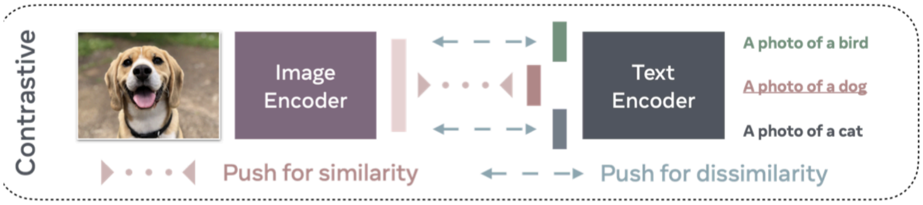
Varun Chitturi , Brandon Kleinman , Aditya Sirohi , Runyu Tian

Milestone 1

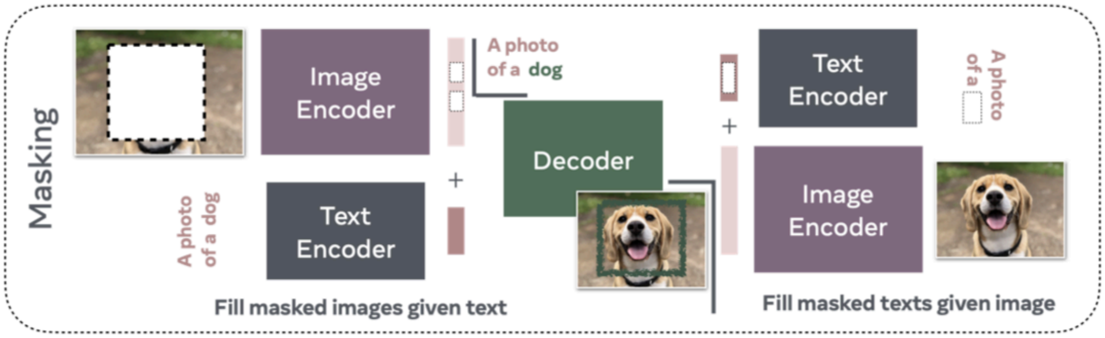
Literature Review

Paper 1: An Introduction to Vision Language Modeling

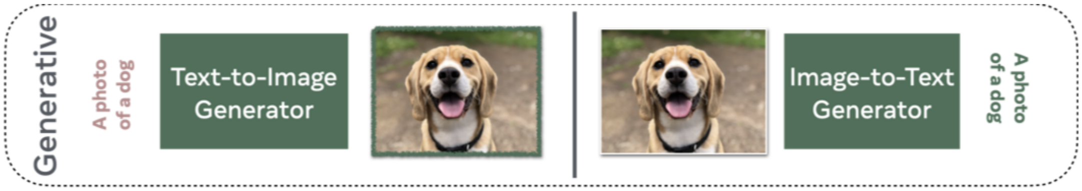
This paper describes 4 different transformer based approaches to vision language modeling. These approaches are by using contrastive training, masked training, generative training, and utilizing a pre trained backbone. These approaches are not mutually exclusive and can be used together to create a single model.



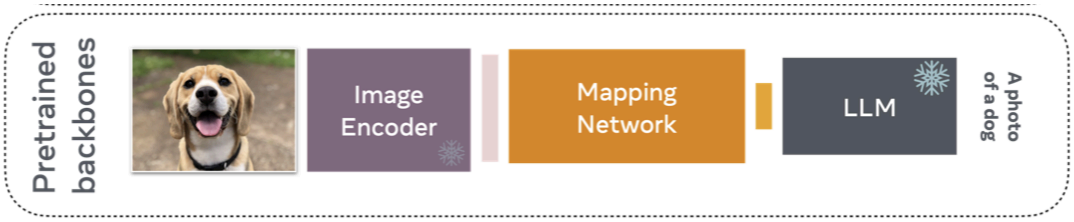
In contrastive training, we train VLM to predict how well an image text pair fits together. The VLM should assign high scores to text that matches/describes an image and low scores to text that doesn't match an image. One example of such a model is CLiP. During training, it uses image and text encoders to obtain contextual embeddings for an image/text pair. It then trains the text and image encoders to maximize the cosine similarity between all correct image-caption embedding pairs and minimize the similarity scores between incompatible pairs. One drawback of CLiP is that it requires huge batches, on the order of 10k, of image-text pairs to train. This would prove intractable for small scale hardware.



