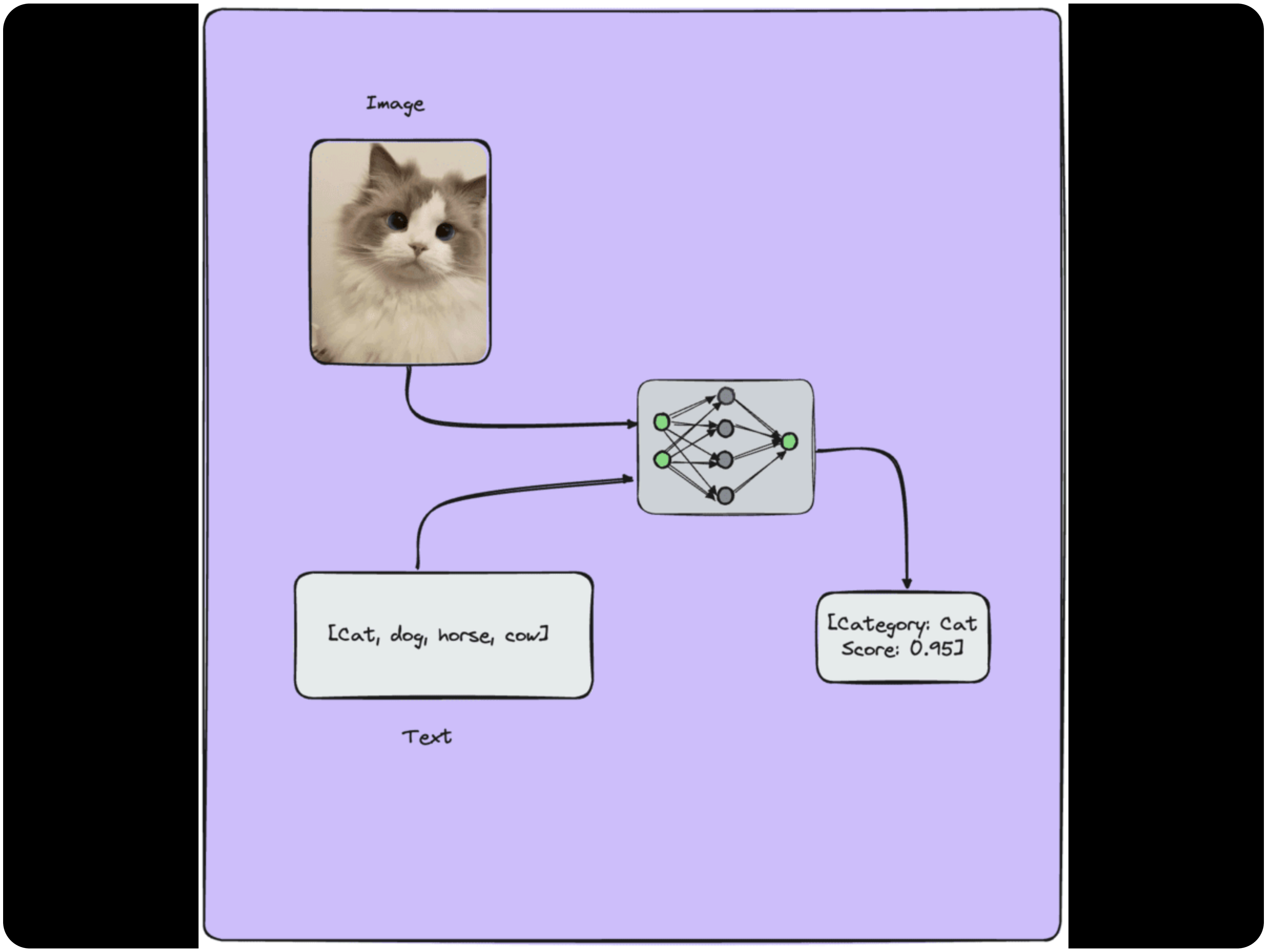
Blog / Tips & Guides / Research Labs
Research

Vision Language Models: Learning Strategies & Applications

April 2, 2024 | 7 min read

Adithya Singh



Vision language models (VLMs) are a type of AI model that combines computer vision (CV) and natural language processing (NLP) capabilities. These models are designed to understand and generate text about images, bridging the gap between visual information and natural language descriptions.

