Text Based Image Retrieval System

Abdulaziz Khaled Jawad

December 5, 2024

Instructor: Dr.Hammam Al-Ghamdi

Contents

1 Introduction

The field of smart devices has a huge importance in our lives, these days almost every adult owns a smart device.

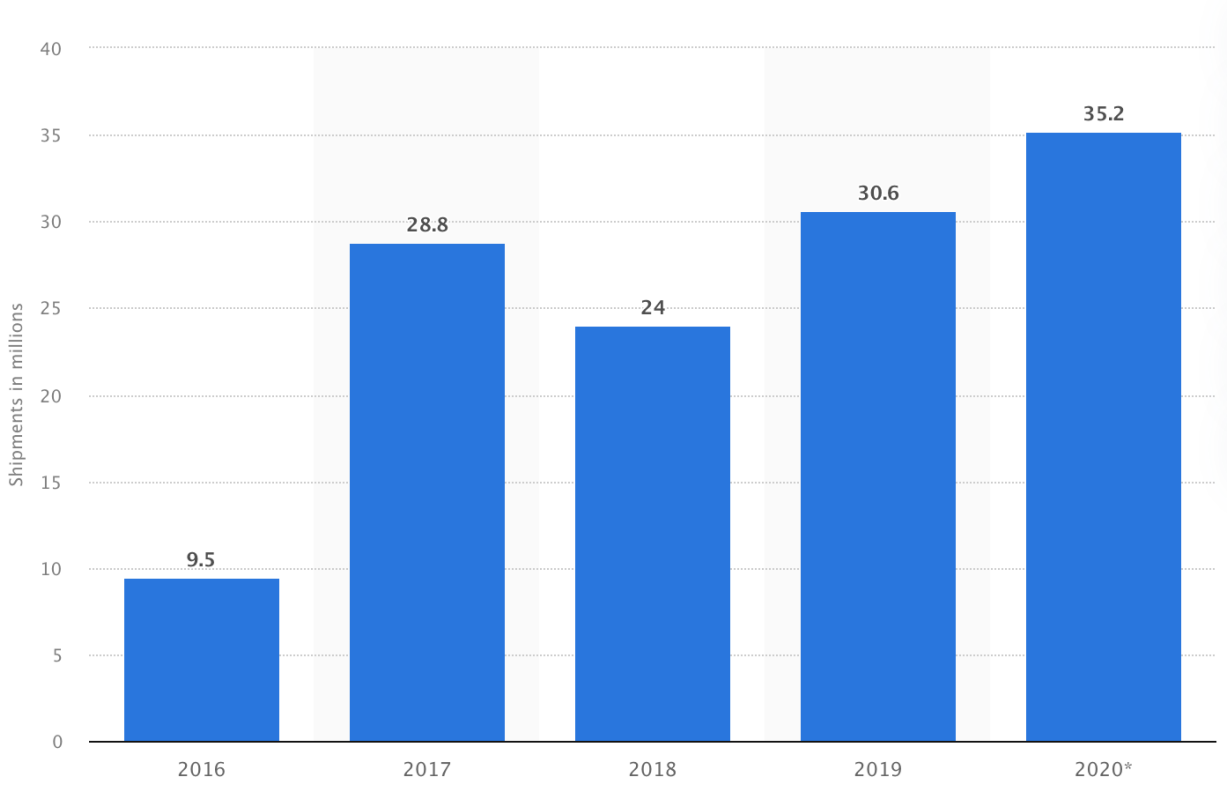


Figure 1: Smart home devices unit shipments in the United States from 2016 to 2020 (in millions)

As the :figure shows the number of smart device shipments in the US is increasing yearly

Almost all smart devices have images, Images are one of the most used things nowadays, with images

people describe their ideas, with images people communicate and share their daily life and news, and with images they react. But with images comes some problems, and one of these problems is the loss of some images, most of the time we want to react or communicate or even complete some applications online and we need the images either we don't find them or find them when it's too late. To back up our problem we asked several smart devise users: (Have you ever wanted to find an image and you didn't find it fast? or didn't find it at all?)

