SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA ENGINEERING AND TECHNOLOGY

A Comprehensive Survey of Deep Learning techniques for Bidirectional Image Context Mapping

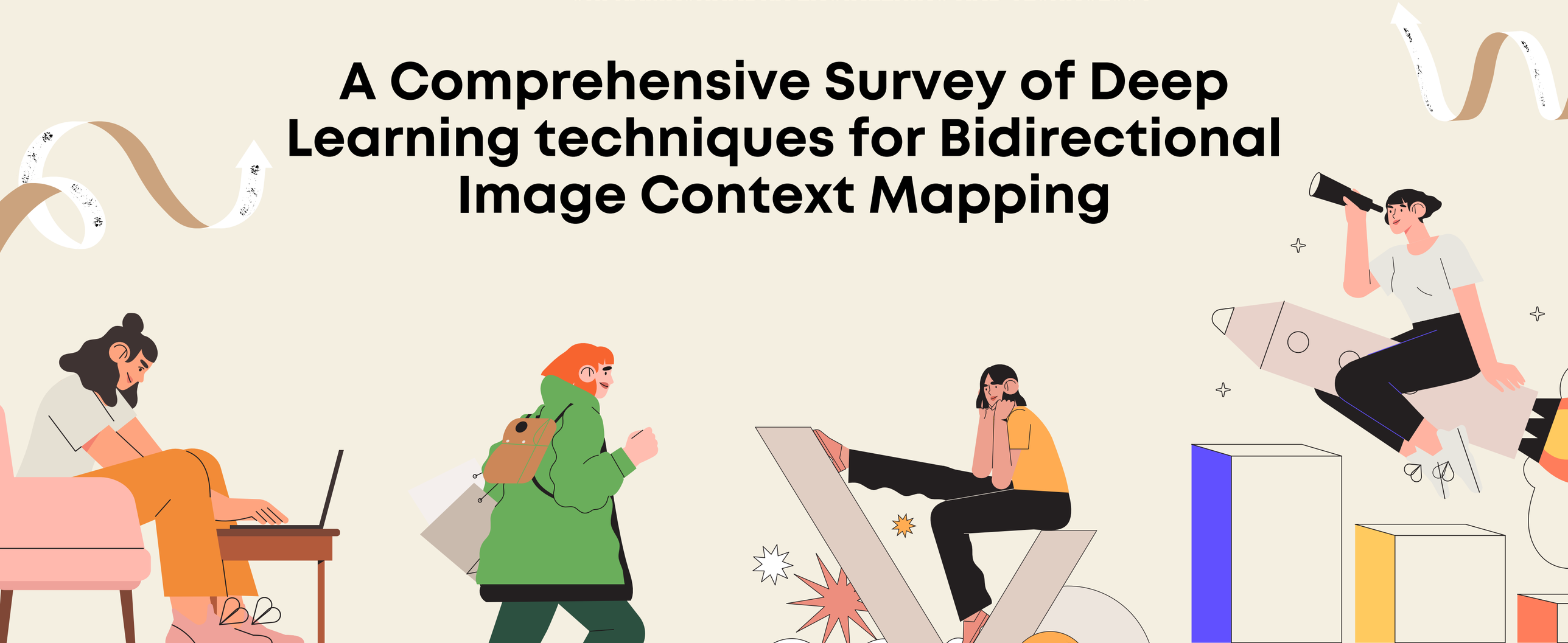


Table of Content

Our Team

Karan V

E0119039

SathishKumar M

E0119052

Vishal L

E0119010

Introduction

Image Processing

There are several classification, regression and generative models

Language Processing

There are several Text generation, Classification, Summarization and QA generation