VLMs can perform a variety of tasks, including image captioning (generating descriptions for images), visual question answering (answering questions about images), and image-text matching (finding similarity between images and text descriptions). They are typically trained on large datasets that contain paired images and text annotations, allowing the model to learn to associate visual features with linguistic expressions.

One of the key challenges in developing VLMs is the integration of both visual and textual modalities in a coherent and effective manner.

Researchers use techniques such as multi-modal fusion, where visual and textual information is combined at different stages of the model architecture, to achieve this integration.

VLMs have applications in various fields, including content generation, accessibility (e.g., for visually impaired individuals), and multimodal understanding (e.g., for autonomous systems that need to understand both images and text). They represent a significant advancement in AI technology, enabling machines to better understand and interact with the visual world.

Applications of Vision Language Models

Vision language models (VLMs) have several real-world applications across different industries. Here are some examples:

1. **Image Captioning**: VLMs can generate descriptive captions for images, making them useful for applications like social media content generation, automated image tagging, and enhancing accessibility for visually impaired individuals.

2. **Visual Question Answering (VQA)**: VLMs can answer questions about images, enabling applications such as interactive educational tools, virtual assistants for image-based queries, and enhanced image search engines.

3. **Content Creation**: VLMs can be used to generate engaging content for marketing, advertising, and storytelling. They can automatically generate captions, headlines, and other textual elements for visual content.

4. **Multimodal Understanding**: VLMs can help in understanding and interpreting multimodal content, such as videos, where both visual and auditory information are present. This can be useful for applications like video summarization and content moderation.

5. **Virtual and Augmented Reality**: VLMs can enhance virtual and augmented reality experiences by providing contextually relevant information or generating interactive elements based on visual input.

6. **Healthcare**: VLMs can assist in medical image analysis, such as identifying anomalies in medical scans or assisting radiologists in diagnosis by providing relevant information based on visual inputs.

Learning Strategies

The Vision Language Models have been a topic of research for a long time now. Researchers have explored multiple learning strategies that could be used to ensure and enhance model learning. Some of these approaches are end-to-end and work towards fusing the textual and visual features, while others consider text and images as separate modalities. Let's take an in-depth look at some of the most popular learning strategies.

Contrastive Learning

Contrastive learning relies on teaching the model about differences between similar and different inputs. The idea is to present the model with inputs in pairs—similar pairs are called positive pairs and dissimilar pairs are known as negative pairs.

The model learns to extract meaningful representations of the input pair and project them into a lower dimensional space. The model tries to project the representations of the similar pair closer to each other and consequently the dissimilar representations are projected far apart.