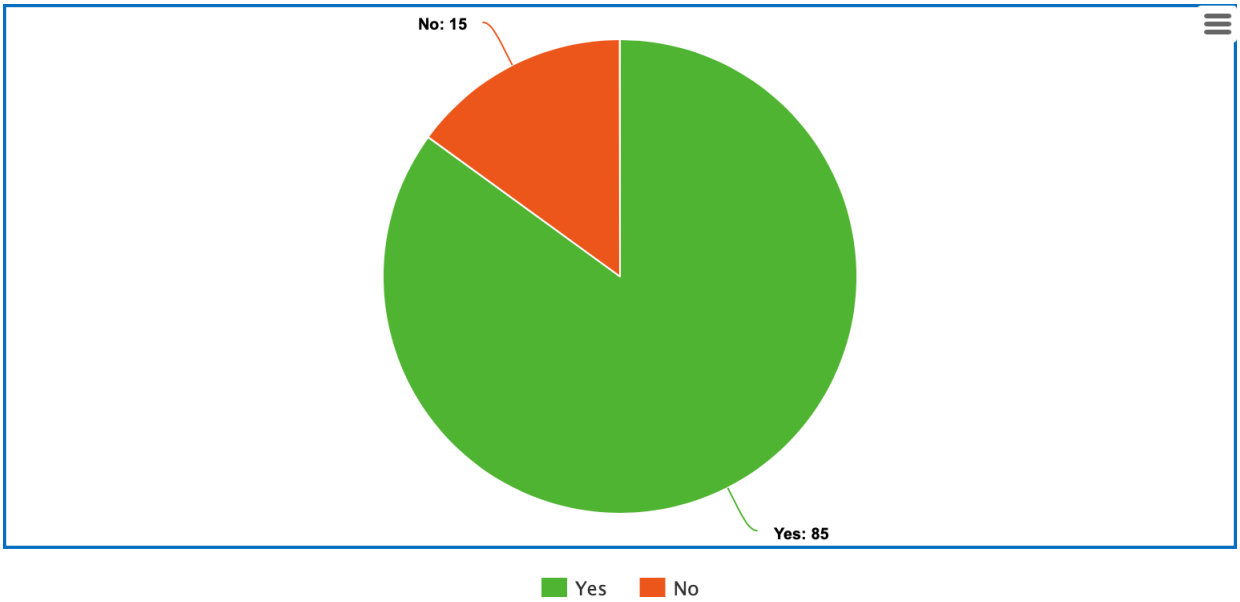


Figure 2: Those were the answers that we've received. During this project, we intend to develop an easy and fast way to :find the images needed when and as soon as needed with the help of AI(artificial intelligence).

We found that if we created an image search engine that takes the image description as text and returns the image, we would solve this problem.

2 Literature Review

With technological advancement, these image retrieval techniques went through a continuous development shift from Text-based image Retrieval to Content-Based Image Retrieval, and finally bi-directional retrieval. Most of such evolvements take place by sorting out the failures of one methodology and enhancing more perfect aligning amongst their textual queries to visual data accordingly.

Text-to-image retrieval can be considered one pioneer area aligning the textual input with the corresponding images. The paper "A Cross-Modal Coherence Model for Text-to-Image Retrieval" [6] developed the Cross-Modal Coherence Model, which improved retrieval accuracy by taking temporal and causal relations between objects in images with coherence-aware mechanisms. This is a good approach, but its high computational intensity limited it scalability for either real-time applications or large datasets. In 2021, Radford et al. came up with CLIP : "Learning Transferable Visual Models From Natural Language Supervision."[8] Without coherence-aware training, alignment of texts and images was considered in a shared embedding space while training on a large dataset of image-text pairs to achieve extremely impressive zero-shot performances for different retrieval tasks. However, it usually has a shortage in the fine-grained contextual relationships of image objects for capturing interactions because of over-reliance on global alignment. Thus, capturing detailed semantic understanding in applications remains limited.

As the limitations in text-based methods became certain, researchers started to rely on low-level visual features of color, shape, and texture applied in conventional Content-Based Image Retrieval or CBIR methods. For instance, the paper "Content-Based Image Retrieval Using Color, Shape and Texture Descriptors and Features"[9] employed RGB color models, Canny edge detection, and Gray-Level Co-occurrence Matrices for improving performance in CBIR . While another paper "Image Retrieval Based on Deep Learning"[7] explored the use of deep learning models like CNNs to enhance feature extraction and bridge the semantic gap between pixel-level data and user queries. While these methods worked perfectly on image-only data sets, they were not designed for textual queries and hence could not become suitable for cross-modal tasks. This shortcoming then continuously grew as the demand grew higher for systems that combine both textual and visual data in them.

