VQA

VISUAL QUESTION ANSWERING

421230 - Nitish Kumar
421274 - Vivek Ranjan
421247 - Rushikesh Amol Pandge

Under the Guidance of **MR. M.G. Karthekeyan sir**

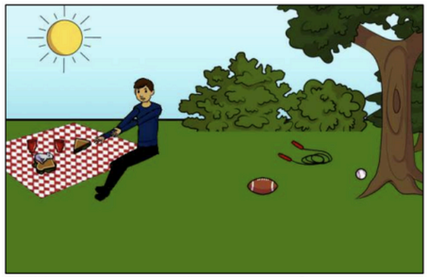
MINI-PROJECT 2024 NATIONAL INSTITUTE OF TECHNOLOGY, ANDHRA PRADESH

INTRODUCTION

- Being able to look at an image and ask questions to our computers can prove to be useful to people in numerous ways.
- A visually impaired person can move about more independently by simply asking questions about his neighborhood, whereabouts etc.
- Another major practical implication of VQA is human-computer interaction in order to get visual content - a person in a foreign country can explore the place by just enquiring what his eyes can see.

What color are her eyes?
What is the mustache made of?

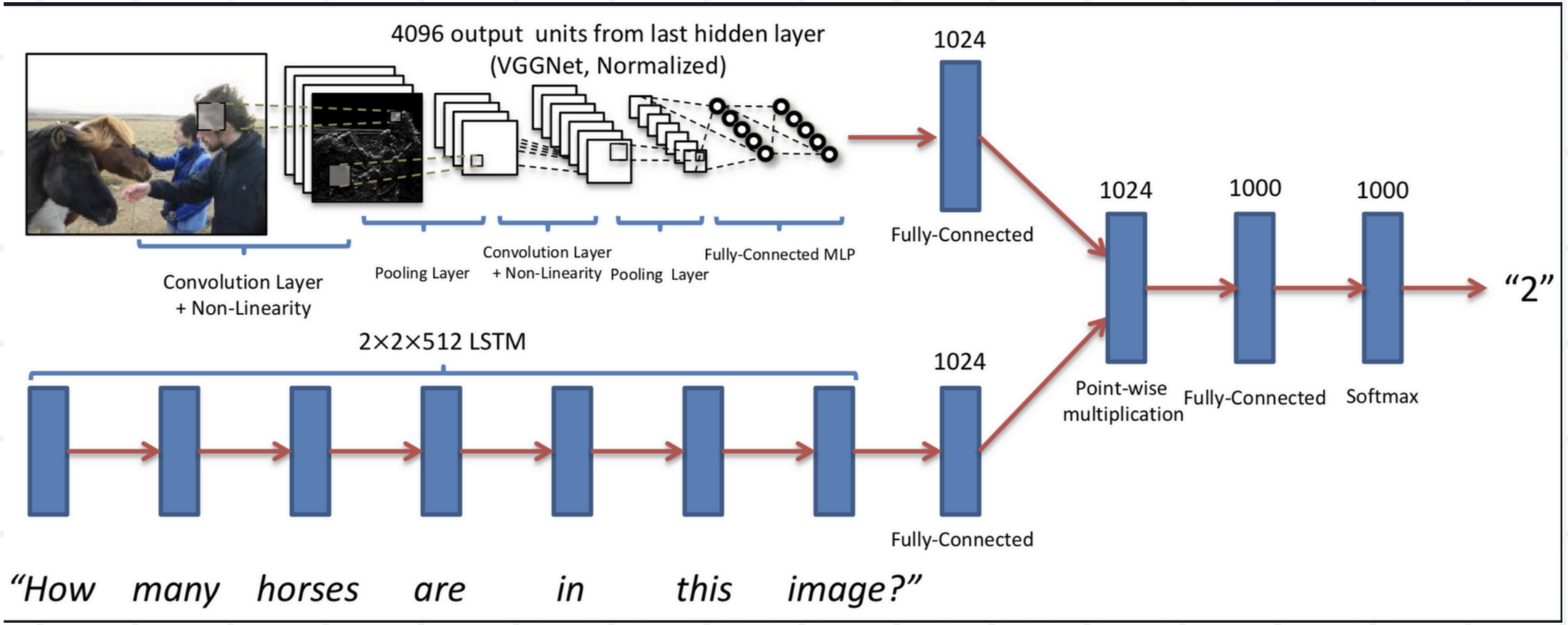
How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

INITIAL MODEL



- **Convolutional layers:** These layers are used to extract features from the image. The text mentions a convolution layer followed by a non-linearity (likely a ReLU activation function) and a pooling layer.