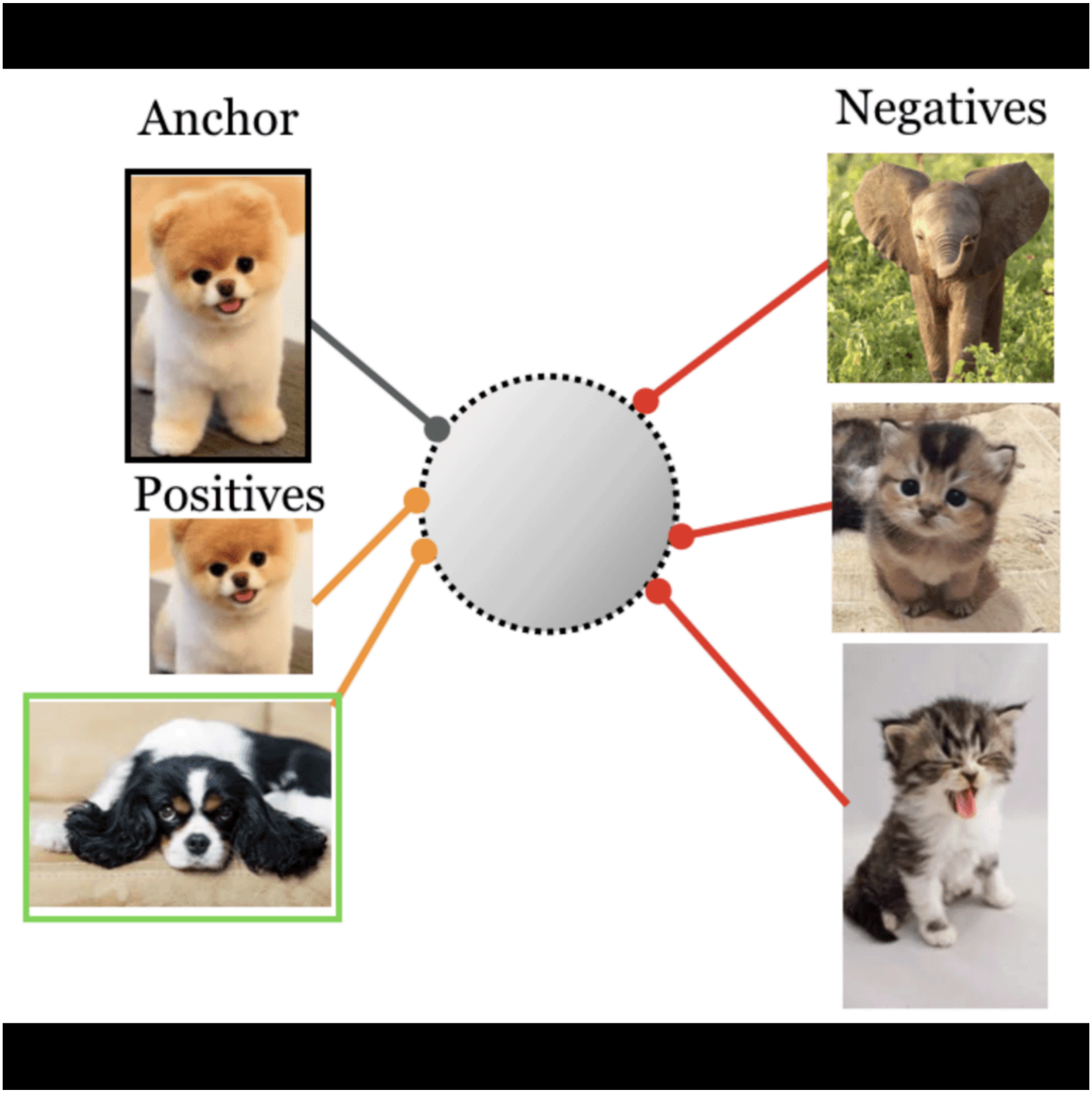


Fig 1. Contrastive learning strategy (**Source**)

The distances between the projections can be measured using manhattan and eucledian heuristics. The contrastive learning can be conducted as supervised, semi-supervised, or self supervised manner. This alleviates the need of large size annotated datasets.

PrefixLM

Prefix language modeling (PrefixLM) is a technique that uses a fixed-length prefix of a sequence of tokens (e.g. words or characters) to predict the next token in the sequence. In the context of training a vision language model, the prefix is used to provide context to the language model, so that it can generate more accurate and informative captions for images.