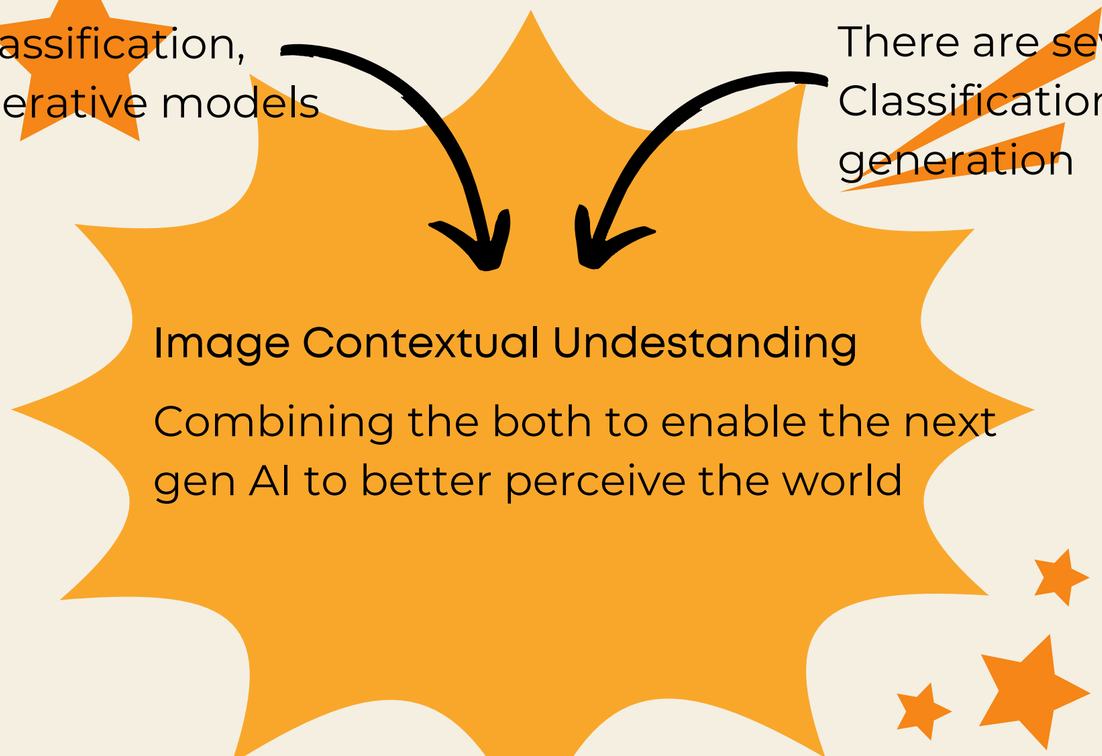


Image Contextual Undestanding

Combining the both to enable the next gen AI to better perceive the world

Problem Statement

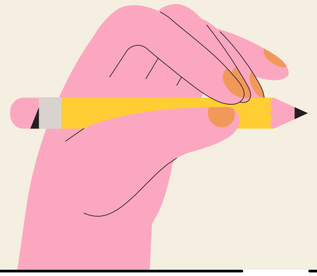
Contextual Understanding

Contextual data predictions in image video and audio have been difficult due to complexity of human language, with various phonetics and meaning of the same words producing various context

Bidirectional Retrieval

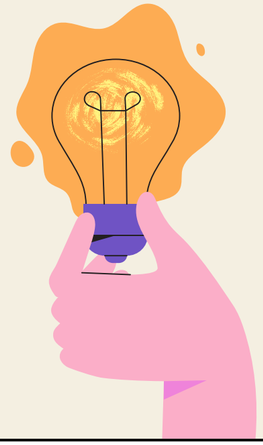
All the existing models try to perform only one way either Text to image or Image to Text

What We Did



VAE

VAE is a deep unsupervised generative approach for genrating latent distribution



Novel Model

Friendly and professional customer service specialist with extensive experience

ViT & Beam Search

Visual Transformers to create a unified model to create accurate context mapping

Literature Survey

Publication : International Journal of Applied Engineering Research - RI Publication

Title : Image Captioning - A Deep Learning Approach(2019)

Overview :

The model performs image captioning by basic encoder decoder with a image feature extraction layer by VGG 16 weights. Evaluated the algorithm using a dataset of 500 images from Flickr8k Dataset and achieves 0.683 Bleu Score

Limitations :

- Test Images are very less so Bleu score is not reliable.
- Computationally expensive
- Attention mechanism is not added to the model
- Feature extraction is only tried using VGG 16
- LSTM memory units and total params of the model are very less

Literature Survey

Publication : IEEE Access

Title : EyeSee: Camera to Caption with Attention Mechanism(2020)

Overview :

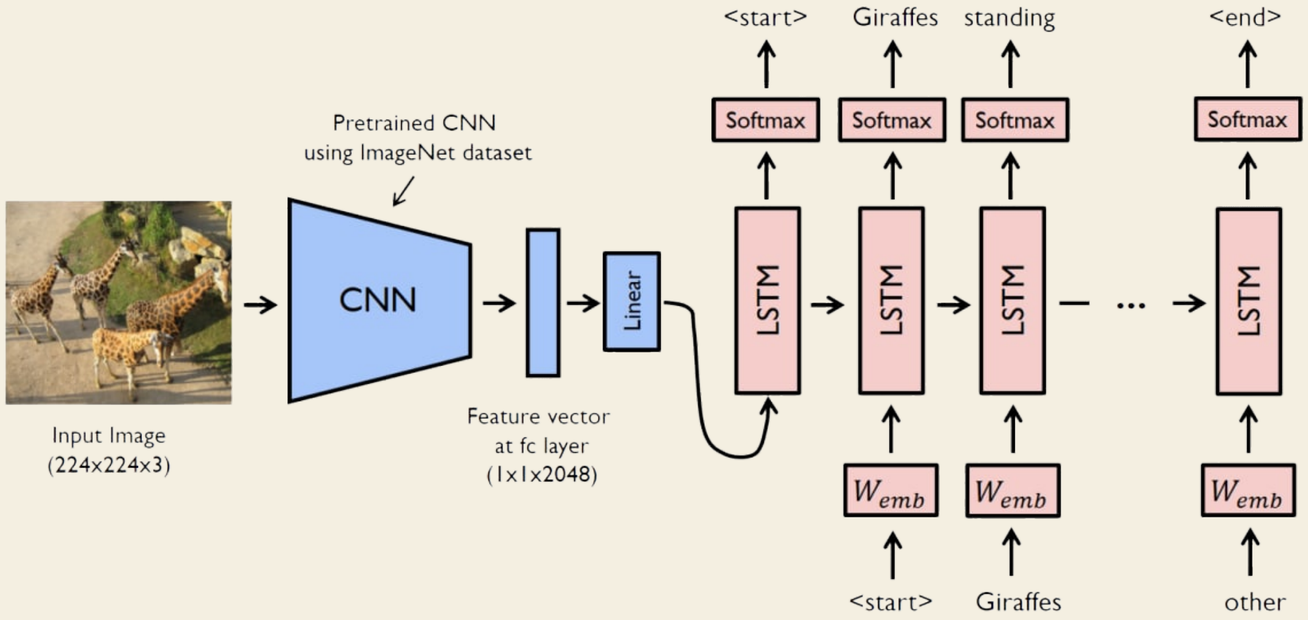
The model performs image captioning by Show, Attend and Tell research paper model and encoder includes the pretrained CNN, InceptionNet, along with a fully connected layer which is made up of a dense layer and a ReLU layer. The decoder, it consists of two cascading elements being the attention function and the RNN(GRU cells)

