Automatic Image Captioning

Generating Descriptive Captions for Images Using Transformers

(Transformers, Vision Transformers and Data Efficient Image Transformers)

ESI-SBA | Artificial Intelligence and Data Science

Introduction

Automatic image captioning is a multidisciplinary task that combines computer vision and natural language processing techniques to construct deep learning systems capable of generating textual descriptions for images.

In this project, we will elaborate on our approach using an Encoder-Decoder architecture based on transformers and vision transformers to construct a model capable of generating captions for images.

What are the challenges?

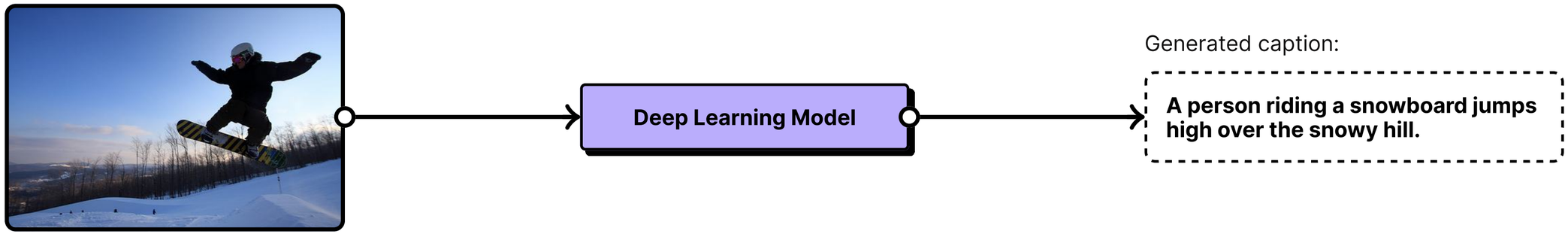
Achieving human-like captioning capabilities requires overcoming significant hurdles, including accurately recognizing and **interpreting complex visual scenes**, **understanding contextual relationships**, and **generating coherent and natural language descriptions**.

What are the objectives?

Our primary objective of this project is to develop a deep learning model capable of generating accurate and natural language descriptions for a diverse set of images.

We aim to leverage the power of transformer architectures and their ability to effectively capture and integrate visual and textual information.

Input Image:



Used tools

PyTorch for preprocessing, training and evaluation

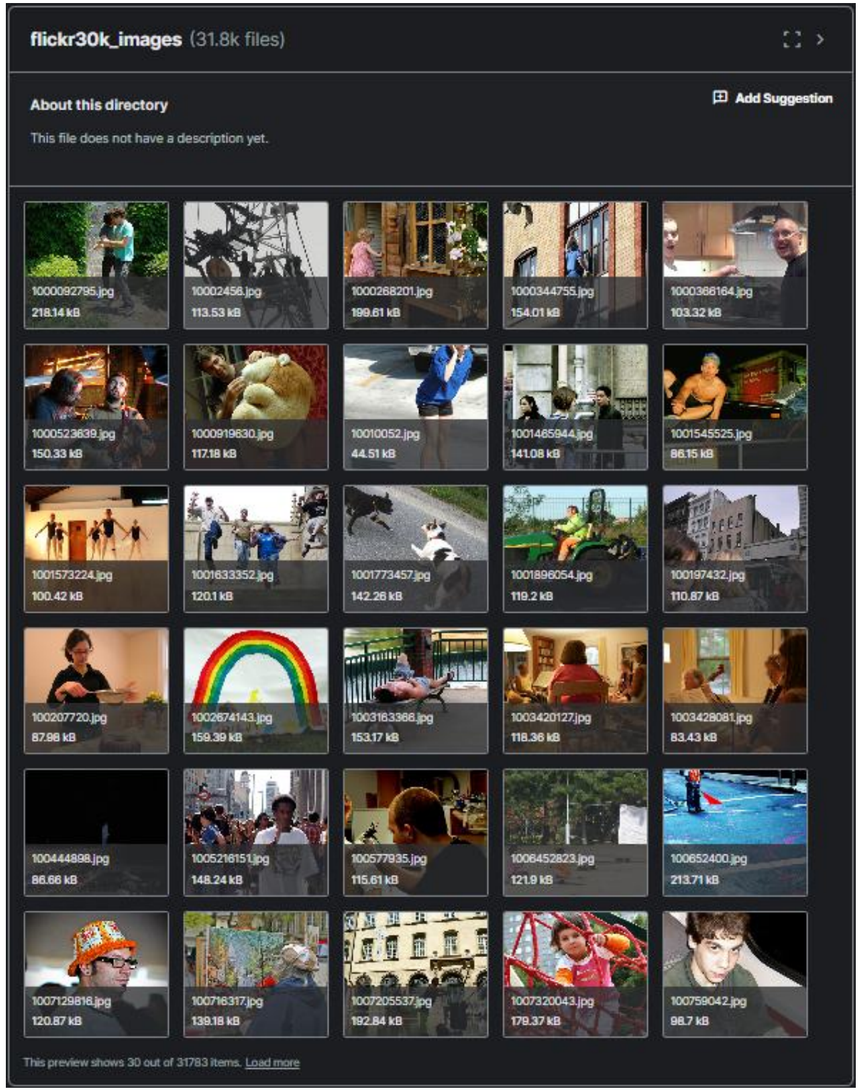
Streamlit for Deployement

Flickr30k dataset

The dataset used to train the model is the Flickr30k dataset with **31,783 images** of various resolutions.

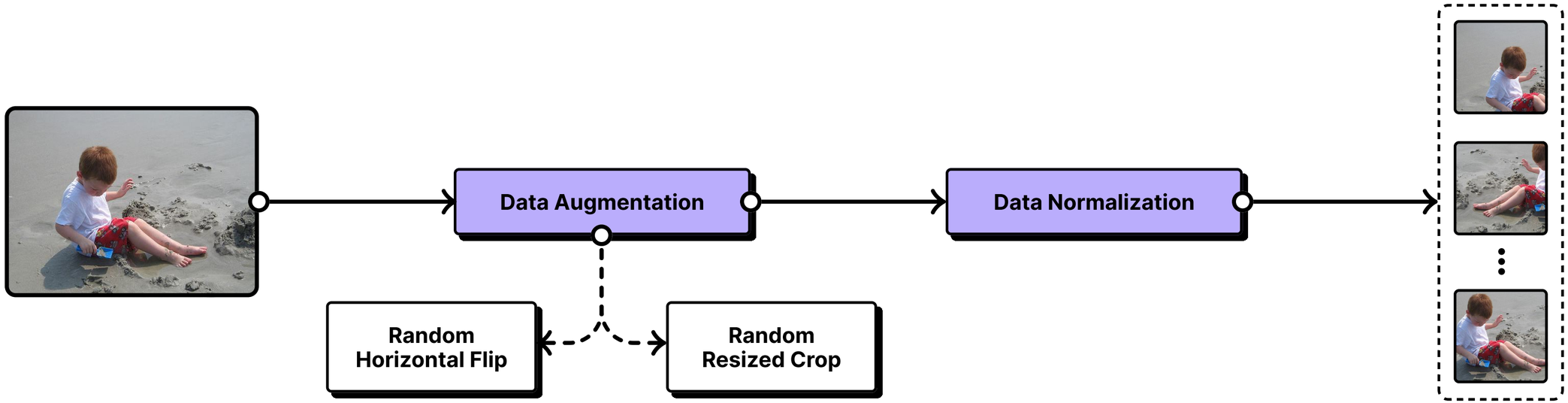
There is exactly **5 captions** per image, totaling **158,915 image-caption** pairs.

The dataset size is **4.39GB** and primarily consists of images depicting people involved in everyday activities and events.



Images Preprocessing & Augmentation

The input images were randomly horizontally flipped, and then a random sub-image of size **384 × 348 pixels** was randomly selected. The pixel values were divided by **255** as a form of normalization.