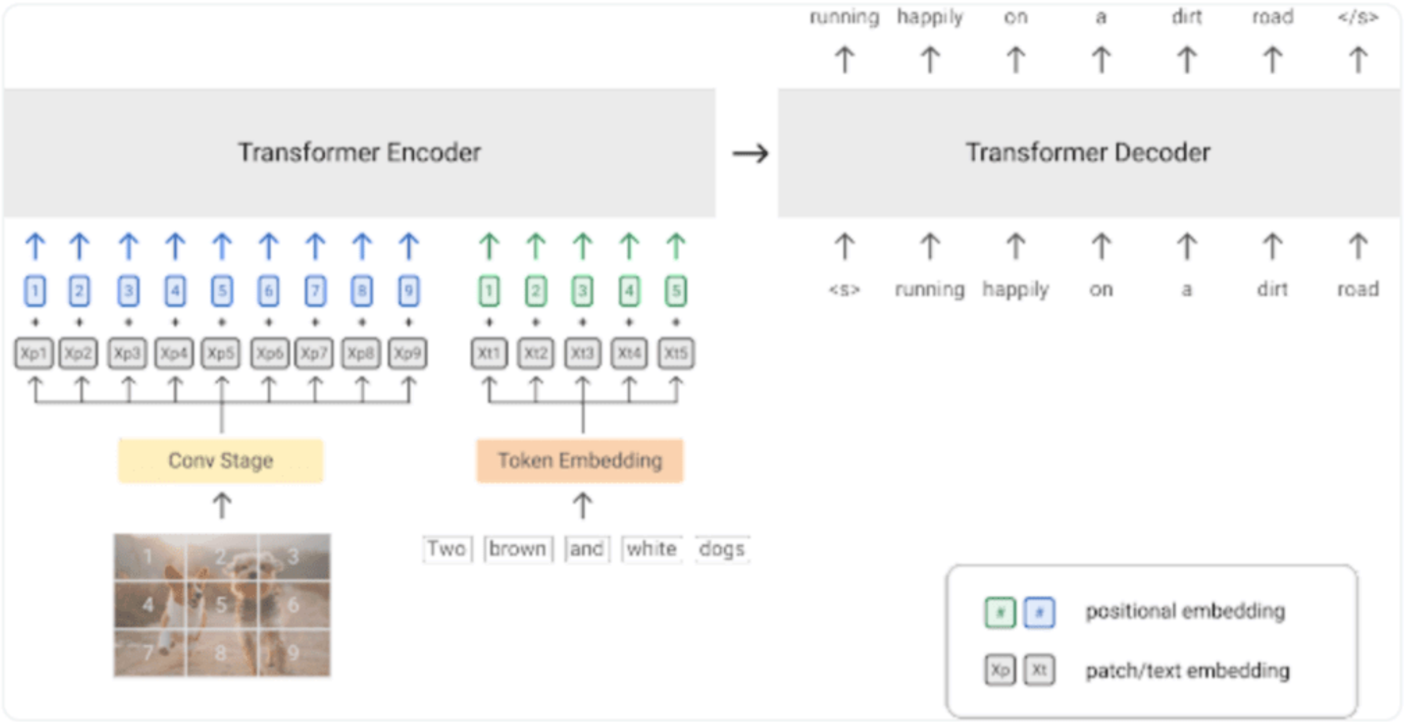


Fig 2. SimVLM architecture for prefix language modeling (**Source**)

The idea is that the prefix provides a starting point for the language model, and helps it to focus on the relevant aspects of the image when generating the caption. By using a prefix that is descriptive of the image, the language model can generate captions that are more accurate and informative, and that better capture the content and context of the image.

Multimodal Fusing with Cross Attention

Eventually, to develop a fully generalized model capable of completing the tasks that require both visual and textual information, we need a model capable of taking images and text as inputs and jointly processing both to perform tasks. The task can be off the shelf image captioning, visual question answering, object detection and semantic segmentation. The model should be able to aggregate both visual and textual features and contextualize them.