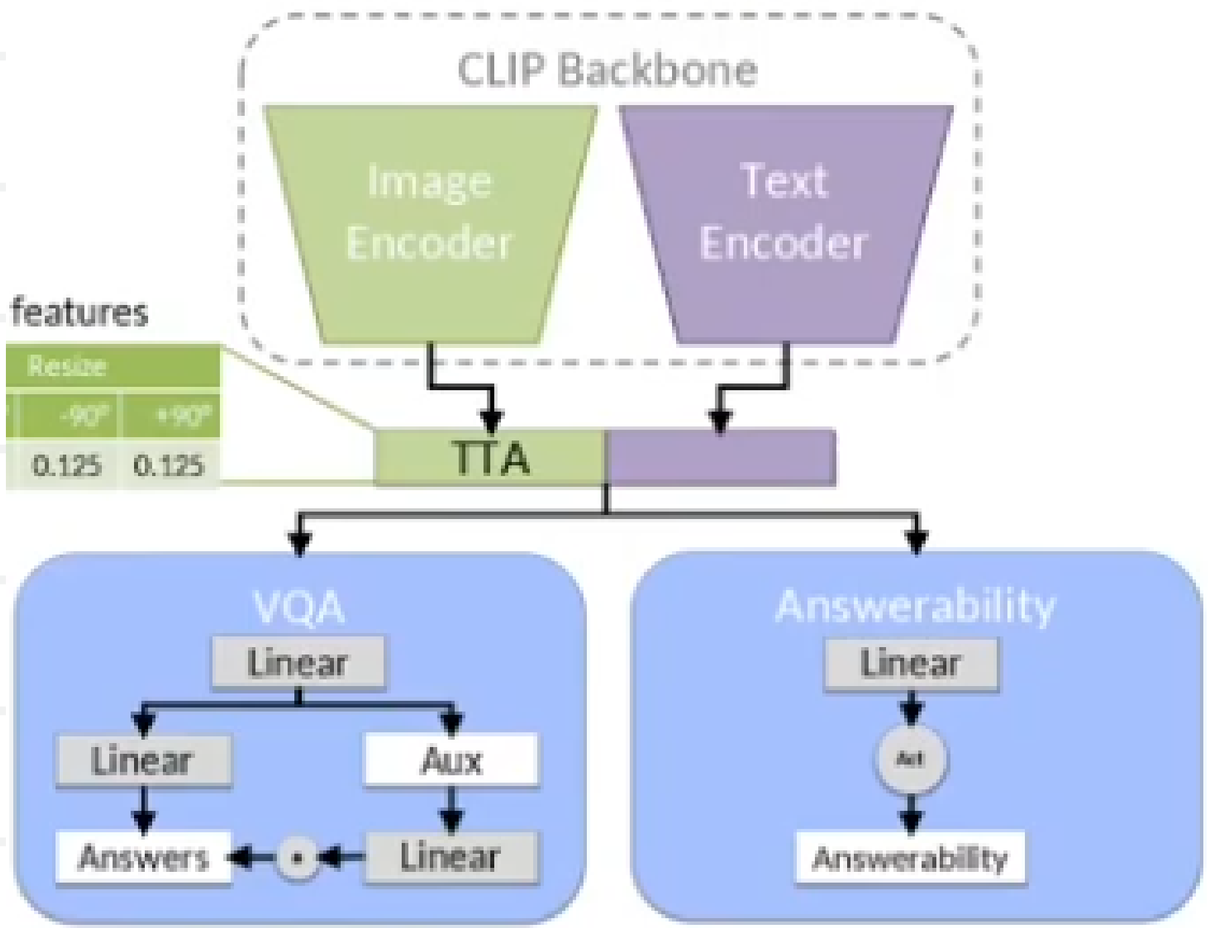
- **Long short-term memory (LSTM)** : This is a type of recurrent neural network (RNN) that is capable of learning long-term dependencies in data. In image we used a 2x2x512 LSTM layer.

- **Fully-connected layers:** These layers are used to classify the image. The text mentions two fully-connected layers, one with 1024 units and another with 1000 units, followed by a softmax layer.