Captions Preprocessing

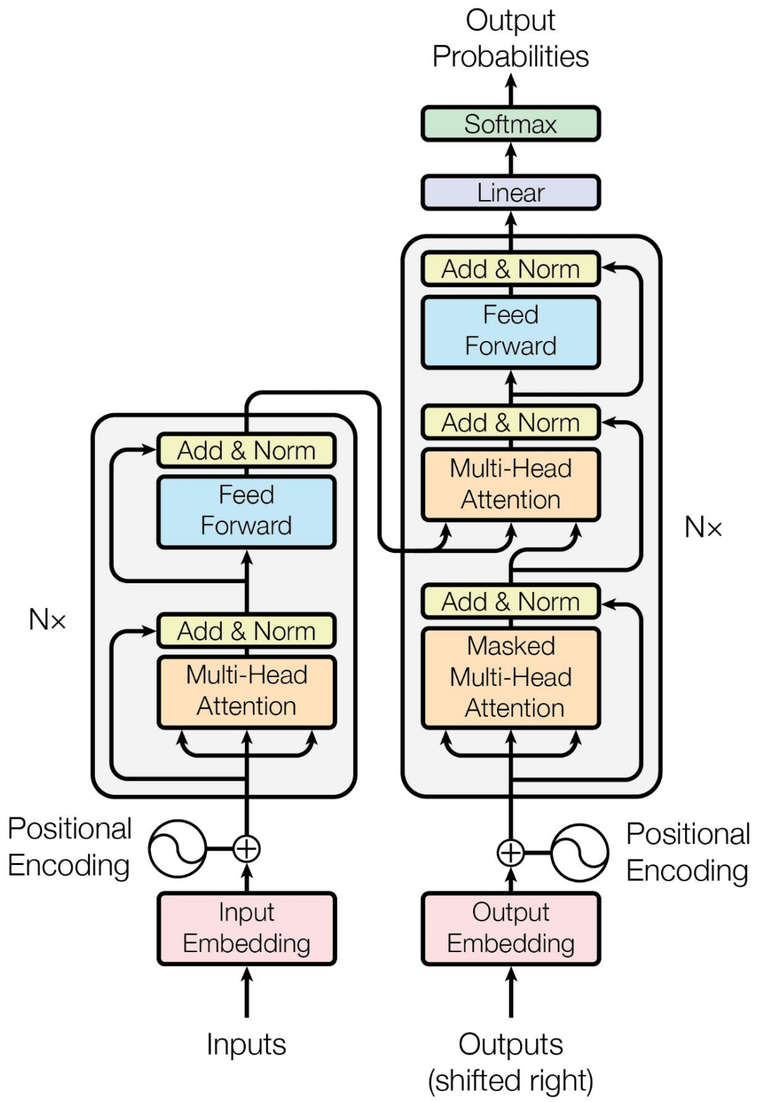
- **Caption:** “A group of men , women and children .”
- **Lowercasing:** “a group of men , women and children .”
- **Punctuation removal:** “a group of men women and children”
- **Space tokenization:** [‘a’ , ‘group’ , ‘of’ , ‘men’ , ‘women’ , ‘and’ , ‘children’]
- **Vocab indexing:** [4 , 50 , 18 , 24 , 25 , 3 , 42]
- **Padding sequences:** [1 , 4 , 50 , 18 , 24 , 25 , 3 , 42 , 0 , 0 , 0 , 0 , … , 0 , 2]

To reduce the vocab size, a token has to appear in at least **8 documents** to be added to the vocabulary

Transformers

Transformers, introduced in 2017 by **Google** in their paper [Attention is All You Need] are an **encoder-decoder** architecture designed for sequence-to-sequence tasks.

They leverage multi-head attention mechanisms to handle long-range dependencies more effectively than previous models.