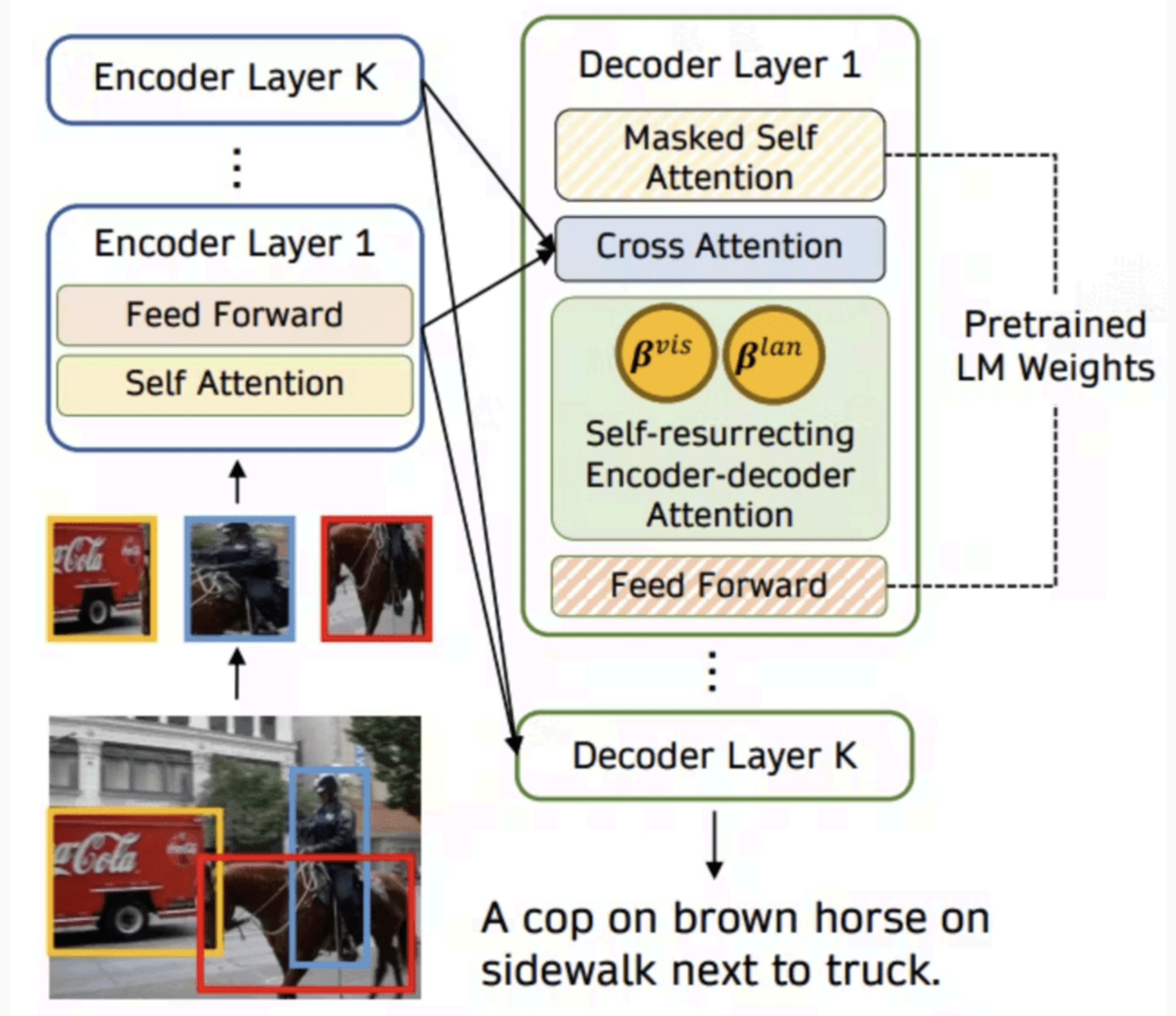


Fig 3. VisualGPT model for multimodal cross attention (**Source**)

Multimodal fusing with cross attention is a technique that allows the model to learn these relationships by fusing the visual and linguistic information in a way that takes into account the context and relevance of each modality.

Learn about image comparison algorithms

The technique works by first encoding the visual and linguistic information using separate encoders, such as a CNN for the visual information and a transformer for the linguistic information. The encoded representations are then fused using cross attention, which allows the model to learn the relevance of each modality and generate a weighted representation that takes into account the context and relevance of each modality.

Research on VLMs

The literature trends of VLMs have gained considerable traction with advent of well established large language models capable of processing large context windows. Text-to-image models have accelerated the bridging between visual and textual cues with diffusion models.

Let’s take a look at some of the most fundamental research experiments to make VLMs a reality.

CLIP

CLIP model (Contrastive Language-Image Pre-Training) works by learning a robust representation of images and text. The model is trained on a large dataset of image-text pairs, where each pair consists of an image and a corresponding text caption.

The model uses a text encoder and image encoder to convert large datasets of image-text pairs to respective feature embeddings. Then the algorithm trains the model to learn to minimize the loss between correct image-text embedding pairs, and maximizing the loss between incorrect pairs. This is pre-training phase, and once completed, the model can be used to make a zero-shot classifier over a dataset without any explicit training.