Limitations :

- Computationally expensive as they are proposing for a mobile app
- Small Pretrained models should be used
- Model is domain specific as it is trained on flickr8k and flickr30k

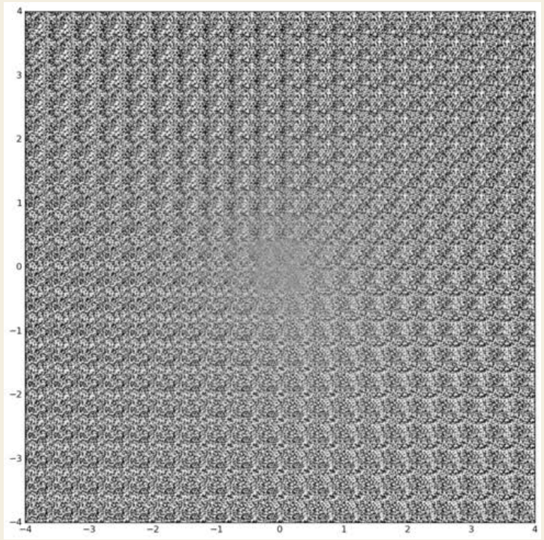
Encoder Decoder

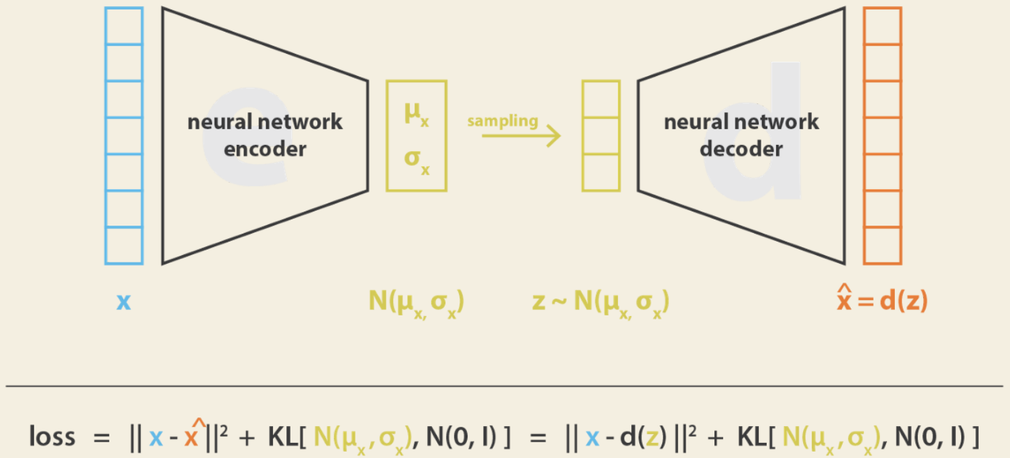
- The encoded vector aims to encapsulate the information for all input elements in order to help the decoder make accurate predictions.
- Each recurrent unit accepts a hidden state from the previous unit and produces and output as well as its own hidden state



Variational Auto Encoder

- The encoder downsamples the images and the bottle neck layer is responsible for sampling the latent distribution vector
- The Decoder upsamples the sampled latent vector to diffrent representation of the input.