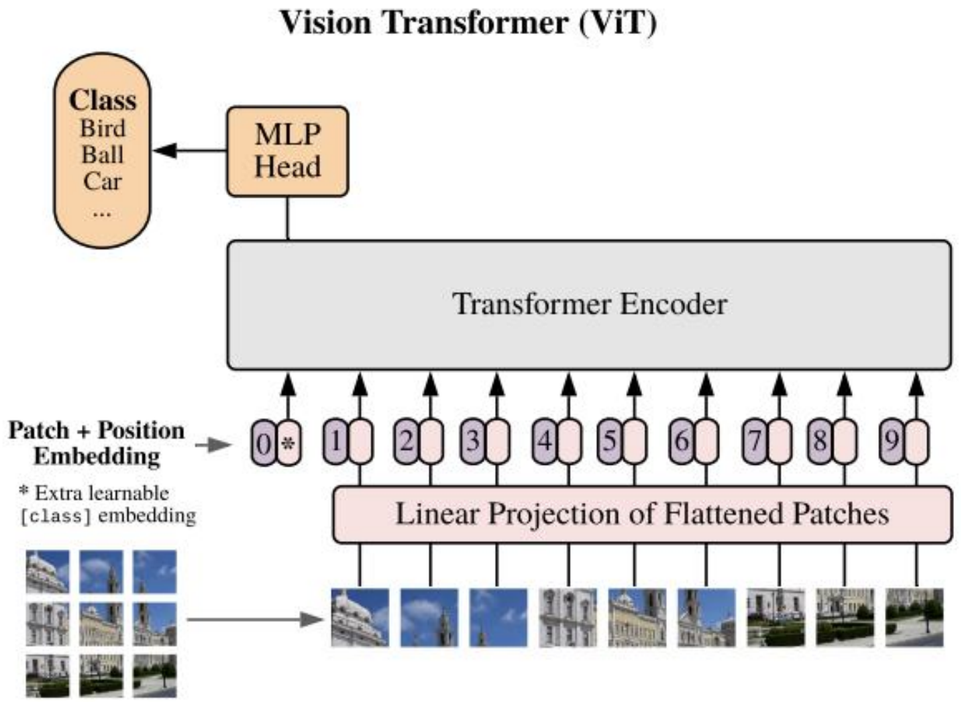


Vision Transformers

Vision transformers are **encoder-only transformers** adapted for computer vision tasks.

The idea is to break down an image into a series of patches, which are flattened into a 2D matrix. A linear projection layer is used to transform each flattened patch into a lower-dimensional vector. Finally, a positional embedding is applied.

The output is then fed to a regular transformer encoder. The encoder outputs are passed through a Multi-Layer Perceptron to produce class probabilities.



Data Efficient Image Transformers (DeiT)

DeiT, introduced in 2021 by **Facebook AI Research**, stands for Data Efficient Image Transformer.

It enhances the classic Vision Transformer to reduce the requirement for vast amounts of training data and the computational requirements of the model using **teacher-student** strategy.

The student model is first trained to mimic the behavior of a larger teacher model, which has been pre-trained on a vast amount of data, using a special **distillation token** added to the input sequence.

After this initial distillation phase, the student model is further fine-tuned on the original task using the ground truth labels.

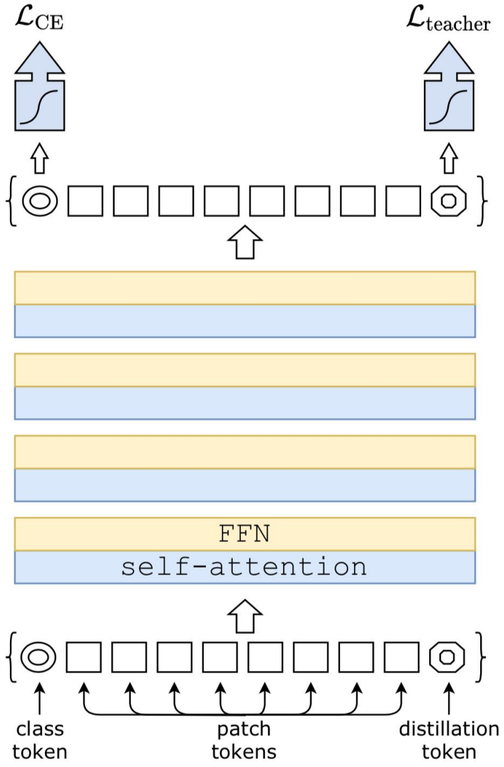
Data Efficient Image Transformers (DeiT)