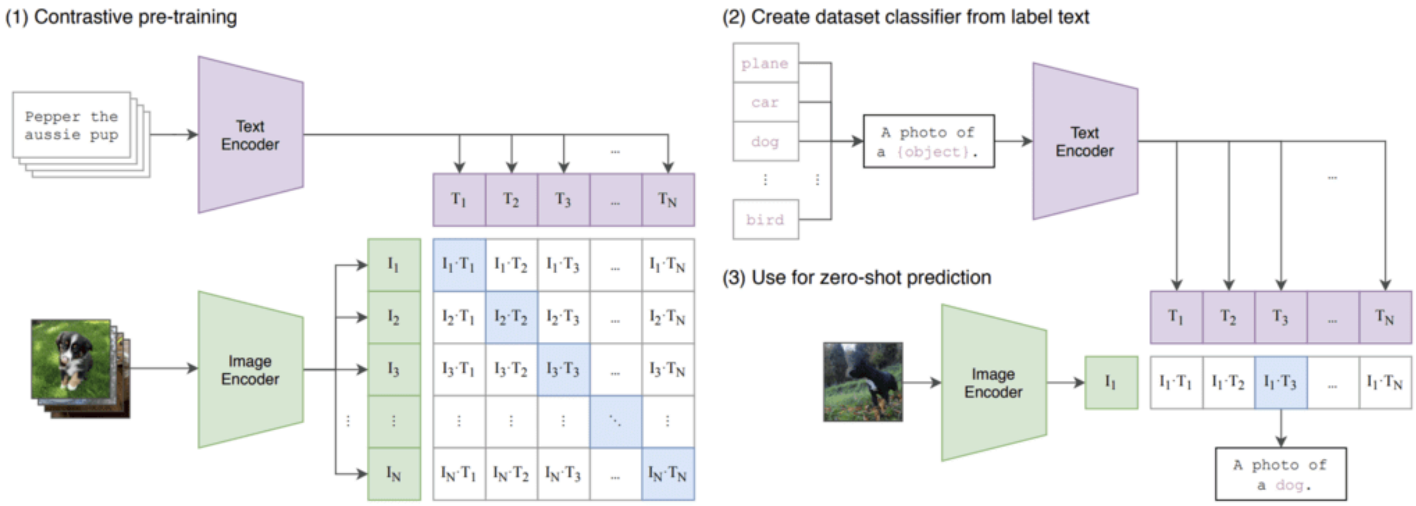

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Fig 4. CLIP model (**Source**)

In testing phase, the model utilizes the classes in the dataset to generate a rudimentary caption like “A photo of [object]” for each class label. Each caption is processed by the text encoder and its textual embedding is added to the set of embeddings. Similarly the image encoder processes the input image to generate the image embedding.

Since the CLIP is pre-trained to find the similar image-text pair embeddings, the image embedded is paired to the most appropriate embedding (the class associated to this embedding is considered to be most likely classification). Thus, the model is able to perform image captioning off-the shelf without explicitly on the training dataset.

Multi-Modality Cross Attention Network

The paper proposes a model capable of finding the right pairs of image and captions. This model is trained to find semantically similar image and text pairs. The paper intends to perform this matching by taking both visual attention and textual attention and fusing them together to perform joint training.

The key innovation of multi-modality cross-attention networks is the use of cross-attention mechanisms to learn the relationships between the different modalities. Cross-attention allows the model to selectively focus on the most relevant parts of the input data, and to generate output that takes into account the context and relevance of each modality.

In a multi-modality cross-attention network, the input data consists of both text and image data. The text data is typically represented as a sequence of words or tokens, while the image data is represented as a grid of pixels. The network consists of multiple layers, each of which performs a specific function.