In order to overcome the difficulties, bi-directional models of image-text retrieval are implemented by utilizing text-to-image retrieval and image-to-text retrieval. PFAN++ introduced the Pyramid Feature Attention Network to align the text with specific image regions, improving retrieval accuracy through multi-level feature abstractions. MagicLens employed self-supervised learning on a large dataset of image-text triplets, thus enabling it to handle open-ended instructions and

retrieve images based on complex relationships [11]. These were increases forward, really, showing how well the models could handle nuanced queries. However, with added unmanageable complexity and demands, such models remain rather impractical for simple tasks where efficiency may be an important component in the real world.

While MagicLens is very advanced at matching images and text , it takes two inputs , text and image . then retrieve an image based on the inputs. However , CLIP model can take text as an input and retrieve an image based on that text input .This makes MagicLens less suitable for simpler tasks that only need basic text-to-image matching .Also , MagicLens uses more detailed relationship to handle search request for images , which is more than necessary for straightforward tasks. This gap suggests an opportunity to use CLIP as it more direct for text-to-image retrieval, CLIP also has zero shot accuracy for new unseen data making it more flexible . Our project takes advantage of CLIP's strengths to provide a faster, more user-friendly solution for everyday retrieval tasks .

3 Status of the Art (SOA)

3.1 Relevant Models:

For TBIR problem, the SOA models have focused on Vision-Language models that work for both visual and textual data. So here's some key models that :its our downstream analysis:

- ALIGN (Vision-Language Pretraining): ALIGN is a multi-modal vision and language model that has been trained on 1.8B dataset of image-text pairs. It can be used for image-text similarity and for zero-shot image classi:ication. ALIGN features a dual-encoder architecture with Ef:icientNet as its vision encoder and BERT as its text encoder. What makes ALIGN such a good model is it training, because it has been trained on a large noisy data, but that costs them lower zero-shot accuracy which makes the model a bit more domain-speci:ic.[06]

- ViLT (Vision-and-Language Transformer Without Convolution or Region Supervision):ViLT is the simplest architecture by far for a vision and language model as it uses the transformer module to both extract and process visual features insted of a separate deep visual embedder such in other models (e.g., ALIGN). This design basically leads to signi:icant runtime and parameter ef:iciency. The model may underperform in visual feature extraction tasks since it doesn't use a dedicated network.[07]

3.2 Applicability to Your Project:

Both of the models mentioned share the same idea but different approaches. Our project will be using CLIP's model and modify it (:ine-tuning), by doing more training on the dataset we've chosen (Flickr8k), so that more image retrieval accuracy achieved, and less textuall-input variance occurs.

4 Dataset

4.1 Dataset Description:

For the data selection we've chosen Flickr (Flickr8k) dataset which was created and released by the University of Illinois at Urbana-Champaign. It contains images with their corresponding captions that describes what the image is about.

The Flick8k dataset can be found on Kaggle or downloaded by request on their Website. The dataset consists of 8,000 images (JPEG), Each image associated with :ive captions (Texts), resulting in total of 40,000 captions.[08]

4.2 Dataset Relevance:

The Flickr8k dataset is highly relevant for our TBIR (Text-Based Image Retrieval), because it consists of both images and corresponding textual descriptions. Which seems perfect for the model we've selected (CLIP) due to it design to understand the relationships between text and images. Combining visual and textual data will help the model to learn strong connections between the texts and images, allowing for more accurate image retrieval when given a text input.