- **Soft Distillation:**

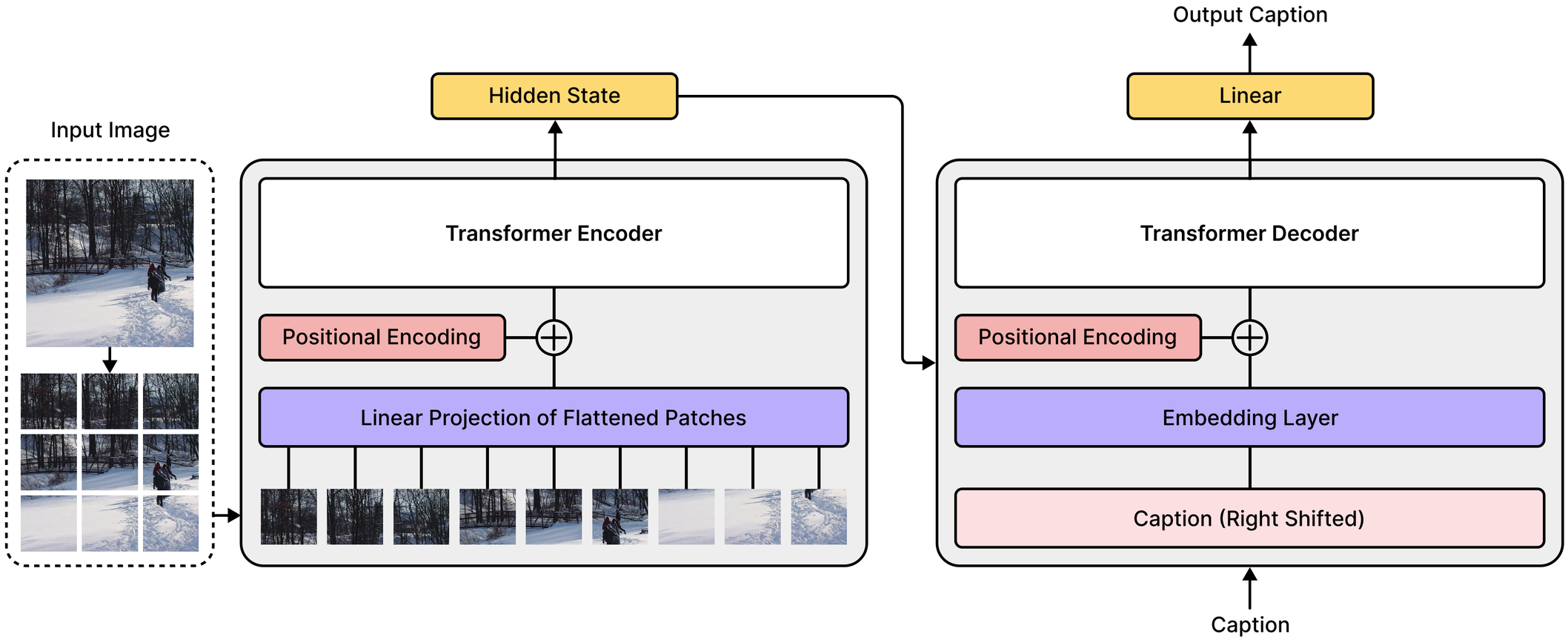
$$\mathcal{L} = (1-\lambda) * \mathcal{L}_{ce}(f(Z_s), y) + \lambda * \tau^2 * KL(f(\frac{Z_t}{\tau}), f(\frac{Z_s}{\tau}))$$

- **Hard Distillation:**

$$\mathcal{L} = \frac{1}{2}\mathcal{L}_{ce}(f(Z_s), y) + \frac{1}{2}\mathcal{L}_{ce}(f(Z_s), y_t)$$



