Fine Tuning Image Transformers

Mohit Mehta, Inder Preet Singh, Pulkit Khandelwal

New York University
mm12318@nyu.edu, isw2029@nyu.edu, pk2660@nyu.edu

Introduction

The BERT Image Transformer, or BEIT (Bao et al. 2022), is a versatile model developed by Microsoft that brings together the power of transformer architectures and the potential of visual tasks to create a unified, performant tool for image understanding and language generation. This document delves into the application of the BEIT model for two distinct but interrelated tasks - automatic image captioning on the COCO Captioning (Common Objects in Context) dataset (Chen et al. 2015) and Visual Question Answering (VQA) on the COCO VQA dataset (Agrawal et al. 2016). The COCO dataset, known for its wide-ranging everyday scenes and object types, provides a robust testing ground for the model's image captioning capabilities. Concurrently, the VQA dataset challenges the model's prowess in interpreting visual content and generating appropriate responses to contextually based questions. With its inherent integration of BERT's language understanding and transformer's attention mechanisms, the BEiT model emerges as a promising approach to these tasks. This report will provide insights into harnessing the BEIT model for automatic image captioning on the COCO dataset and visual question answering on the VQA dataset, underscoring the model's proficiency in bridging the gap between visual perception and language understanding.

Literature Review

In the course of our research, we embarked on a thorough investigation of various cutting-edge models using transformers alongside attention for images. Our primary focus was on the BEiT (BERT Image Transformer), ViT (Vision Transformer) (Dosovitskiy et al. 2021), and DEiT (Data Efficient Image Transformer) (Touvron et al. 2020) models, each trained on Masked Image Modelling, with BEiT being the first true model to capitalize on Masked Image modeling which is quite similar to BERT pretraining task.

The DeiT model, touted for its data efficiency in image classification tasks, was a natural choice for our initial experiments. Despite its design intended to optimize performance in data-limited scenarios, it fell short in our tests on both the COCO and VQA datasets. The primary drawback observed was its relative inability to effectively bridge the gap between image understanding and language generation, a critical element for both image captioning and VQA tasks.

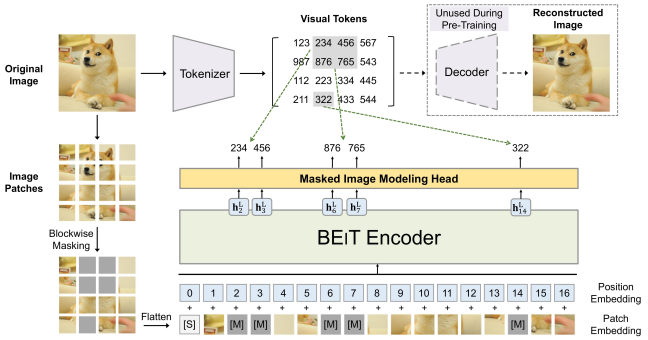

Figure 1: BEiT Architecture

Our exploration then led us to the BEIT model, an architecture developed by Microsoft that synergies BERT-style transformer pretraining with visual tasks. Remarkably, the BEIT model outperformed both DeiT and ViT on the COCO and VQA datasets. Its strength lies in its design, which inherently integrates BERT's language understanding with the attention mechanisms of transformers. This integration enabled the BEIT model to better understand the context of images, generate more accurate and contextually appropriate captions, and answer questions about images with higher precision.

In light of our findings, the BEIT model stands out as a promising direction for future research in the field of image captioning and visual question answering. Its superior performance underscores the value of tightly coupling visual perception with language understanding, a feature that may guide the development of more effective models in the future.

Dataset

The COCO (Common Objects in Context) dataset is a benchmark for computer vision tasks such as object detection, image segmentation, visual question answering (VQA), and image captioning. It contains a diverse collection of images depicting everyday scenes with objects in their contextual settings. The dataset offers broad coverage of object categories, facilitating comprehensive model training and evaluation. It includes meticulous annotations such as bounding boxes, segmentation masks, keypoints, and textual descrip-

tions.

The COCO dataset is split into three subsets: train, val, and test. The train subset contains a large number of images for primary model training. The val subset consists of previously unseen images used for model validation and fine-tuning to evaluate generalization. The test subset comprises images withheld from both training and validation, ensuring unbiased assessment of model performance on novel examples.

For VQA tasks, the COCO dataset provides question annotations corresponding to each image. These annotations form the basis for developing models that accurately understand and respond to questions about the visual content.

The COCO dataset also includes annotations specifically designed for image captioning tasks. These annotations are utilized to train image captioning models to generate accurate and contextually relevant captions.