LESS IS MORE

LINEAR LAYERS ON CLIP FEATURES AS POWERFUL VIZWIZ MODEL

MODEL



- Each Linear layer consist of Layer Norm followed by Drop out with probability of 0.5 followed by fully connected layer of size 512.
- Cross Entropy Loss for answer and answer type
- Binary Cross Entropy Loss for answerability

DATASET

The VizWiz dataset consists of images taken by **blind or visually impaired individuals** using smartphones, accompanied by text questions seeking assistance. **Sighted workers provide textual annotations** in response, aiding users in understanding the content of the images. This collaborative collection process provides valuable data for research in computer vision and accessibility, offering insights into real-world challenges faced by visually impaired individuals.

IMAGE ENCODER

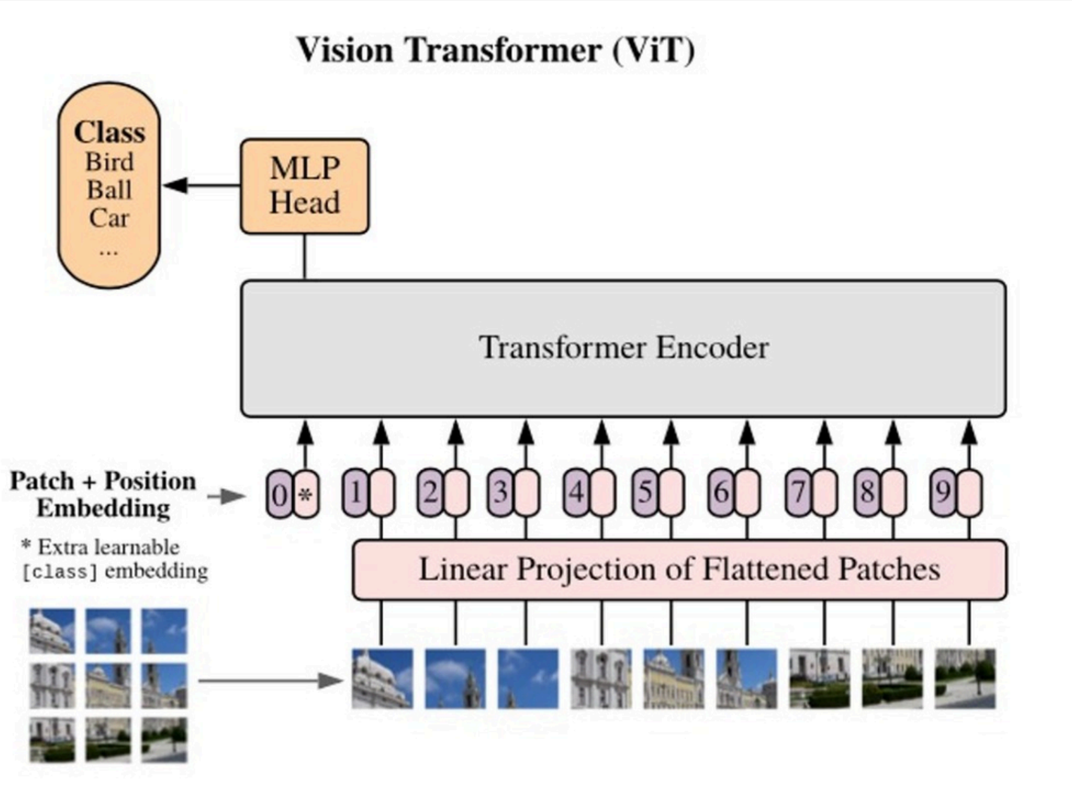
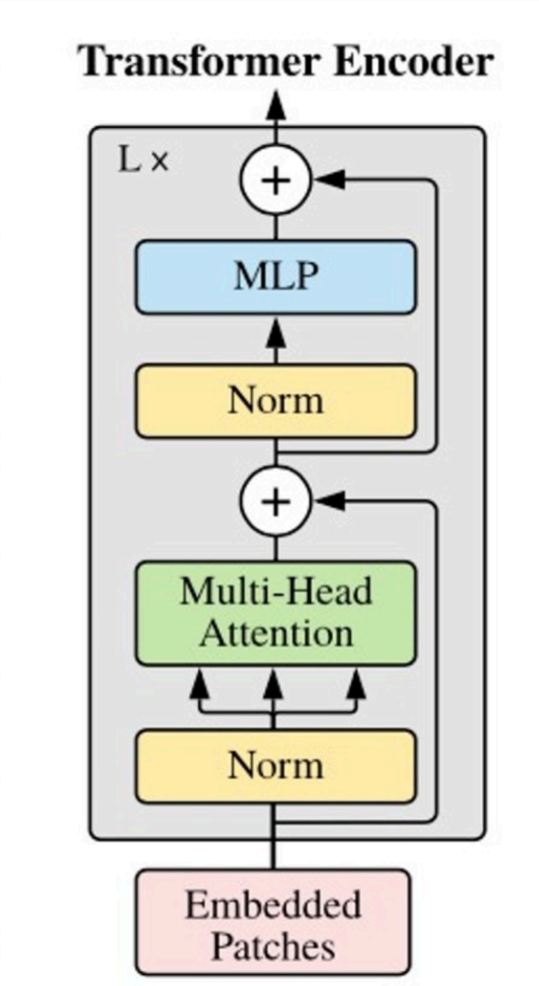