Encoder-Decoder architecture with a DeiT-III



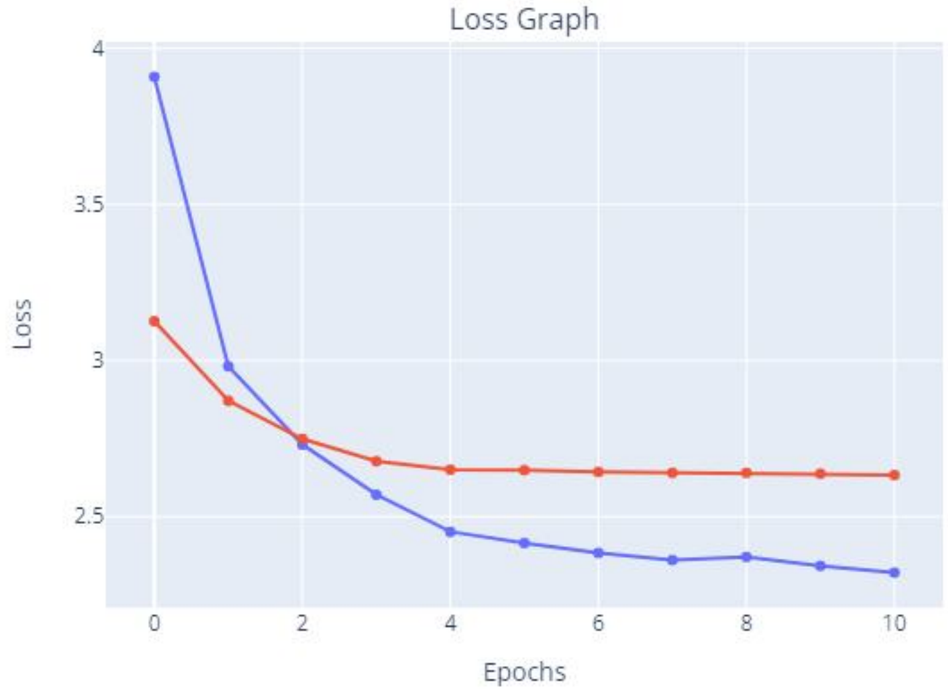
Training Experimentation

The model's encoder was initialized using weights of **pre-trained DeiT-III** on the **ImageNet-22k** dataset, and we further fine-tuned it on our task instead of freezing it, while the decoder's weights were initialized randomly.

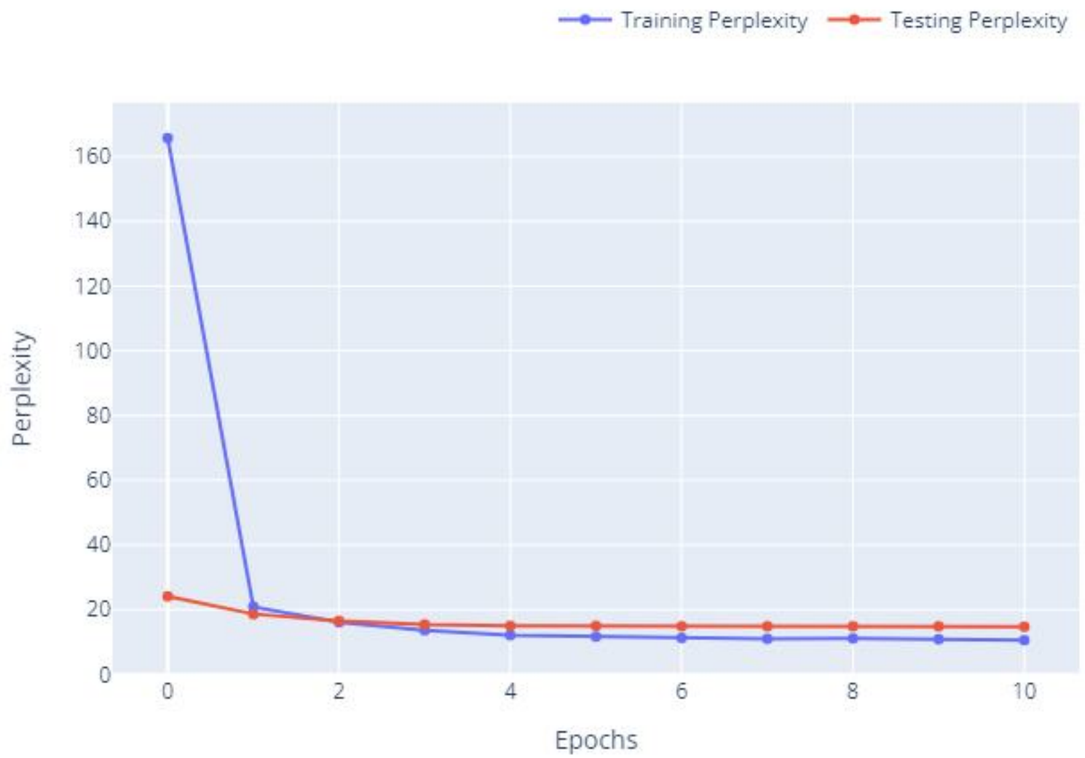
The model was trained for **11 epochs** across three different sessions using the **AdamW optimizer** with **learning rate of 10−4** and a **weight decay of 10−4**, a **batch size of 8**, and a **linear learning rate decay**.