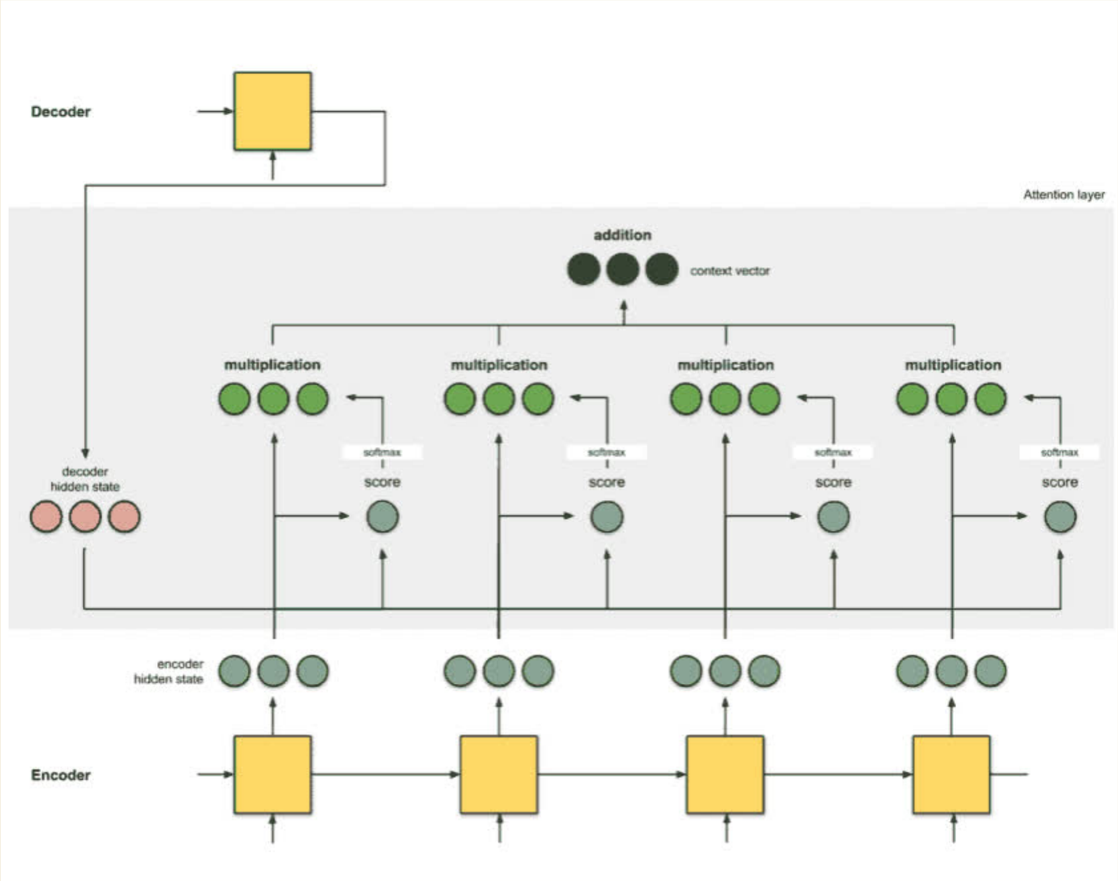




$$\text{loss} = \|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$

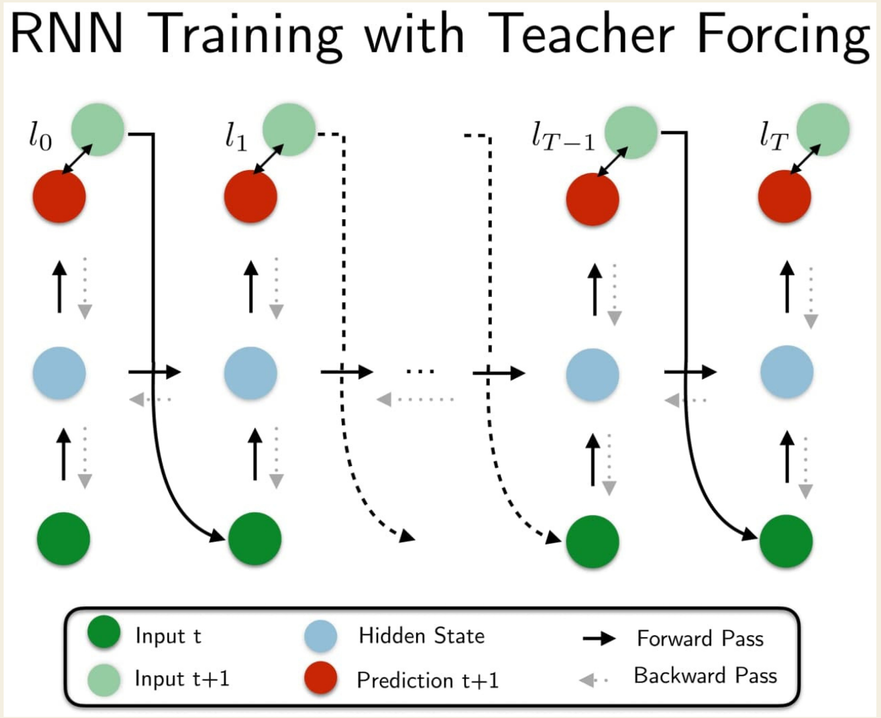
Attention-Encoder Decoder

- This effect enhances some parts of the input data while diminishing other parts.
- It assigns a probablity distributed weights to each part of the enocder at each time steps.