Image Encoder of the CLIP model is responsible for processing and encoding images. It takes an image as input and transforms it into a vector representation. The encoding process involves extracting meaningful features from the image, such as shapes, colors, textures, and patterns. These features are then transformed into a numerical format that the model can understand and work with. The goal of the image encoder is to capture the essential visual information in the image in a compact and informative vector form.

TEXT ENCODER



The text encoder in the CLIP model processes and encodes textual input, such as descriptions or captions. It converts the text into a numerical vector representation by analyzing the semantic meaning and context of the words and sentences. The text encoder understands the relationships between words, phrases, and sentences, allowing it to generate vector embeddings that capture the semantic similarities and differences between different text inputs. This encoding is crucial for enabling the model to understand and relate text descriptions to corresponding images.

LOSS FUNCTION

LOSS FOR ANSWER AND ANSWER TYPE

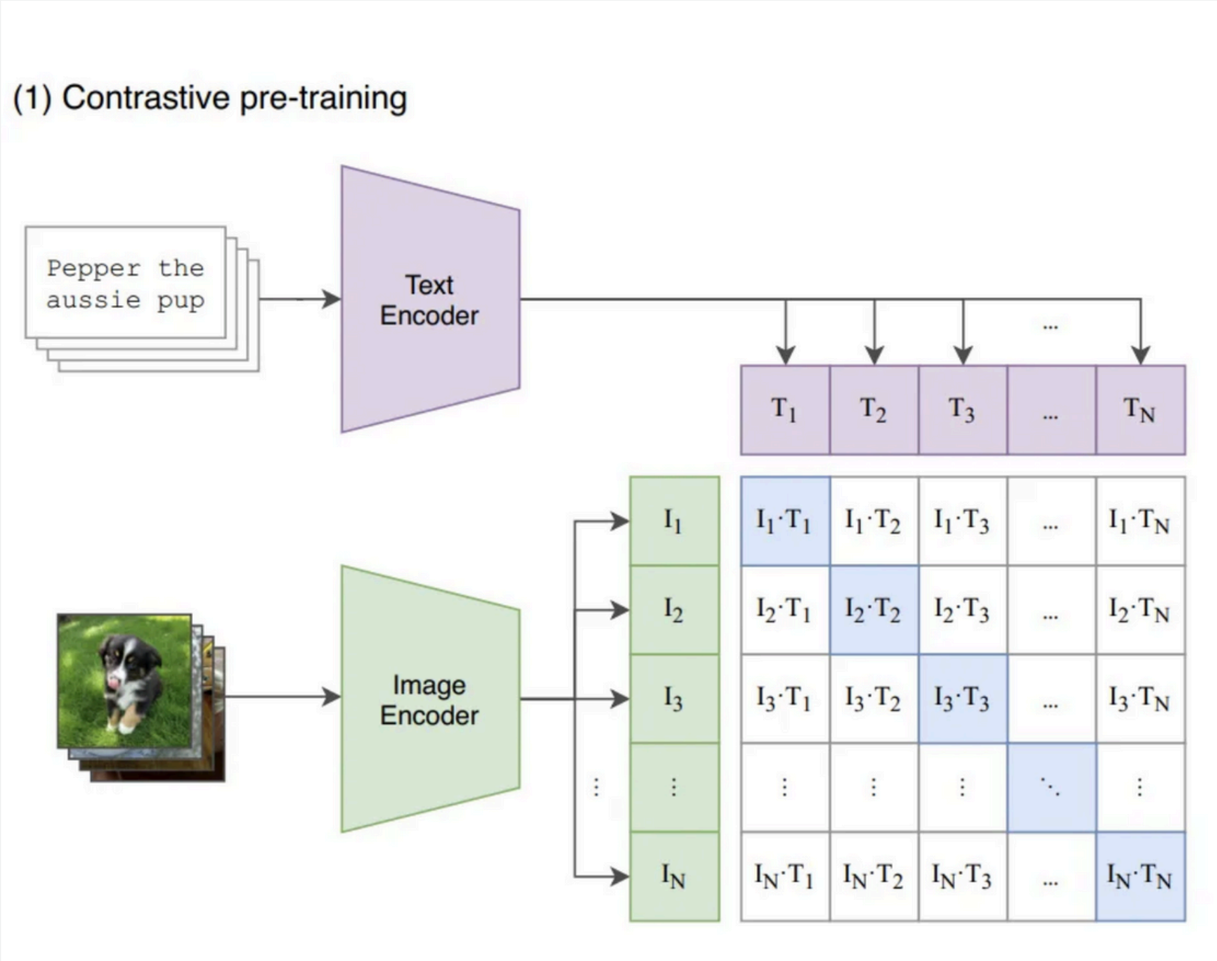
$$L_{\text{CE}} = -\sum_{i=1}^{n} t_i \log(p_i), \quad \text{for n classes,}$$