In masked training, the model tries to reconstruct images with masked patches given some unmasked text. Another approach is to mask text captions and try to reconstruct the caption given the unmasked image. One example of a model that uses masked training is FLAVA. FLAVA consists of 3 transformer based components. The first two of which are an image encoder (ViT) and a text encoder (text transformer) that are trained using masking approaches. The third component is a multimodal encoder that combines the outputs of the image and text encoder with linear projections and attention mechanisms. The multimodal encoder is trained using both multimodal and unimodal masked modeling losses along with a contrastive objective. FLAVA achieves SOTA performance across 35 different multimodal tasks and benchmarks.



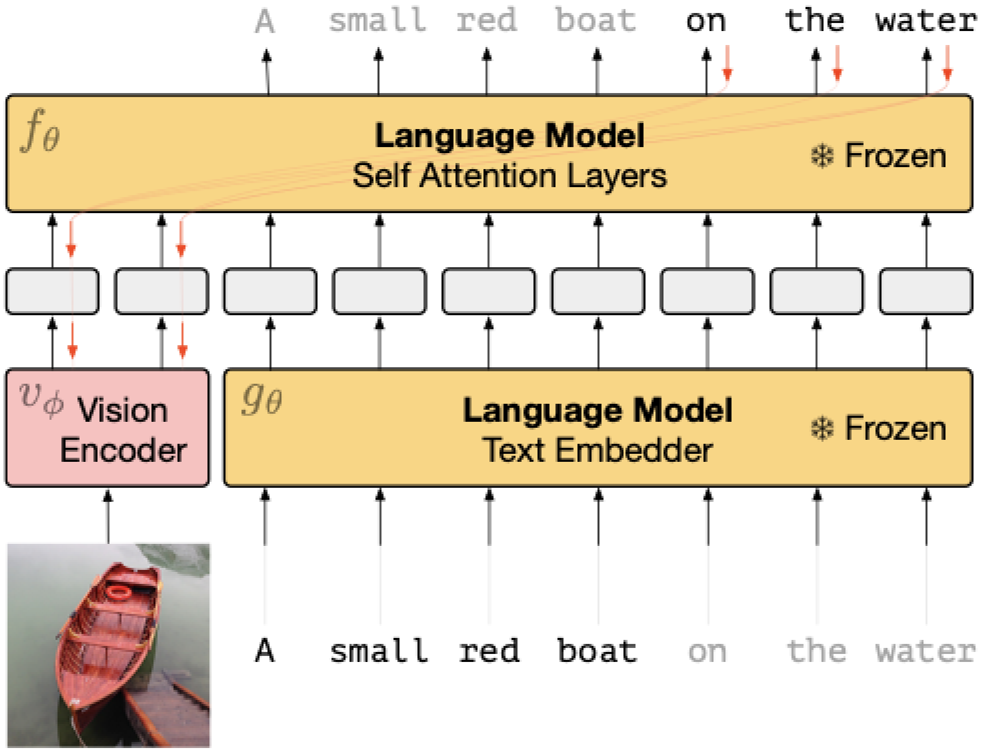
In generative training, the models are trained to explicitly generate text and/or images. CoCa (Contastive Captioner) is a text generation model that is trained using both a contrastive loss (like in CLiP) and a generative loss. CoCa uses an image encoder and text encoder to encode an image-text pair to image-text vectors. CoCa then uses a multimodal decoder to decode an image-text embedding to generate a caption. The generated caption is then evaluated using a generative loss.



The last approach is one in which we use a pre-trained LLM and essentially adapt it into a VLM. By leveraging pretrained LLMs it is possible to learn vision tasks easier, with less computational overhead. Specifically, we only need to learn a mapping between text modality and images and let the LLM take care of the rest. Frozen is a model that leverages a pre-trained LLM. It uses a lightweight vision encoder that maps visual features of an image into text embeddings. For the vision encoder, Frozen used a NF-ResNet-50 and a linear mapping which is trained from scratch. To train the model, a simple text generation objective was used on the Conceptual Captions dataset. Frozen achieves modest performance but its strength is that it requires relatively little training and has high zero/few shot performance and can adapt to new tasks very quickly.

Paper 2: Multimodal Few-Shot Learning with Frozen Language Models

The paper presents a new approach of training a new multimodal model extending the principles of auto-regressive transformers to images. Frozen, this new approach, takes a pre-trained language model and an image encoder to get few-shot learning for new tasks. The principle of Frozen is to keep the weights of the self attention layers and language model "frozen" but back propagate the gradients to the image encoder during training. It can perform zero-shot performance on tasks that it was not trained on such as vision question answering (VQA).



The model uses a pre-trained deep auto-regressive language model and adds a visual encoder (e.g., NF-ResNet-50 in the paper) that outputs a continuous sequence to the transformer in an embedding format with the same dimensionality as the token embedding (visual prefix). The paper stated that the best number of tokens for the visual prefix is 2.