Teacher Forcing Mechanism

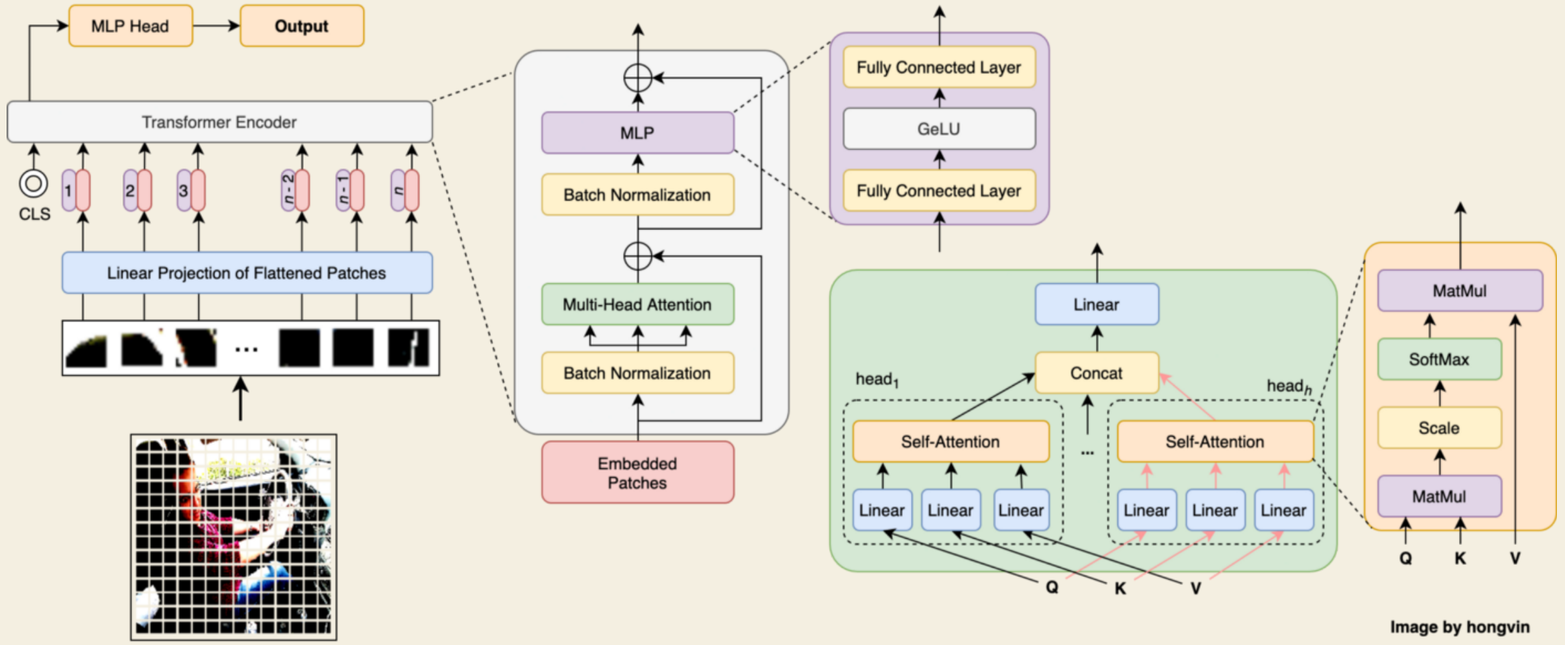
Teacher forcing is a strategy for training RNN that uses ground truth as input, instead of model output from a prior time step as an input. Models that have recurrent connections from their outputs leading back into the model may be trained with teacher forcing.

