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, analysis of the spatial understanding of VLMs, and understandings about prompt sensitivity and prompt engineering. 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.

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.

Finally, although image synthesis models are not strictly VLMs as I have defined them (they do not take visual signals as input), it is worth discussing them here. DALL-E (Ramesh et al., 2021) is perhaps the most well know of these. It works by…

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 results in less bias toward texture, higher accuracy at ImageNet, and higher robustness to image distortions.

Like Geirhos et al., Naseer et al (2021) apply a similar experiment to ViTs. They find that compared to CNNs, ViTs exhibit less reliance on texture, whether trained on the stylized ImageNet or not, and that training ViTs on the stylized ImageNet reduces texture-reliance even more. They also find that by introducing a ‘space token’ and adapting attentive distillation (Touvron et al., 2021), a single ViT can exhibit both texture bias and shape bias with separate tokens. They also examine the effect of disturbing the structure of an image by removing positional encodings from a ViT. They find that positional encodings actually contribute minimally toward giving ViTs structural information about an image, and that ViTs are more robust to having positional encodings removed than CNNs are to having patches of an image shuffled. They conclude that “positional encoding is not absolutely crucial for right classification decisions”

Another method of gaining insight into the inner workings of a Transformer model is through visualization. A visualization of the decision process of a Transformer should explain what regions of an image and provide the most information to allow the model to make its prediction. Visualization can help with debugging the models, verify that they are fair, and enabling downstream tasks. In gradient based visualizations, the gradient of each layer are typically multiplied by the input activations, as in Shrikumar et al. (2017). Attribution propagation-based visualization work by reclusively decomposing a network classification decision into contributions of its input elements, as in Montavon et al. (2017). Chefer et al. (2021) introduces a novel visualization method specifically for Transformers which uses an attribution propagation-based scheme. They compute scores for each attention head in each layer and integrate those scores by incorporating both relevancy and gradient information. Their visualization method is able to handle both positive and negative attributions, and it produces different classes visualized.

Spatial Understanding

Spatial understanding is the ability to understand, reason about, and use visual and spatial relationships between objects. In humans, this skill is innate. Spatial understanding is essential for complex understanding of visual scenes, and so a good bi-modally trained language models must have a solid sense of spatial awareness. Furthermore, spatial awareness is a key task for downstream tasks, such as robotics (Shridhar et al., 2021).

~~One specific aspect of interpreting VLMs in concretely understanding what abstract information they encode, such as spatial relations.~~

A study into the spatial understand of a visionless model is Ghanimifard and Dobnik (2017). They train an LSTM to predict the two-dimensional probability distribution of various spatial relations (eg. ‘left’, ‘right’), as determined by participant humans. Their model is trained on data that is grounded in spatial relations, which means input word-embeddings are concatenated with encoded location features in the form of 7x7 *spatial templates*. They find that when their model is trained on single-world relations (eg. ‘above’) and composed relations (eg. ‘above or below’), it is able to predict the distribution of unseen single-words that have only appeared in composed relations. Likewise, their model is able to predict unseen composed relations that contain seen single-word relations. Furthermore, they find that introducing distractor word which do not contribute to the spatial semantics of a relations does not actually significantly distract their model or worsen performance.

Zhang et al. (2020) evaluate large, pre-trained, transformer based language models like BERT (Devin et al., 2019) and eLMO (Peters et al., 2018) on their ability to represent scalable attributes like length, mass, and price. They trained linear probes on-top of the pre-trained models to recover a distribution over one of those three attributes. Their conclusion was that pre-trained models are bad at encoding length, mass, and price information. Bhagavatula et al. (2020) also evaluates large pre-trained models for their spatial understanding. They introduce a new dataset, ART, which evaluates the abductive commonsense reasoning of language models. They find a significant gap between BERT performance and human performance, and that BERT especially underperforms on test instances involving spatial or numerical descriptions.