During training, the vision encoder parameters are updated using image captioning data. The following equation is used to maximize the log likelihood of the image encoder:

$$\begin{aligned}\log p_{\theta,\phi}(\mathbf{y}|x) &= \sum_{l} \log p_{\theta,\phi}(y_l|\mathbf{x}, y_1, y_2, \ldots, y_{l-1}) \\ &= \sum_{l} f_{\theta}(i_1, i_2, \ldots, i_n, t_1, t_2, \ldots, t_{l-1})_{y_l}\end{aligned}$$

Frozen has the ability to access encyclopedic knowledge about an image and uses the outputs of the image model and interleaves them with the language model to determine what is presented in the image. Additionally it has the ability to perform VQU with few-shot learning and the performance of the model improves as more samples from each class are provided. The model can take trained classes and then use new sample inputs and knowledge in the language model to answer questions about a new, unseen image.

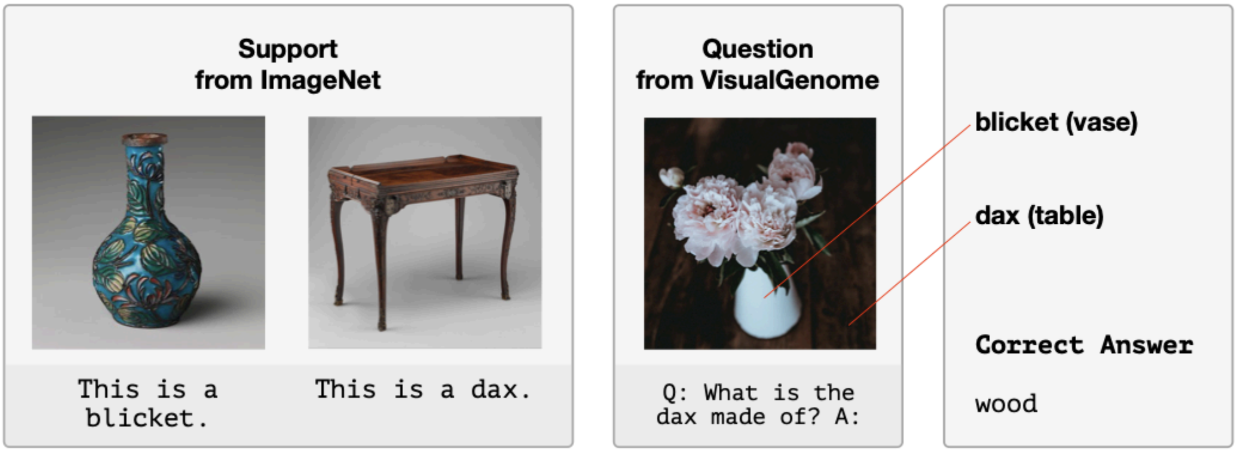
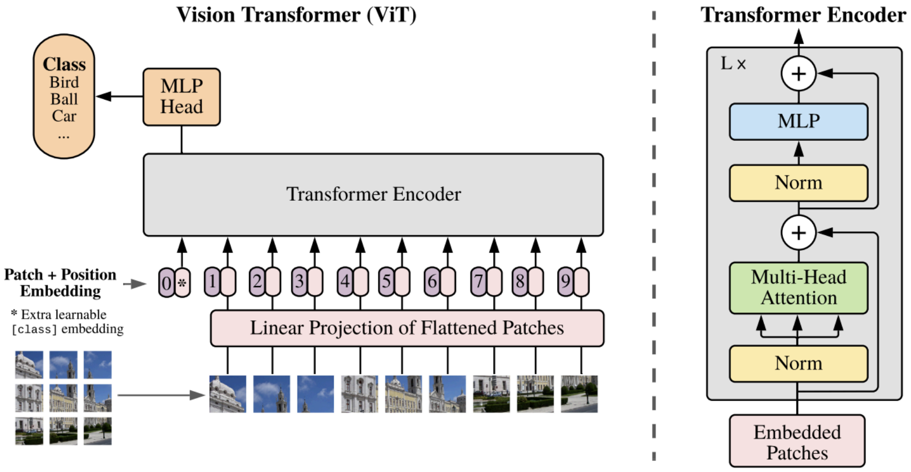


Figure 7: Example of a Fast-VQA task.