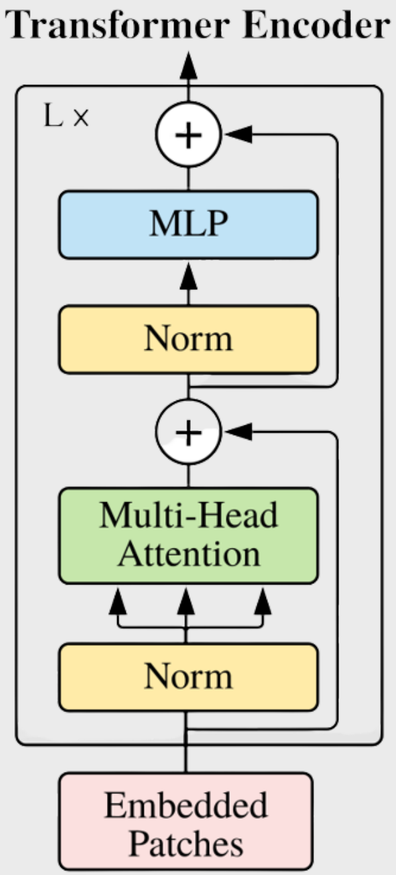


Vision Transformers

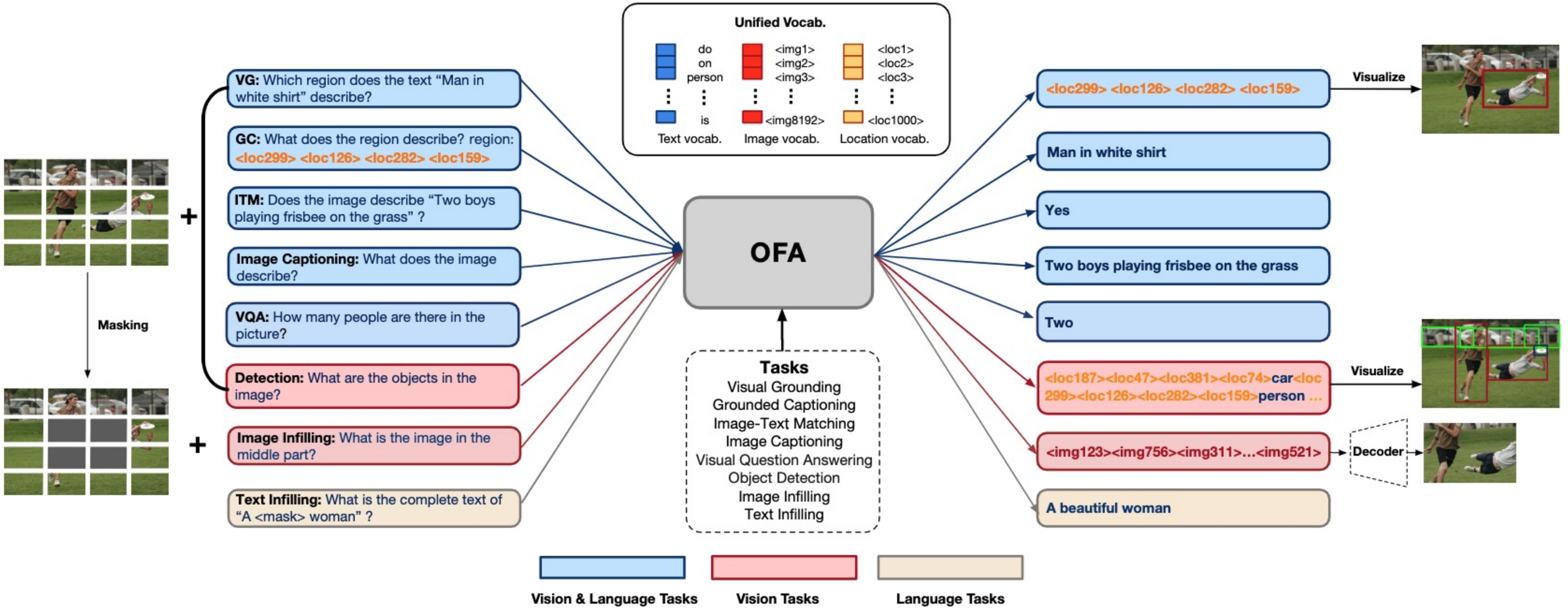
Overview

The Vision Transformer, or ViT, is a model for image classification that employs a Transformer-like architecture over patches of the image. An image is split into fixed-size patches, each of them are then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder.




Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

OFA-Variation



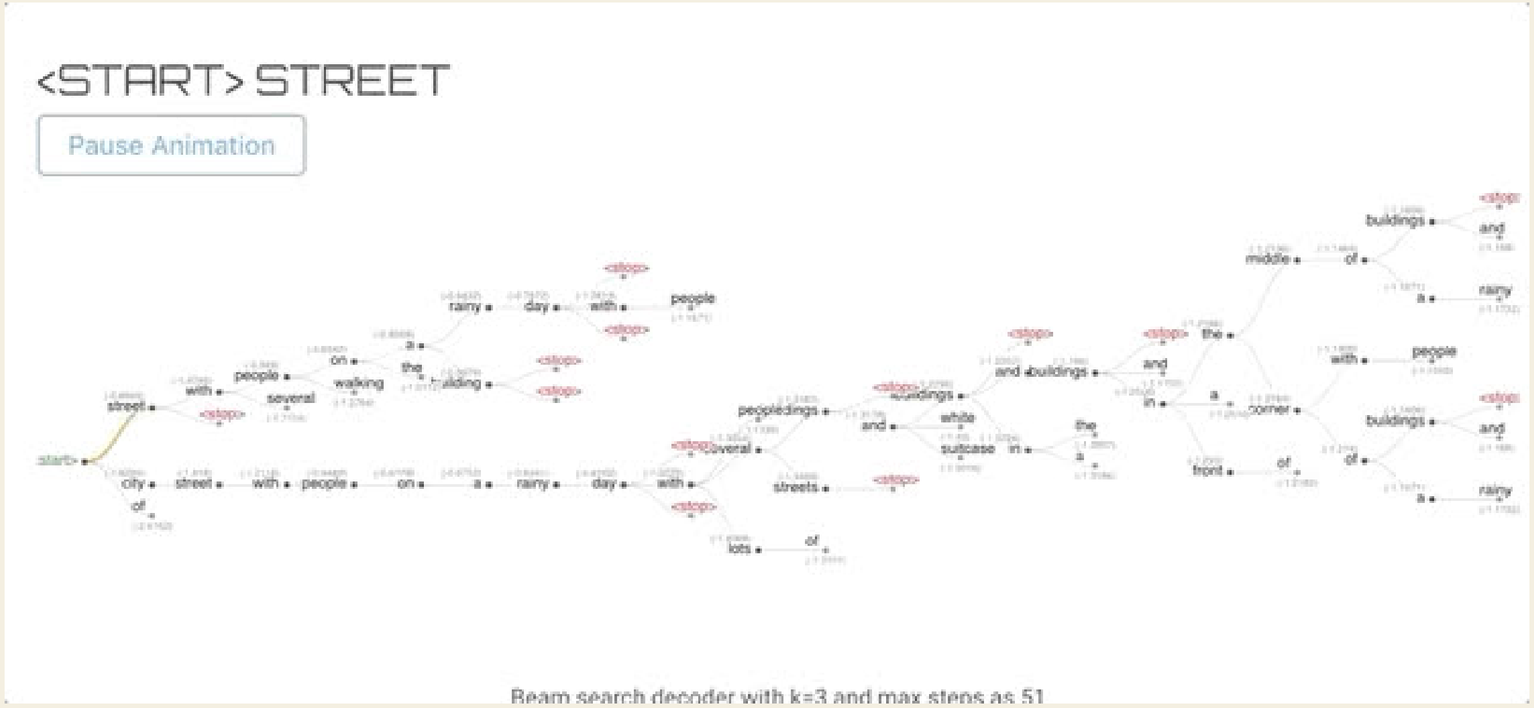
OFA-Variation

Image-Text Retrieval: “The man in blue shirt is wearing glasses.”