~~Spatial BERT (Dan et al., 2020) is (one of) the first model that specifically uses Transformer architecture with the goal of accurately representing spatial relations.~~

The studies above describe models which are trained only on text information (or text information and spatial templates). In contrast, Collell et al. (2018) apply spatial understanding to a model which also gets input signals the vision domain, but only indirectly. For their task, a model must predict the coordinates and size of the bounding box of an object relative to another object, given the names of the two objects, a word denoting their spatial relation, and the position of the other bounding box. They find that their models generalize well and can predict the position of unseen objects. This indicates that the word embeddings provided as input have information about spatial properties embedded in them, or that objects with similar word embeddings have similar spatial distributions. They conclude that in general, “models acquire solid common sense spatial knowledge”.

As for analyzing studies about models which take visual signals as input directly, Liu et al. (2019) introduces CLEVR-Ref, itself a modified version of CLEVR (Johnson et al., 2017). CLEVR-Ref a diagnostic dataset aimed at evaluating a VLM’s ability to identify (by bounding box or segmentation) the objects in an image which are uniquely referred to by some referring expression. Referring expressions may pick out objects by referring to their size, shape, color, etc. Referring expressions may also pick out objects by referring to their relation to other objects, such as “second sphere from the left” or “sphere to the left of the cube” They found that models they tested had higher performance when tested on referring expression that mentions the color, shape, or visibility. This suggests that these features are more salient or easier to learn for VLMs than the size, ordinality, or material of an object.

Cirik et al. (2018) point out that state-of-the-arts VLMs involve complex neural parameterizations which make it difficult to interpret what is actually being learned, and they criticize existing referring expression datasets on the basis that they can be solved by exploiting biases in the datasets. They show that shuffling the order or words in a referring expression, discarding preposition, or even discarding the entire referring expression (using only the image is input) only results in a minor drop in VLM performance at locating the referred object. This suggests that existing referring expression dataset can be solved by exploiting statistical biases.

VSR (Liu et al., 2022a) is a dataset specifically aimed at evaluating the spatial skills of VLMs. Each instance in their dataset consists of one image, a caption, and a binary label. The binary label reflects if the caption truly or falsely describes the image. They find a significant gap between human and VLM performance, suggesting that VLMs have a hard time at learning spatial relations. Furthermore, they find the number of training instance for a given relations does not correspond strongly with performance at that relation. The also find that VLMs have a hard time generalizing spatial relations to unseen concepts, and that spatial relations involving orientation are particularly challenging.

Liu et al. (2022b) introduce another dataset for evaluating the spatial skills of VLMs. Other such datasets, like VSR, focus mostly on explicit relations, like ‘above’ and ‘to the left of’. Liu et al.’s dataset additionally tests implicit relations, like the relation between a person and a bike while the person is riding the bike (ie. ‘on’). Their dataset consists of images paired with both explicit captions (eg. ‘person on a bike’) and implicit captions (eg. ‘person riding a bike’). They also evaluate image synthesis models by providing an explicit or implicit caption and evaluating the resulting image. They find that models which take visual signals as input have more accurate and consistent spatial commonsense than text only models. They also find that image synthesis models have even more accurate and consistent spatial commonsense than non-synthesis VLMs.

PaintSkills (Cho et al., 2022) is a diagnostic data set for evaluating image synthesis models. They test four visual reasoning skills. They find that Transformer image synthesis models are better at recognizing and counting objects than they are at recognizing colors and understanding spatial relations. Conwell and Ullman (2022) also evaluate image synthesis models for their visual reasoning skills, specifically focused on spatial understanding. They give DALL-E 2 (Ramesh et al., 2022) various text prompts involving spatial relations (eg. ‘teacup in a box’), and had human participants decide if the generated images matched the text prompts. Their conclusion was that DALL-E 2 lacks commonsense relational understanding. Part of DALL-E 2’s challenge in generating accurate images was that it was much more biased towards producing frequently observed relations (eg. ‘spoon in a teacup’ is easier to generate than ‘spoon on a teacup’).