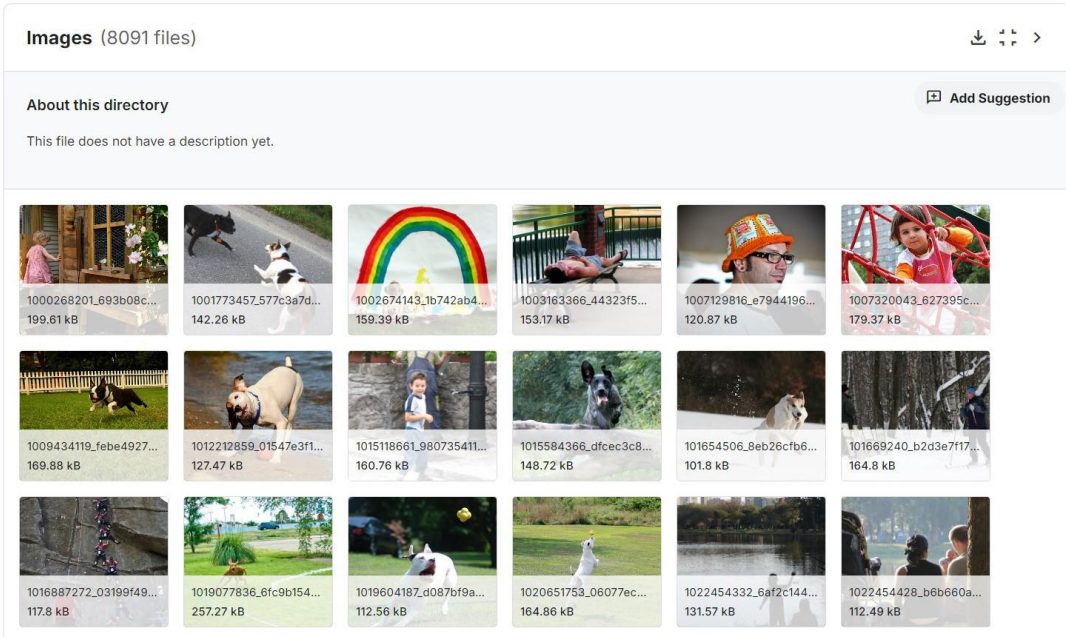


Figure 3: Snapshot of The Dataset

Figure 4: First image in the Dataset

```
image,caption
1000268201_693b08cb0e.jpg,A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg,A girl going into a wooden building .
1000268201_693b08cb0e.jpg,A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg,A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg,A little girl in a pink dress going into a wooden cabin .
```

Figure 5: Sample of the captions related to the First Image

5 Baseline Model

The baseline model for this project will be CLIP (Contrastive Language–Image Pretraining) developed by OpenAI, CLIP is a multimodal model that can deal perfectly with both images and text, allowing it to do either TBIR (Text-Based Image Retrieval) or CBIR (Content-Based Image Retrieval), so it already has a very strong linking ability that suits our downstream analysis.

Since CLIP has been pretrained on 400 million image-text pairs collected from a variety of publicly available sources on the internet, it does not require retraining from scratch for image retrieval. It can be Fine-tuned on the Flickr8k dataset, ensuring that it adapts to the speci:ic image-caption pairsavailable.[09]

6 Proposal of Baseline Model Selection

6.1 Architecture:

We'll be using the CLIP baseline but after fine-tuning it to fit our dataset by freezing the low-level feature layers and train the few last layers (High Level Feature) which are the Transformer and the Text Projection Layers.

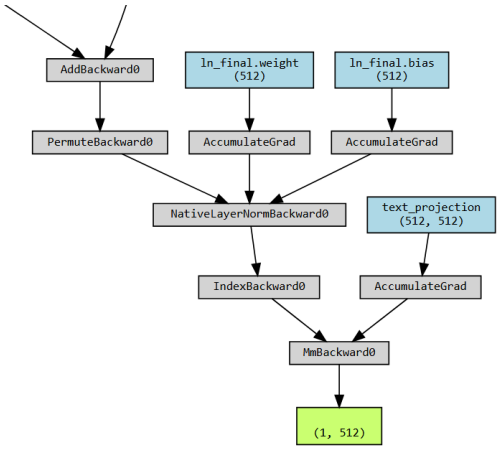


Figure 6: Sample of the training layers

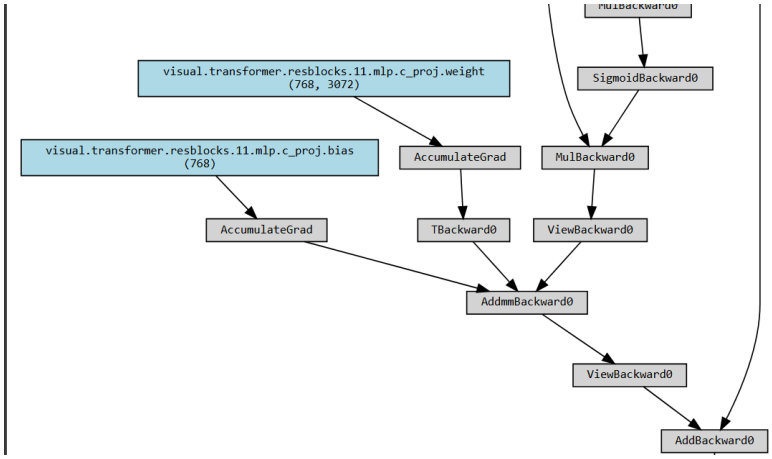


Figure 7: Sample of the training layers

```
Trainable Layers (25): ['text_projection', 'visual.t
Non-Trainable Layers (277): ['positional_embedding',
Trainable Layers: ['text_projection', 'visual.tr
Number of Trainable Parameters: 10502400
Number of Non-Trainable Parameters: 140774913
```