When using the BeiT-3 model for VQA and image captioning on the COCO dataset, the COCO 2014 train images and corresponding annotations are used for model training. The COCO 2014 val images and annotations are employed for fine-tuning and validation. Finally, the COCO 2015 test images are used to evaluate the model's ability to generate accurate captions or answer questions about previously unseen images.

Methodology

Model Architecture

The architecture of the BEiT model is designed to encode an input image into contextualized vector representations. In figure 1, the diagram involves several key components:

Image Representations: Images are represented in two forms: image patches and visual tokens. The 2D image is split into a sequence of patches so that a standard Transformer can accept image data directly. These image patches preserve raw pixels and serve as input features in BEiT.

Visual Tokens: Similar to natural language, the image is represented as a sequence of discrete tokens obtained by an "image tokenizer". The image tokenizer is learned by a discrete variational autoencoder (dVAE) and includes a tokenizer and decoder module for training the model parameters.

Backbone Network: The backbone network of the model is a standard Transformer, as used in the Vision Transformer (ViT) model. The input to the Transformer is a sequence of image patches, and the model also incorporates standard learnable 1D position embeddings.

Pre-training with Masked Image Modeling (MIM): BEiT is pre-trained using a Masked Image Modeling (MIM) task (Wang et al. 2022). This involves randomly masking some percentage of image patches and predicting the visual tokens corresponding to the masked patches. Rather than randomly choosing patches for the masked positions, an algorithm called Blockwise Masking is used. This MIM task, inspired by masked language modeling used in natural language processing, aims to overcome the issue of focusing on short-range dependencies and high-frequency details by predicting discrete visual tokens, which summarize the details to high-level abstractions.

For downstream tasks such as image classification and semantic segmentation, task-specific layers are appended to the pre-trained BEiT and the parameters are fine-tuned on the specific datasets.

Hyper-Parameters

In addition to the model architecture above, we used the following hyperparameters.

Layer Decay: In the context of BEiT, the layer decay parameter is set to 1.0, indicating that the learning rate remains consistent across all transformer layers. This decision reflects the understanding that all layers in the BEiT model equally contribute to the representation of image features and the corresponding downstream tasks, thus each layer requires equal learning opportunities.

Learning Rate: The learning rate for the BEiT model is set to a relatively small value (1e-5). This low learning rate ensures that the model's fine-tuning process proceeds in a careful and controlled manner, thus preventing significant disruptions to the pre-learned representations within the model.

Weight Decay: The weight decay is set to 0.01, indicating that the model applies a small penalty to the magnitude of the weights during training. This L2 regularization helps prevent overfitting and keeps the model generalized, which is particularly important in the BEiT model due to its large parameter space and the diverse image representations it must capture.

Warmup Epochs: A single warmup epoch is used in the BEiT model's training. This warmup phase gradually increases the learning rate from a very small value to the set learning rate, helping to ensure a stable optimization process. This stability is crucial in the early stages of training BEiT, as abrupt changes in image representations can hinder the model's ability to effectively learn from the image patches and visual tokens.

Optimizer

The BEiT model primarily employs the AdamW optimizer for its training process, a variant of the Adam optimizer, noted for its effective adaptive learning rate adjustments. The default parameters for AdamW in the BEiT model are an epsilon value of 1e-8 and beta coefficients of [0.9, 0.999]. The learning rate is set to 5e-4 by default.

While AdamW is the default choice, the model's design also accommodates other optimization algorithms such as Stochastic Gradient Descent (SGD), and incorporates a momentum factor of 0.9. Techniques like gradient clipping can also be used to prevent extreme gradient values during training.

The model uses a weight decay parameter of 0.05 as a regularization term, helping to prevent overfitting and improve model generalization. In conclusion, the BEiT model's optimizer configuration is aimed at efficient and stable training, with the flexibility to adjust key parameters to suit specific requirements.

Data Augmentation

In our methodology, data augmentation plays a crucial role in enhancing model robustness. We use RandAugment, a technique that applies random transformations to training images, increasing data diversity. Additionally, we employ 'bicubic' interpolation during image resizing in training, a method that generally yields high-quality results. Together, these strategies enrich our training data, promoting better model learning and performance.

Loss Function

For our model, we have utilized a couple of distinct loss functions, which are integral to the learning process. These functions are part of the timm, or PyTorch Image Models, library, an assortment of tools and models designed to expedite and enhance the use of PyTorch for computer vision tasks.