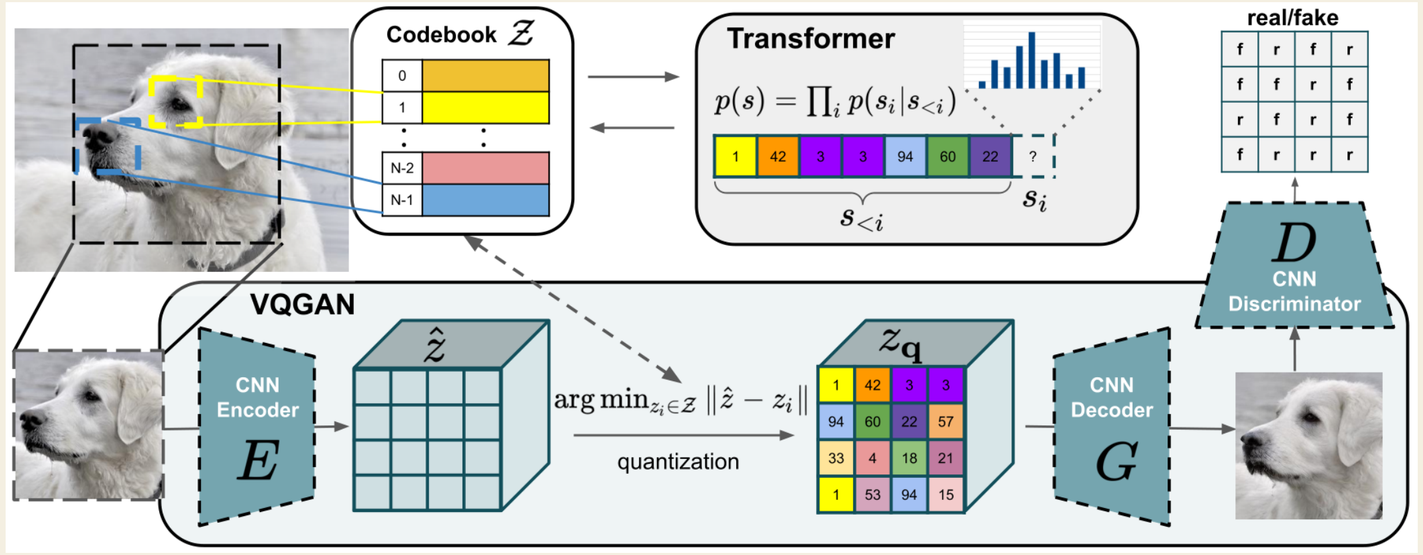
Beam Search

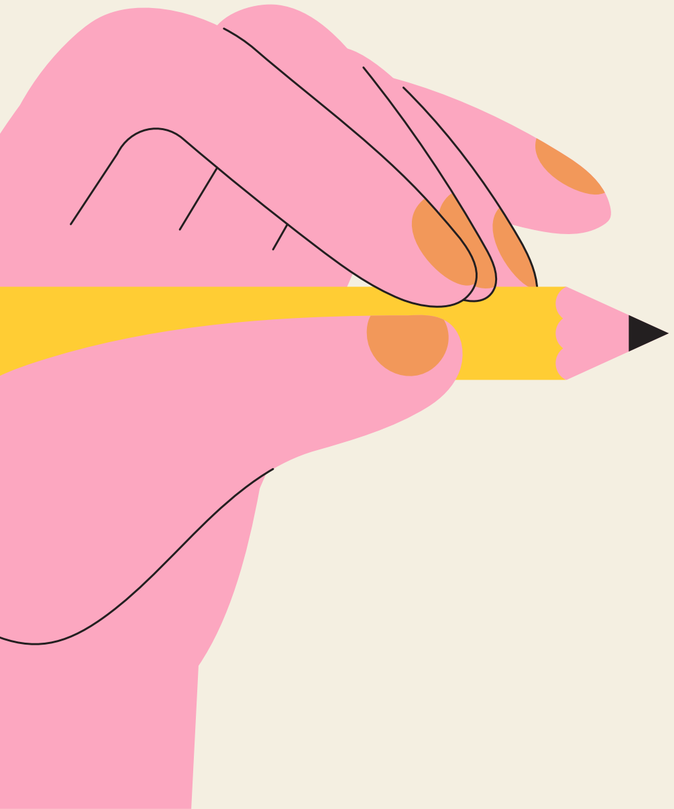
Overview

The beam search algorithm selects multiple tokens for a position in a given sequence based on conditional probability.