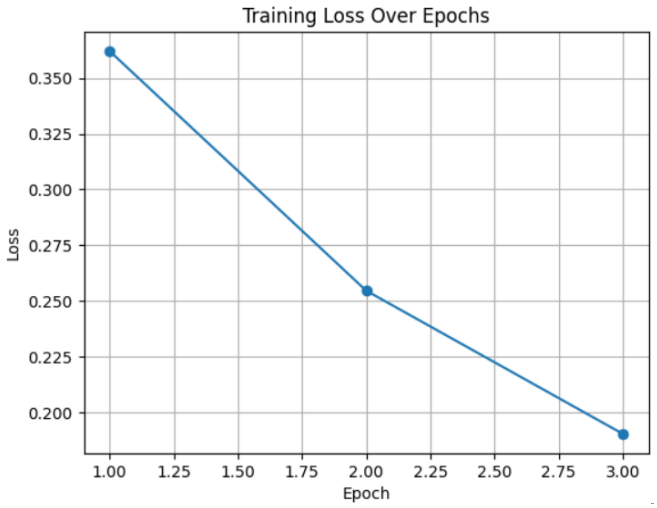
Figure 8: Number of training layers and parms

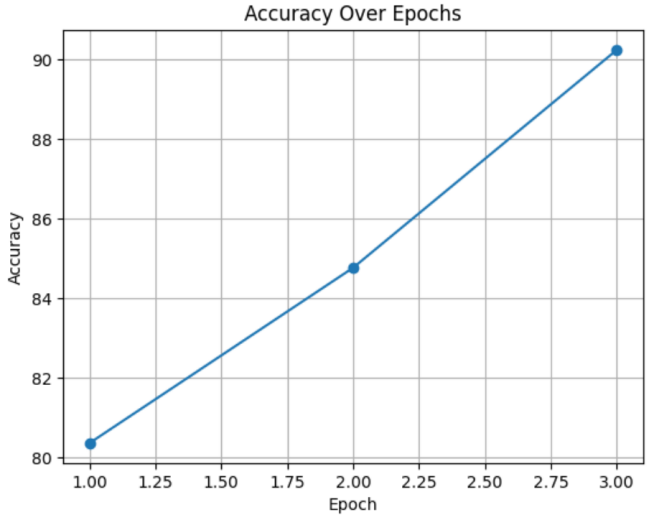
6.2 Added Value:

After fine-tuning the text layers we got extraordinary results better than the baseline, that show more accurate recall and accuracy and helped to understand the textual input far more better.

6.3 Results:

Here's some of the graphs that shows how the training went and our Proposed model results.





```
Final Top-5 Accuracy: 91.30%
```