The training was done on an **NVIDIA GeForce RTX 3050 Ti** Laptop GPU, and it took an average of **115 minutes** per epoch, making the total training time around **21 hours**.

Training & Results



Training & Results



Training & Results

BLEU score (Bilingual Evaluation Understudy):

The BLEU score is a metric used to evaluate the quality of machine-translation models or models that are trying to solve similar tasks.

It works by comparing separate parts of a **candidate text** (model's output) with a **set of references texts**, and assigns a score to each part, then these scores are averaged to give the final score.

The final score ranges from **0** to **1**, with higher scores indicating better performance.

BLUE-1	BLUE-2	BLUE-3	BLUE-4
0.6335	0.4656	0.3403	0.2451

Training & Results

ROUGE score (Recall-Oriented Understudy for Gisting Evaluation):

The ROUGE score is a metric used to measure the similarity between a **machine-generated summary** and **reference summaries**, It does so by comparing overlapping **n-grams**.

ROUGE scores range from **0** to **1**, with higher scores indicating a greater similarity between the automatically generated summary and the reference summaries.

	ROUGE-1	ROUGE-2	ROUGE-L
F1	0.5135	0.2569	0.4728
Precision	0.5477	0.274	0.5029
Recall	0.5106	0.2576	0.4711

Testing the Model

To further test our model's ability to generalize on unseen data, we handpicked **20 images** from the internet. These images contained similar objects to those in the training set but had different resolution ranges. Below are the results for some of them :

A man in a blue shirt and white hat sits on a park bench

A man and a dog are walking on the beach

A little girl is drawing on a table

Three dogs running through a grassy field

Summary & Conclusion

The ability to automatically generate accurate and human-like textual descriptions for images is a significant milestone in the field of artificial intelligence.

Through this project, we have successfully bridged the gap between visual and linguistic representations by developing a robust automatic image captioning model.

Our approach, which combines vision transformers and regular transformers, has demonstrated exceptional performance in understanding complex visual scenes and generating natural language descriptions.

Github Repository

You can access the full project **here** if you want to see all the details and test the model.

Thanks for your Listening! ☺

- Abdelnour FELLAH - (ab.fellah@esi-sba.dz)
- Abderrahmene BENOUNENE - (a.benounene@esi-sba.dz)
- Adel Abdelkader MOKADEM - (aa.mokadem@esi-sba.dz)
- Meriem MEKKI - (me.mekki@esi-sba.dz)
- Yacine Lazreg BENYAMINA - (yl.benyamina@esi-sba.dz)