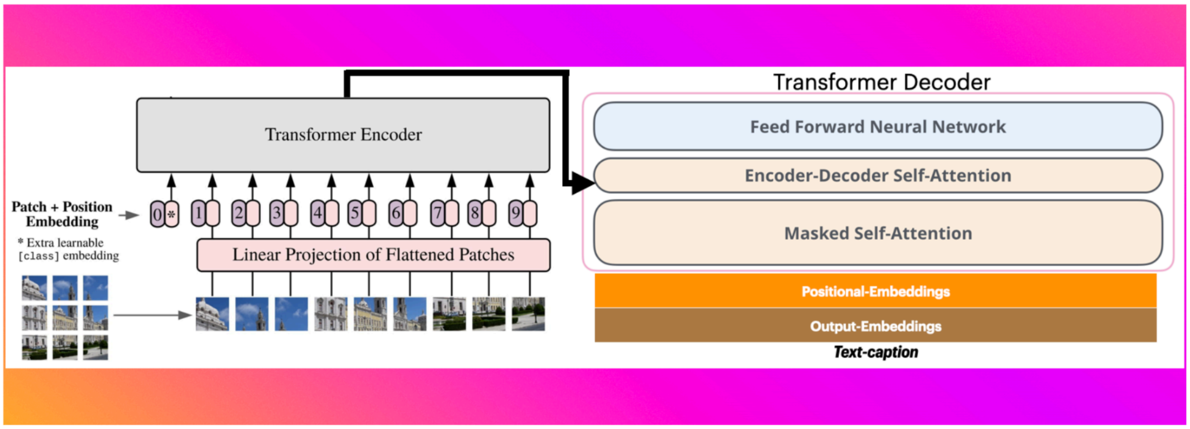
Paper 3: An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale

- This paper introduces Vision Transformer (ViT), that brings transformer architecture that's popular in NLP to Computer Vision (image recognition/classification). The model works by dividing images into fixed-sized patches, which are then processed with the self-attention mechanism. Compared with the traditional CNN model, ViT is able to achieve great performance while requiring substantially fewer computational resources for training.

- Extending ViT to image captioning relies on its mechanism to represent visual images as token sequences, which is very similar to word embeddings in NLP tasks. By combining ViT with language models (eg: transformers decoder / GPT), we could generate descriptive captioning.



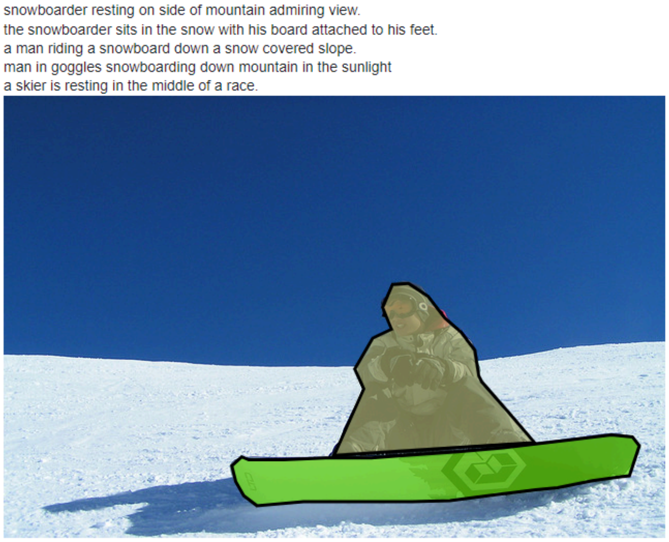
- Overall, in this setup, ViT provides an encoding of images' spatial and semitic information, which is then fed to the language model for contextually relevant captions.
- Example usage from HuggingFace community:


Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption

Dataset

https://cocodataset.org/#home

The COCO (Common objects in context) dataset is a large scale object detection, segmentation, and captioning dataset. It includes 330,000 images of which over 200,000 are labeled, has 1.5 million object instances, 80 different object categories, 91 “stuff” categories, 5 captions per labeled image, and 250,000 people with keypoints. This dataset is particularly valuable for Vision-Language Models (VLMs) as it pairs images with human-written captions, enabling models to learn associations between visual content and language. For this project, using COCO supports our objective to assess and refine the model’s performance in image captioning, ensuring that the VLM can effectively generate fluent descriptions and answer questions about real world scenes. Example datapoint:



snowboarder resting on side of mountain admiring view.
the snowboarder sits in the snow with his board attached to his feet.
a man riding a snowboard down a snow covered slope.
man in goggles snowboarding down mountain in the sunlight
a skier is resting in the middle of a race.

(Just as a side note, computers generally learn much more effectively and require far fewer data points when captions are super detailed. Five short, one-line captions might perform significantly worse than fewer data points with multi-paragraph descriptions detailing the image content. May be interesting to see the difference between)