learn cross-modal information in an end-to-end manner and which reduces memory requirements by sharing parameters across layers and modalities.

CLIP (Radford et al., 2021) is another two-stream VLM , however it has no third transformer to learn cross-modal representations. Instead, it learns cross-modal connections with a training objective that aims to align the otherwise separate language encodings and vision encodings. During training, CLIP takes as input N image-text pairs with the goal of maximizing the cosine similarity between the N correct image-text pairs while minimizing the cosine similarity of the N2 – N incorrect pairs. The language encoder used is similar to the original ViT, with modifications described in Radford et al. (2019). Two different vision encoders were tested, the first being a Transformer based on ViT and the second being ResNet (He et al., 2015). It was trained on 400 million text-image pairs from the internet. CLIP was able to achieve very impressive one-shot performance at image classification, although has a somewhat high training computation cost.

In contrast to multi-stream VLMs, single-stream VLMs use a single series of Transformers to process multi-modal inputs. VisualBERT (Li et al. 2019) is one such model, and it takes language and vision features as input into the same Transformer which must learn the relations between the two domains through the pre-training tasks of filling in missing language tokens and differentiating between true and false image captions.

VL-BERT (Su et al., 2019) is another single-stream model, and it takes either word or region-of-interest features as input. Furthermore, VL-BERT is trained both on both vision-language datasets as well as text-only datasets.

Unicoder-VL (Li et al., 2019a) achieves state-of-the-art performance by using three task-agnostic pre-training tasks: masked language modeling, masked object identification, and vision-language matching.

Unified VLP (Zhou et al., 2020) uses a single transformer network for both the encoding and decoding instead of independent encoder and decoder networks. It is able to be fine-tuned on very different tasks, from image captioning to VQA.

UNITER (Chen et al., 2020) is yet another single-stream VLM, and it uses four pre-training tasks: masked language modeling, masked region modeling, image text matching, and word-region aligning. It also uses conditional masking on language and vision pre-training, conditioned on the other domain input.

Finally, Oscar (Chen et al., 2020) aims to solve the issue that previous VLMs must learn to align image-text semantics by “brute force”. It does this using detected object tags in the training data, along with the image and text itself. Oscar performs better at vision question answering, NLU, and image captioning than models which do not use object tags during training.

Interpretability and Analysis

The dominance of Transformer architecture in NLP is clear, and with each new variant of a Transformer based VLM comes a slightly higher performance on common VL tasks like ImageNet (Deng et al., 2009). However, performance at real-world tasks is often too complicated to be measured with a single metric (Molnar, 2022). Furthermore, for models which perform task that have dramatic outcomes (eg. diagnosing caner), it is important to understand what cause a certain prediction (eg. a visible cluster of malignant cells). Models are more useful and safer to humans when we can understand why it makes the decision and predictions that it does, what aspects of its input it pays attention to, and how the qualities of its input affect its output.

In an attempt to understand what aspects of an image are attended to by computer vision models, Geirhos et al. (2018) evaluate a CNN’s performance at classifying images in ImageNet after various image manipulations, such as removing texture information (silhouette) or applying the texture of one object to the shape of another. They found that CNNs rely on texture much more than shape to identify objects. Furthermore, they found that training on a stylized version of ImageNet have less bias toward texture, higher accuracy at ImageNet, and higher robustness to image distortions.

Inspired by that^, Naseer et al (2021) apply a similar experiment to ViTs.

Spatial Understanding

Prompt Sensitivity

SUTS for CVLMs

Overview

Image Generation

Relation Detection

Sentence Generation

CLIP Evaluation

Connection to CLMS

ReferencesDiagram

Description automatically generated