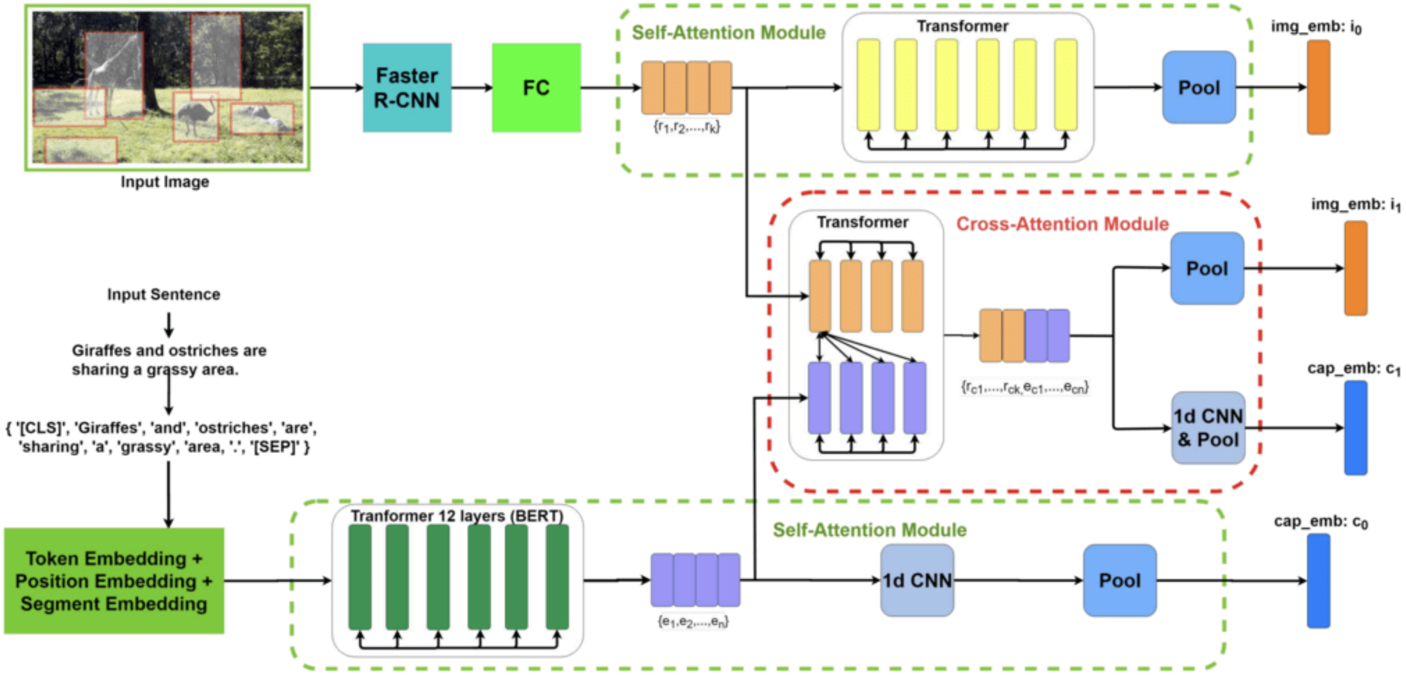

Input Image
Faster R-CNN
FC
Self-Attention Module
$\{r_1, r_2, \ldots, r_k\}$
Transformer
Pool
img_emb: i_0
Transformer
Cross-Attention Module
Pool
img_emb: i_1
$\{r_{c1}, \ldots, r_{ck}, e_{c1}, \ldots, e_{cn}\}$
1d CNN & Pool
cap_emb: c_1
Input Sentence
Giraffes and ostriches are sharing a grassy area.
{ '[CLS]', 'Giraffes', 'and', 'ostriches', 'are', 'sharing', 'a', 'grassy', 'area, '.', '[SEP]' }
Token Embedding + Position Embedding + Segment Embedding
Tranformer 12 layers (BERT)
$\{e_1, e_2, \ldots, e_n\}$
Self-Attention Module
1d CNN
Pool
cap_emb: c_0

Fig 5. Multi-modality cross attention network (**Source**)

The model has two modules, self attention module and cross attention module. Both modules take image text pair as inputs. The self attention module processes images and text separately and produces two embeddings using a self attention mechanism, image embedding, and caption embedding (i0, c0). The self attention of the image and texts are fused by the cross attention module to generate another pair of image-text embedding pair (i1, c1).

The similarity score is computed using this pair: i0 · c0 + α (i1 · c1).

Finally triplet loss is computed for the similarity score is computed. The model learns the weights to minimize the loss for the image-text pair that correspond and maximize the loss for pairs that do not correspond.

Conclusion

The research around vision language models has been gaining more and more traction. These models are capable of performing computer vision tasks using simple text instructions like image captioning, visual question-answering, object detection and image segmentation.

Learn more about transformers for computer vision.

There are multiple strategies to train a vision language model such as contrastive learning, prefix language modeling, and multi modal fusion. These models can be used in various real world applications like content creation and virtual reality.