Figure 9: Results

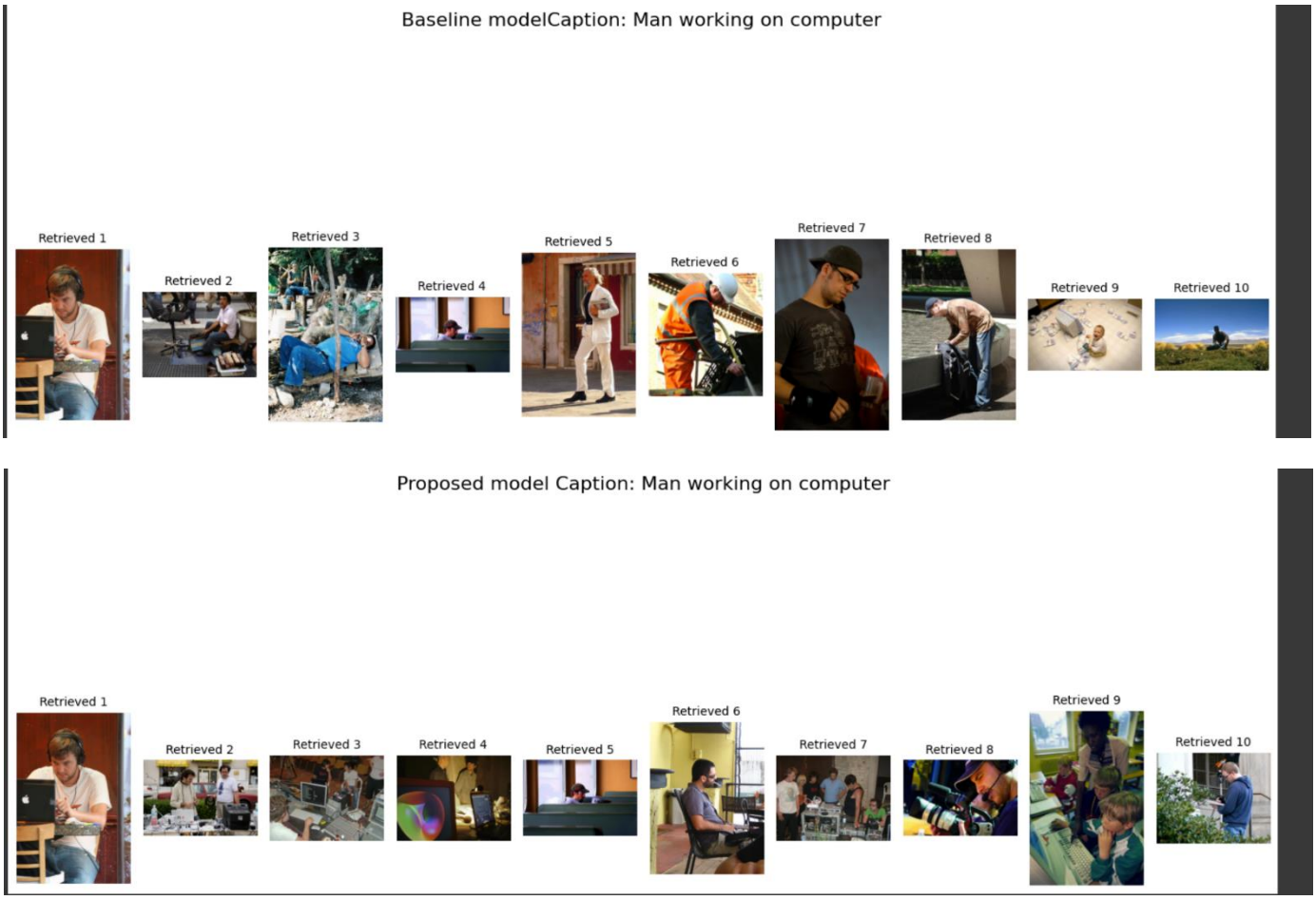


Figure 10: Comparison between Baseline and Proposed model

```
No checkpoint found, starting training from epoch 0
Epoch 1: 100%|          | 5057/5057 [1:45:20<00:00,  1.25s/batch, loss=0.342]
Checkpoint saved for epoch 1
Epoch 2: 100%|          | 5057/5057 [1:44:35<00:00,  1.24s/batch, loss=0.199]
Checkpoint saved for epoch 2
Epoch 3: 100%|          | 5057/5057 [1:44:39<00:00,  1.24s/batch, loss=0.235]
Checkpoint saved for epoch 3
```

Figure 11: Training process

7 Conclusion

In the end, we completed all the research we needed and gathered all the knowledge needed to make a complete and useful project. From the many sources that we have searched during the time that we took to study this project, we have found that this model will take a large amount of time during the training phase, and it did but eventually the results were worth it by exceeding the baseline model and getting very accurate images retrieved I believe that we've accomplished the project objectives and aims.

8 References

01 https://www.statista.com/statistics/798370/us-smart-home-units/

02 https://sh-tsang.medium.com/review-align-scaling-up-visual-and-vision-language-representationlearning-with-noisy-text-2970ce0c4065

03 https://proceedings.mlr.press/v139/kim21k/kim21k.pdf

04 https://www.kaggle.com/datasets/adityajn105/:lickr8k/data

05 https://openai.com/index/clip/

06 https://ojs.aaai.org/index.php/AAAI/article/view/21285

07 https://www.aasmr.org/jsms/Vol12/JSMS%20April%202022/Vol.12No.02.26.pdf

08 https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

09 https://www.researchgate.net/profile/Mutasem_Alsmadi2/publication/339037221_Content-Based_Image_Retrieval_Using_Color_Shape_and_Texture_Descriptors_and_Features/links/619636a0d7d1af224b01c481/Content-Based-Image-Retrieval-Using-Color-Shape-and-Texture-Descriptors-and-Features.pdf

10 http://www.smiles-xjtu.com/html/Our%20Paper/PFAN-TMM-online.pdf

11 https://arxiv.org/pdf/2203.15867

12 https://arxiv.org/pdf/2403.19651