where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

LOSS FOR ANSWERABILITY

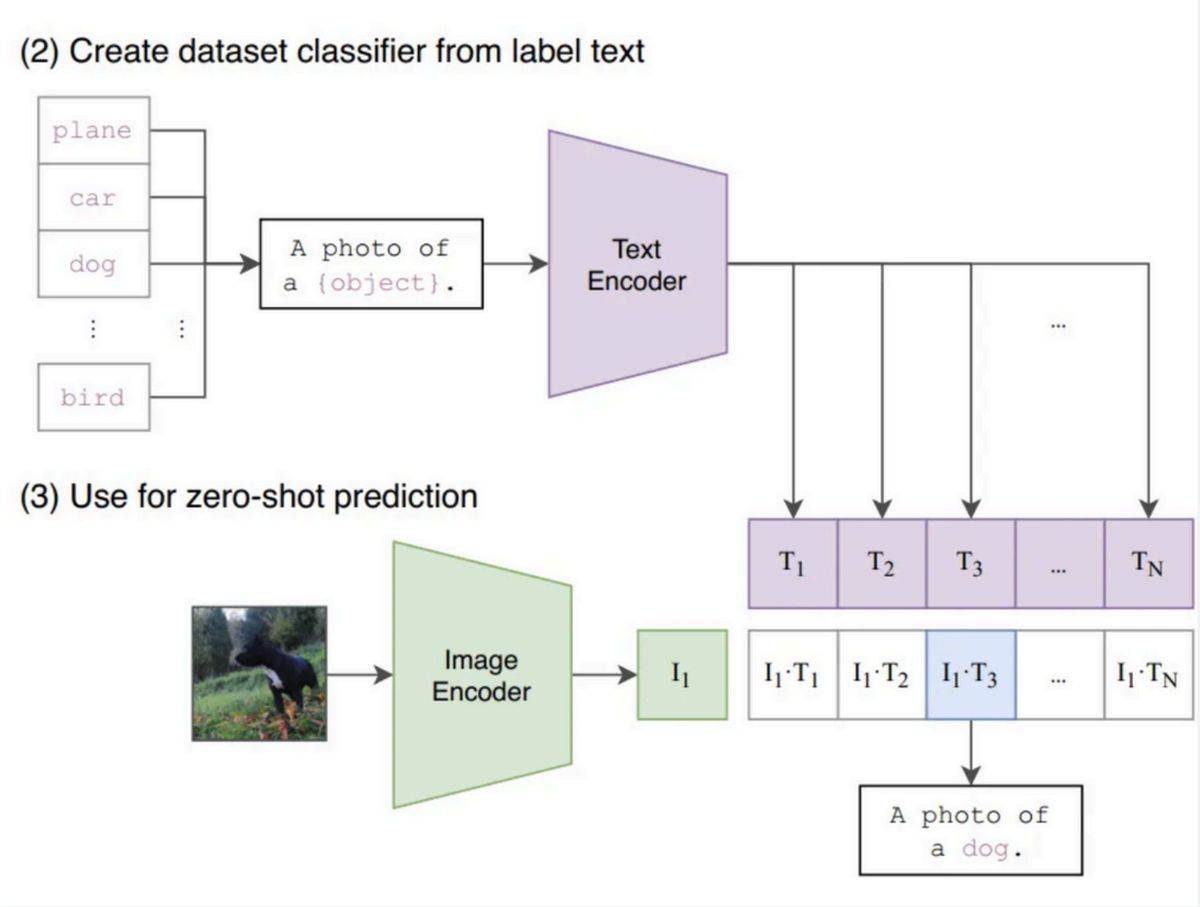
$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log (1 - \hat{Y}_i))$$