LabelSmoothingCrossEntropy: This is a derivative of the conventional cross-entropy loss. It's particularly useful in classification tasks where the model is trained to be less confident in its predictions. By assigning a small amount of confidence to incorrect labels, it helps in preventing the model from making overconfident predictions, thus contributing to improved generalization.

SoftTargetCrossEntropy: This loss function is typically used when the targets are probabilities (soft targets) instead of hard labels. It's beneficial in scenarios such as knowledge distillation, where we want the model to learn from the soft output of another model.

These loss functions play a pivotal role in optimizing our model's predictions during the training phase, enabling it to learn more complex patterns and contribute to its overall performance.

Scheduler

The BEiT model uses a cosine scheduler for managing the learning rate during training. It starts with a warmup phase where the learning rate linearly increases from a start value to a base value. Post warmup, the scheduler adjusts the learning rate following either a cosine or linear schedule, as specified by the user. In the cosine schedule, the learning rate follows a half-cosine curve, gradually reducing from the base value to a final value. This scheduler provides an efficient control over the learning rate throughout the training process.

Figure 6 Explanation The BeiT-3 Visual Transformer model was evaluated on a test dataset for Visual Question Answering (VQA) using the Common Objects in Context (COCO) dataset. The evaluation was conducted over 83 batches, and the model achieved an average accuracy of 83.45% across the batches. The individual batch scores ranged from 80.94% to 85.14%. This demonstrates the model's capability to accurately recognize and answer questions based on visual inputs. The results indicate promising performance for VQA tasks, highlighting the potential effectiveness of the BeiT-3 Visual Transformer model in accurate visual recognition and question-answering tasks.

Finetuning

COCO Captioning

For fine-tuning on COCO captioning, we attach a fully connected layer on top of transformer heads as done similarly for BERT. The fully connected layer takes 768 as input with 64010 as output. As we are computationally bound, we first load the BEiT weights provided by Microsoft and freeze the weights of BEiT, only training the fully connected layer. This results in 270,751,754 overall parameters, with 49,223,690 trainable parameters. We used a batch size of 32, with the number of training examples being 113,350

Metrics Used

- **ROUGE_L:** ROUGE_L (Lin 2004) is an NLP metric measuring the longest common subsequence in system and reference summaries. It's useful for evaluating structure similarity in tasks like automatic summarization and machine translation.
- **Bleu_1, Bleu_2, Bleu_3, Bleu_4:** BLEU scores evaluate the quality of machine-generated text. The number indicates the size of n-grams considered (1-grams for Bleu_1, 2-grams for Bleu_2, etc.), making BLEU scores useful for comparing machine translation models.
- **CIDEr:** CIDEr (Vedantam, Zitnick, and Parikh 2015) is used in image captioning tasks. It measures the similarity between machine-generated and reference captions, factoring in human consensus on caption quality.
- **METEOR**: METEOR is used in machine translation evaluation. It considers linguistic phenomena like synonymy, stemming, and paraphrasing, leading to a higher correlation with human judgment on translation quality.

COCO VQA

The BEiT model is fine-tuned on the COCO dataset for Visual Question Answering (VQA). The visual transformer layers of the model are frozen to leverage their pre-training on a large dataset to capture visual features effectively. The fine-tuning process utilizes two A100 GPUs to address computational constraints and employs a batch size of 32. The training process involves 113,350 labeled samples from the COCO dataset to optimize the model's performance in the VQA task.

Results

COCO Captioning

In order to evaluate the COCO captioning dataset, we used the Karpathy Split. Figure 2 highlights the different validation metrics vs the epochs. Additionally, we also sampled some validation dataset and retreive the predicted label alongside with true labels shown in Figure 3, 4, 5.

COCO VQA

The model achieved an average accuracy of 83.45% and a loss of 2.1% across the batches. The results are summarized in Figure 6, which displays the accuracy scores. Furthermore, Figure 7 showcases the predictions of the BeiT