VQGAN+CLIP

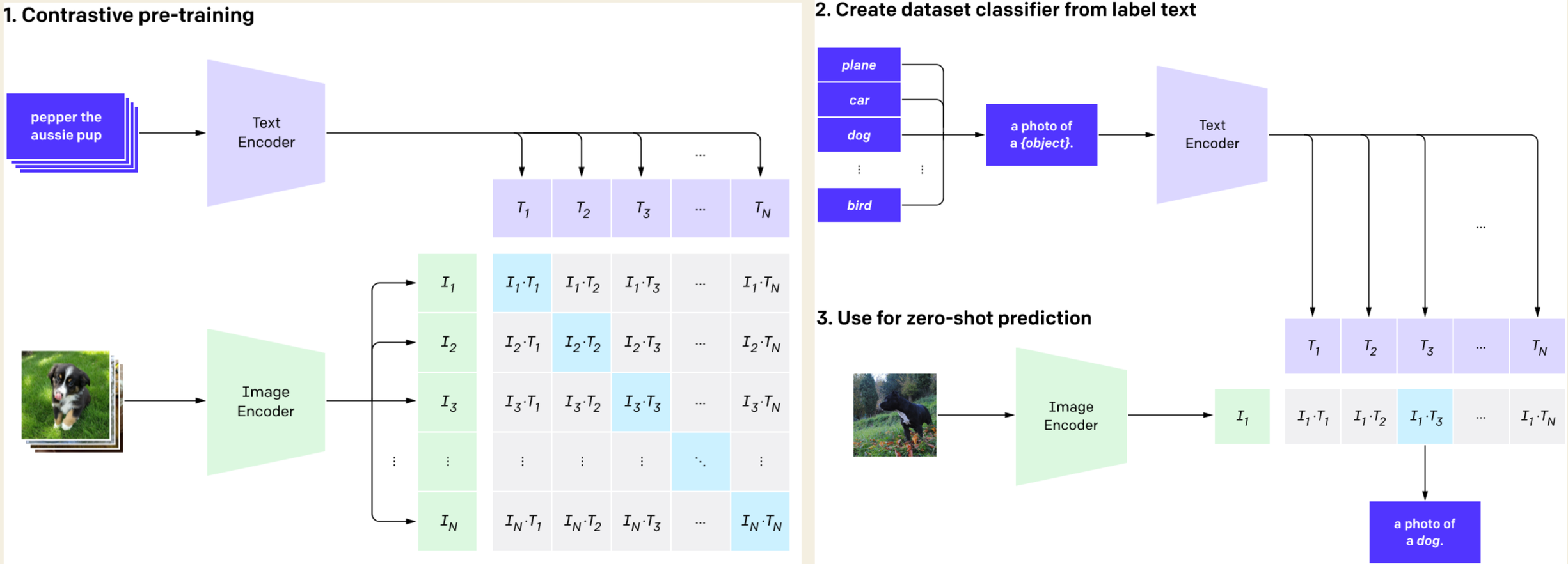




Combining the effectiveness of the inductive bias of CNNs with the expressivity of transformers enables them to model and thereby synthesize high-resolution images.

CLIP

- CLIP (Contrastive Language–Image Pre-training) is a companion third neural network which finds images based on natural language descriptions, which are what's initially fed into the VQGAN.
- CLIP is not a generative model. CLIP is "just" trained to represent both text and images very well .

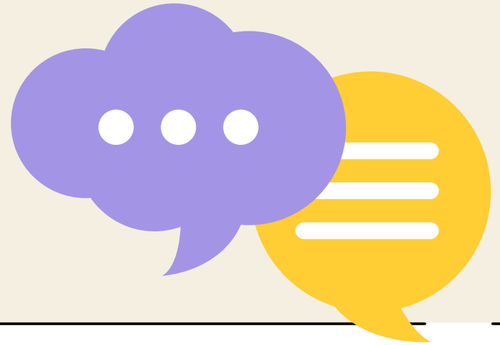


Comparing The models

Model	Bleu-1	Year
Novel Model	0.49	2022
VAE	0.32	2013
Attention based model	0.56	2014
ViT+GPT2	0.2636*	2018
OFA-Variation	0.435*	2022

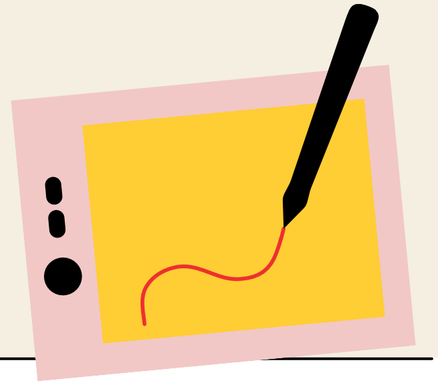
*The Bleu-1 score depends upon the testing data size which is too high for marked results

Future Endeavor



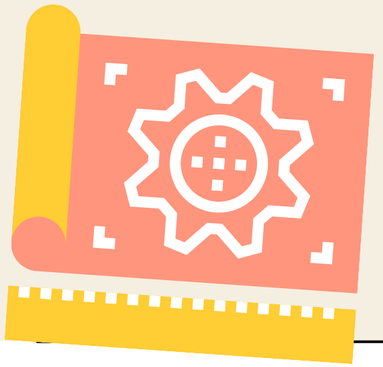
Visual Question Answering

Semantic task that aims to answer questions based on an image.



One Shot VGAN

Generating high resolution images from scratch without re-training



Visual Grounding

Locate the most relevant object or region in an image, based on a natural language query

CONCLUSION

THE GOOD

From scratch model understands the context of the scene with at most precision

THE BAD

Although contextual understanding is good, the captions produced by the model are very bad due to lack of grammar modelling

THE BETTER

The OFA Variation performs exceptionally with very good captioning but the performance of the model with respect to hardware is intensive

References

1. T. P. Ramsewak, D. Appadoo and Z. Mungloo-Dilmohamud, "EyeSee: Camera to Caption with Attention Mechanism," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411571.
2. Lakshminarasimhan Srinivasan , Dinesh Sreekanthan , Amutha A.L,"Image Captioning - A Deep Learning Approach", 2019 International Journal of Applied Engineering Research