CLIP MODEL



It involves creating positive and negative pairs of instances (e.g., images and their corresponding text descriptions), embedding them into a vector space, and using a contrastive loss function to encourage similar instances to be close together and dissimilar instances to be far apart. By minimizing this loss, the model learns to capture semantic relationships between different modalities, making it capable of tasks like image classification and text-based image retrieval.

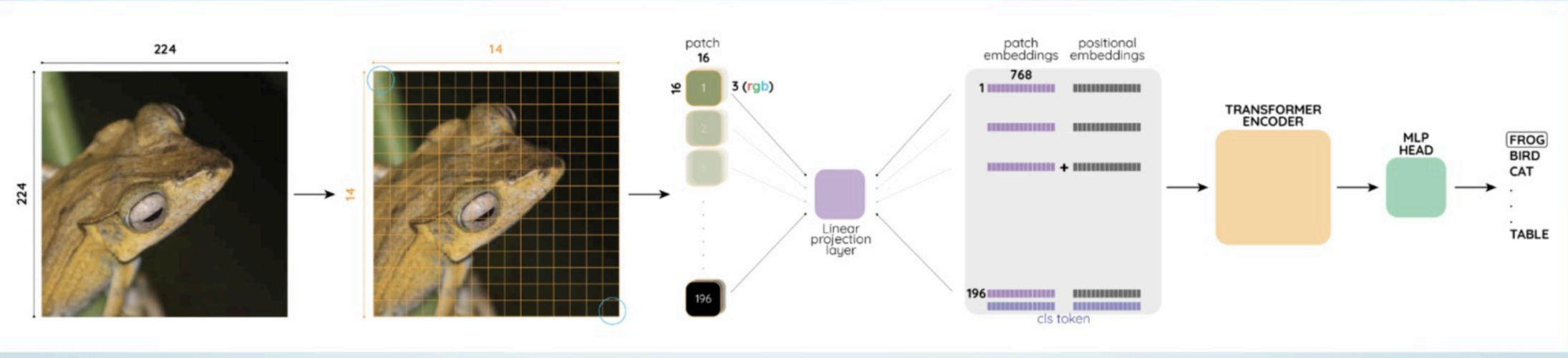
CLIP MODEL



- Creation of pairs of instances that are either similar (positive pairs) or dissimilar (negative pairs). By creating these pairs, the model learns to differentiate between instances based on their semantic similarities or differences.

- Once the pairs of instances are created, each instance (e.g., an image or a text description) is passed through the model to obtain its embedding, which is a high-dimensional representation of the instance in a continuous vector space.