Briand et al. (2022) uses an adapted version of the SpatialVOC2K dataset (Belz et al., 2018) to evaluate the spatial understanding of VLMs. The dataset in their experiment consists of minimal semantic pairs of captions, one of which correctly describes an image and one of which incorrectly describes an images (eg. “a cat in a box” versus “a cat to the left of a box”). The goal of the VLM is to identify the correct caption. Additionally, they put each caption through 18 different sentence patterns to generate re-wordings of the original captions (eg. “a cat in a box” and “a picture of a cat in a box”). The goal of this was to identify which linguistic features most affect the model’s performance at encoding spatial relations. They evaluated CLIP’s performance and this task and found that CLIP does not produce robust encodings of the spatial relations between objects in a scene. Echoing results about DALL-E 2 found by Cho et al. (2022), Briand et al. that CLIP has a bias to select certain relations over others (eg. preferring ‘on’ rather than ‘under’, regardless of input image). This suggests that CLIP may be selecting the more frequently observed relation than attending to the facts of the image. Finally, their results showed that CLIP mostly ignored negation, treating a sentence with a negative polarity item the same as the positive version.

Prompt Sensitivity

As language model grow larger and larger, training a new model for each new task is not practical. Furthermore, a model which is good at many tasks is desirable, since general knowledge from one task may be applicable other tasks. Prompt engineering is effectively the process of designing inputs which elicit good performance from a pre-trained language model which has not been specifically trained at the given task. Good prompt engineering allows for good zero-shot performance at the given task.

Petroni et al. (2019) was an early work in the area that showed that large pre-trained language models can act as a relational knowledge base when given correctly worded text inputs. For example, rather than using complex NLP pipelines to extract information and store it in a knowledge base, such as the capitals of countries, one could simply ask a pre-trained LM to fill in the masked token of a text prompt like “[MASK] is the capital of France”.

Jiang et al. (2020) showed that large pre-trained language models are highly sensitive to their input text, and simply re-wording a prompt could cause a model to succeed or fail at retrieving a particular piece of information. They also propose two methods of systematically generating diverse prompts which improve knowledge retrieval by 8%. This indicates that pre-trained language models are somewhat knowledgeable, but they are sensitive to how that knowledge is queried.

Liu and Chilton (2022) apply prompt engineering to vision models. They perform a series of experiments in which the develop a set of guidelines for good prompt design specifically for image synthesis models. They found that changing the subject and style keywords had a greater impact of generated images than function words and connecting words. They also found that using style cues had a large impact on generate images, but that it was best if the subject matter matched the given style cues in abstraction.

Zhou et al. (2021) note that simply adding the article ‘a’ in front of the class label of a prompt improves VLM performances at object identification by about 5%. They write that “a slight change in wording could have a huge impact on performance [of VLMs]”. Oppenlaender (2022) established a taxonomy of types of prompt modifiers used by the text-based generative art community to encourage image-generating models to create a desired image. They found that participants in the generative art community change the characteristics of generated images by using subject terms (eg. “an old car in a meadow”), style modifiers (eg. “surrealist painting”), quality boosters (eg. “highly detailed”), repetition (eg. “space whale. a whale in space”), and magic terms (eg. “control the soul”, which produces unpredictable but visually interesting results).

SUTS for CVLMs

Lius et al. (2020) criticize synthetic datasets

Overview

Image Generation

Finding 3D models

Spawn areas

Random position/orientation

Collision detection

Big things only on the street

Camera placement

Background and sky

Screenshot, detect relations

Relation Detection

After object placement, there are two types of relations, frame and world

* + 1. Frame Position

Absolute

Relative

* + 1. World Position

Relative

Taxonomy of objects

Lacks vertical

Sentence Generation

Basic, basic 2

articles

adjectives

Verbs, adverbs

Question, command

Prompt engineering

Ungrammatical

CLIP Evaluation

CLIP is bad at SU, but good at object detection

Connection to CLMS

Knowledge about transformers, regular expressions, knowledge about interpretability,

References