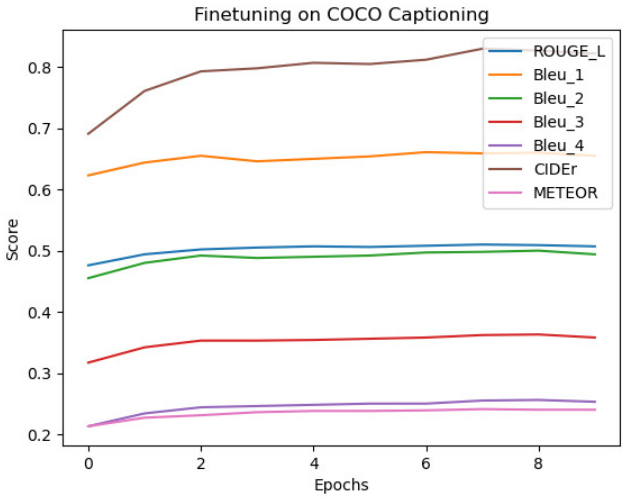

Figure 2: COCO Captioning Metrics

Figure 3: COCO Captioning **True Label:**
1. A close up of a dog sitting wearing a hat.
2. A dog wearing a striped elf hat sits in the snow.
3. A dog is wearing an elf hat in the snow.
4. A dog wearing an elf hat sits in the snow.
5. Brown and white dog in Christmas hat standing in the snow

Prediction Label: A small brown dog wearing a green hat and red hat.

Figure 4: COCO Captioning
True Label:
1. A young man riding a skateboard on top of a skate park.
2. A young man bending down ridding a skateboard.
3. A person riding a skate board near a ledge.
4. A guy riding his skateboard down a paved path.
5. A skateboard is skating down the sidewalk on his skate board.

Predicted Label:
A man riding a skateboard in a park.

model on the VQA dataset. Additionally, we selected a subset of the validation dataset and presented the predicted labels alongside the true labels in Figure 7.

Conclusion

In conclusion, this project explored the potentials of the BEiT architecture, an innovative model that leverages transformers in a self-supervised approach for image processing tasks. Through rigorous experimentation, we found that BEiT offers an impressive performance, demonstrating its proficiency in capturing visual representations effectively. The model was applied on two challenging datasets, COCO Captioning and VQA, where it showed notable results in tasks such as object detection and visual question answering, respectively.

Furthermore, the study highlighted how BEiT, pretrained through the masked image modeling task, could be effectively fine-tuned for downstream tasks, providing an impactful approach to enhance its performance on domain-specific tasks. The model's ability to handle image patches and its use of visual tokens were particularly key to its success in these tasks.

While our findings are promising, there is always room for improvement and further study. Future research might explore different pretraining strategies or delve deeper into refining the model's parameters for even better performance on specific tasks. The study leaves no doubt that transformers, like the BEiT architecture, are poised to play an increasingly important role in the field of computer vision, setting the stage for new breakthroughs and advancements.

Figure 5: COCO Captioning - **True Label:** 1. A small little bathroom with a toilet in it.
2. A white toilet in a generic public bathroom stall.
3. Modern commode in small all white and tiled water closet.
4. A photo of white toilet in a bathroom.
5. THIS IS A PHOTO LOOKING DOWN ON A TOILET
Predicted Label: A white toilet with a white lid and seat.

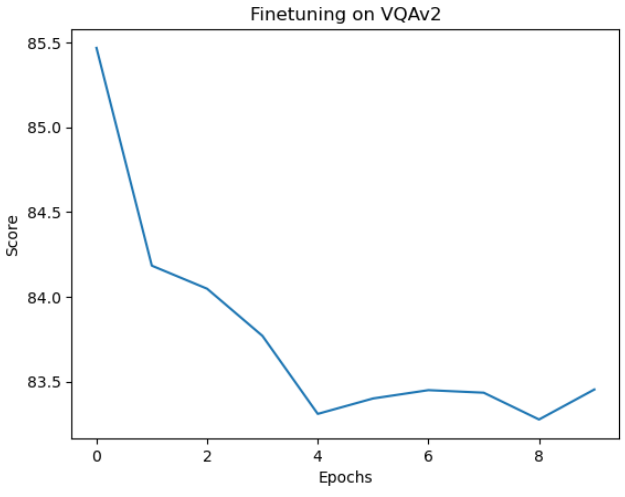

Figure 6: VQA Score

Figure 7: VQA Mobile - 1. What color is the case? Answer: Silver
2. What time is it? Answer: 02:31
3. What device is the video image on? Answer: 2
4. How many phones are there? Answer: 3

Code

The code for the project can be found on GitHub at: https://github.com/sidakwalia/DL-Final-Project

References

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2016. VQA: Visual Question Answering. arXiv:1505.00468.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Touvron, H.; Vedaldi, A.; Douze, M.; and Jégou, H. 2020. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726.

Wang, W.; Li, X.; Huang, Y.; Li, X.; Zhang, N.; Gong, C.; and Liu, W. 2022. Image as a Foreign Language: BEiT Pre-

training for All Vision and Vision-Language Tasks. *arXiv preprint arXiv:2208.10442*.