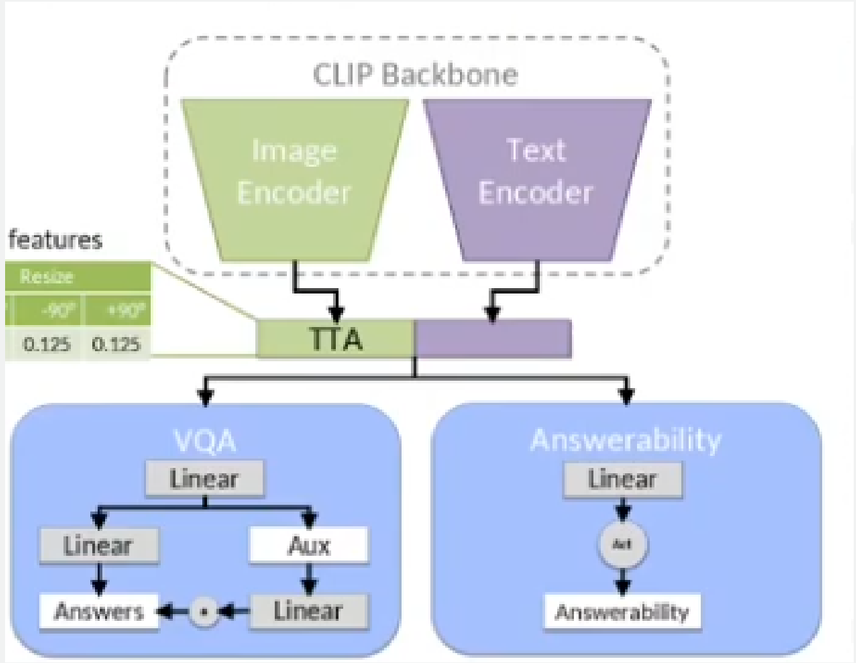
VISION TRANSFORMER (VIT)



It divide input images into smaller, non-overlapping patches, treating each patch as a token akin to words in language tasks. These patches are then linearly embedded into a lower-dimensional space and augmented with positional encodings to convey spatial information.

Through multiple transformer encoder layers, which consist of self-attention mechanisms and feedforward neural networks, the model learns to capture both local and global relationships among patches, enabling it to understand the context of the entire image.

HOW IT WORKS TOGETHER?

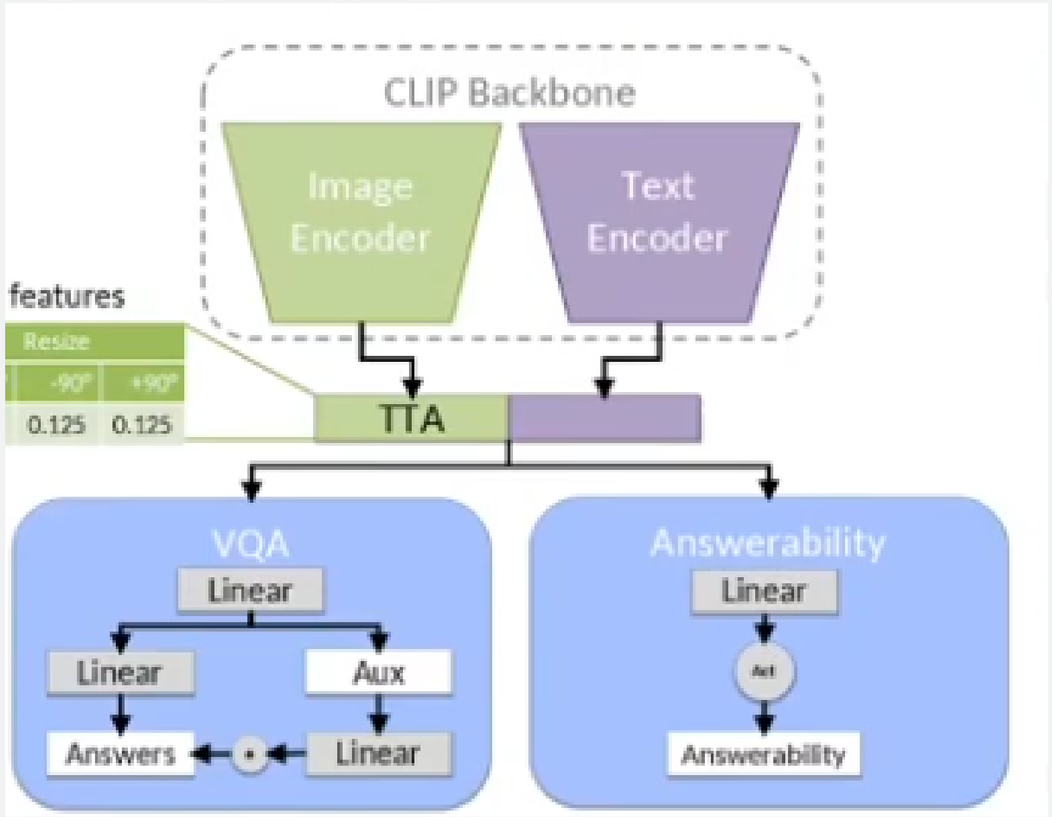


The image and question features are **first flattened and then concatenated** together along their feature dimensions, creating a unified feature tensor. This combined representation is then passed through a sequence of linear layers to perform various tasks.

A linear layer computes the **answerability score**, indicating the model's confidence in answering the question based on the given features. A **sigmoid function** is applied to the output to interpret it as a probability.

The combined features undergo transformation through a linear layer that applies **normalization, dropout, and linear transformations.** Subsequently, another linear layer predicts the **answer type**, generating probabilities for each of the predefined answer types.

HOW IT WORKS TOGETHER?



An additional linear layer generates an **answer mask** based on the **predicted answer type**, which is then applied to the output to filter out irrelevant answers. Finally, the transformed features pass through a second linear layer to produce logits representing the likelihood of each class (vocabulary size) being the correct answer.

EVALUATION

VizWiz Accuracy: Given an image and question about it, the task is to predict an accurate answer. Inspired by the VQA challenge, we use the following accuracy evaluation metric:

$$\text{accuracy} = \min\left(\frac{\#\text{ humans that provided that answer}}{3}, 1\right)$$

Evaluation metric is the minimum between 1 and the number of people who provided the answer minus 1.

MODEL EVALUATION

Metrics	Training	Validation	Test
Accuracy	0.769	0.464	0.537
VizWiz Accuracy:	0.809	0.600	0.685
Answerability Score	0.803	0.797	0.797

```
Training Accuracy: 0.769 | Validation Accuracy: 0.464 | Test Accuracy: 0.537
Training VizWiz Accuracy: 0.809 | Validation VizWiz Accuracy: 0.600 | Test VizWiz Accuracy: 0.685
Training Answerability Score: 0.803 | Validation Answerability Score: 0.797 | Test Answerability Score: 0.797
```

Trial Queries

COLORS

VQA (Visual Question Answering)

Choose File tshirt.jpg

what is the color of the tshirt

Ask Now

Answer: green

Predicted type: other

Answerability: 0.0025631189346313477

VQA (Visual Question Answering)

Choose File roon.jpg

what is the color of the wall

Ask Now

Answer: brown

Predicted type: other

Answerability: 0.0015193819999694824

CURRENCY

VQA (Visual Question Answering)

Choose File rupee.jpg

what is this

Ask Now

Answer: money

Predicted type: other

Answerability: 0.0042435526847839355

VQA (Visual Question Answering)

Choose File euro.jpg

what is this?

Ask Now

Answer: 10 euros

Predicted type: other

Answerability: 0.0032418370246887207

OCR

Choose File butter.png

Please tell me what is this

Ask Now

Answer: butter

Predicted type: other

Answerability: 0.05490767955780029

Choose File cheetos.jpg

what is this

Ask Now

Answer: cheetos

Predicted type: other

Answerability: 0.005253016948699951

ENVIRONMENT

VQA (Visual Question Answering)

Choose File rain.jpg

is this raining outside?

Ask Now

Answer: yes

Predicted type: yes/no

Answerability: 0.004198014736175537

FUTURE SCOPE

1. **Enhanced Multimodal Understanding:**
   - Expanding the CLIP model to incorporate more modalities beyond just images and text, such as audio or video, can lead to a deeper understanding of multimodal data. This advancement could pave the way for applications in multimodal analysis, content recommendation systems, and immersive media experiences.
2. **Continual Learning and Adaptation:**
   - Implementing continual learning techniques within CLIP can enable the model to adapt and improve over time as it encounters new data. This approach would enhance CLIP's ability to handle evolving concepts and dynamic environments, making it more robust and versatile across a range of applications.
3. **Personalized and Contextualized Recommendations:**
   - Leveraging CLIP's capabilities in natural language understanding and image recognition, personalized recommendation systems can be developed. By considering individual preferences, contexts, and intents, these systems can provide tailored content recommendations, enhancing user experience and engagement in various domains like e-commerce, entertainment, and education.

REFERENCES

- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (Year). Learning Transferable Visual Models From Natural Language Supervision. [Journal/Conference], Volume(Issue), Page range.

- Deuser, F., Habel, K., Rosch, P. J., & Oswald, N. (Year). Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model. [Journal/Conference], Volume(Issue), Page range.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (Year). Attention is All You Need. [Journal/Conference], Volume(Issue), Page range.

- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). VQA: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

LET’S SEE OUR PROJECT

THANK YOU!!!