

Introduction to Deep Learning

The BEiT for medical image classification

Group 46 :

|  |  |
| --- | --- |
| Nguyễn Đình Tùng | 22BI13455 |
| Nguyễn Huy Tùng | 22BI13456 |
| Nguyễn Kiên Trung | BA12-178 |
| Lê Quốc Trung | BA12-177 |
| Đỗ Bảo Phúc | BA12-149 |
| Nguyễn Đức Bảo Minh | BI12-278 |

[1. Introduction 3](#_Toc178721795)

[2. Methods 3](#_Toc178721796)

[**2.1** **Image Representations** 3](#_Toc178721797)

[**a.** **Image Patch** 3](#_Toc178721798)

[**b.** **Visual Token** 4](#_Toc178721799)

[**2.2** **Pre-training** 5](#_Toc178721800)

[**2.3** **Fine-tuning** 5](#_Toc178721801)

[**2.4** **Mathematical Foundations & Formulas for BEiT and Related Models** 6](#_Toc178721802)

[**a.** **Self-Attention Mechanism** 6](#_Toc178721803)

[**b.**  **Transformer Layer Calculation** 6](#_Toc178721804)

[**c.**  **Image Segmentation Loss** 6](#_Toc178721805)

[**d.**  **BEiT Pre-training Objective (Masked Image Modeling)** 6](#_Toc178721806)

[3. BEiT in Medical Imaging 7](#_Toc178721807)

[**a.** **Why BEiT is Suitable for Medical Imaging** 7](#_Toc178721808)

[**b.** **Applications of BEiT in Medical Imaging** 7](#_Toc178721809)

[**c.** **Challenges in Applying BEiT to Medical Imaging** 7](#_Toc178721810)

[4. Dataset of medical image classification 8](#_Toc178721811)

[5. Apply BEiT model for chest x-ray and Result: 9](#_Toc178721812)

[6. Reference 11](#_Toc178721813)

* 1. **Introduction**

Transformers have achieved significant performance in computer vision tasks. However, empirical studies show that vision Transformers require more training data compared to convolutional neural networks (CNNs). To address this data inefficiency, self-supervised pre-training has emerged as a promising solution to leverage large-scale image datasets. Several approaches have been explored for vision Transformers, such as contrastive learning and self-distillation.

At the same time, BERT (Bidirectional Encoder Representations from Transformers) has achieved great success in natural language processing (NLP) through its masked language modeling task, which involves masking portions of the input text and then recovering the masked tokens based on the Transformer’s encoding

However, directly applying BERT-style pre-training to image data presents challenges. Unlike text, where a predefined vocabulary exists, images do not have a pre-existing vocabulary for their input units, i.e., image patches. This makes it impossible to use a softmax classifier to predict masked patches as doing with words in NLP. A simple alternative would be to treat the task as a regression problem, predicting raw pixel values of masked patches. However, this method can be inefficient, as it may waste modeling capacity on short-range dependencies and high-frequency details.

In this work, we introduce BEIT (Bidirectional Encoder Representation from Image Transformers), a self-supervised vision representation model inspired by BERT. We propose a novel pre-training task, masked image modeling (MIM), to address the limitations of directly applying BERT-like strategies to image data. In MIM, each image is divided into two views: image patches and visual tokens. The goal is to predict visual tokens (rather than raw pixels) for the masked image patches. Achieve this by tokenizing the image using a discrete variational autoencoder (VAE) before pre-training.

Pretrain BEIT and evaluate it on downstream tasks such as image classification and semantic segmentation. Experimental results show that BEIT outperforms both from-scratch training and previous self-supervised models. Additionally, BEIT is complementary to supervised pre-training, further improving performance with intermediate fine-tuning using labeled data. The self-attention mechanism in BEIT can also learn meaningful semantic regions in images without human annotations.

* 1. **Methods**

Given an input image x, BEIT encodes it to contextualized vector representations.BEIT is pretrained by the masked image modeling (MIM) task in a self-supervised 2 learning manner. MIM aims at recovering the masked image patches based on encoding vectors. For downstream tasks (such as image classification, and semantic segmentation), we append task layers upon pretrained BEIT and fine-tune the parameters on the specific datasets.

* + 1. **Image Representations**

In BEIT, images are represented in two forms during pre-training: **image patches** (used as input) and **visual tokens** (used as output).

1. **Image Patch**

A screenshot of a computer

Description automatically generatedThe input image is split into a sequence of patches, following the method used by vision Transformers. Specifically, an image x∈R^(H×W×C) is reshaped into N=HW/P^2 ​ patches, where H and W are the image height and width, C is the number of channels, and P × P is the patch resolution. Each patch is flattened into a vector and linearly projected, similar to word embeddings in BERT. In the experiments, a 224x224 image is divided into a 14x14 grid, resulting in 16x16 patches.

1. **Visual Token**

Instead of using raw pixels as output, images are tokenized into discrete **visual tokens**, analogous to words in natural language processing. The image is tokenized into z=[z1,...,zN]z = [z\_1, ..., z\_N]z=[z1​,...,zN​], where the vocabulary V contains discrete token indices. The image tokenizer is learned using a **discrete variational autoencoder (dVAE)**, which consists of a tokenizer and decoder. The tokenizer maps image pixels into discrete tokens using a visual codebook, while the decoder reconstructs the input image from these tokens. The Gumbel-softmax relaxation is applied to handle the non-differentiable nature of the discrete tokens during training.

Each image is tokenized into a 14x14 grid of visual tokens, matching the number of image patches, with a vocabulary size of 8192. BEIT utilizes a publicly available image tokenizer for this process.

A diagram of a diagram

Description automatically generated

## **2.2 Pre-training**

A diagram of a masking example

Description automatically generatedBEIT pre-training leverages a **masked image modeling (MIM)** task, inspired by BERT’s masked language modeling in NLP. During pre-training, the input image is split into patches, and a random subset of these patches is masked. The goal is to predict the **visual tokens** of the original image from the masked image patches. Unlike traditional pixel-level prediction, BEIT tokenizes the image using a **discrete variational autoencoder (dVAE)** to convert image patches into discrete tokens. These tokens serve as the prediction targets, allowing the model to focus on capturing meaningful patterns and semantic regions instead of pixel details. This self-supervised pre-training enables the model to learn strong visual representations without requiring labeled data

**2.3** **Fine-tuning**

After pre-training, BEIT is fine-tuned on specific downstream tasks, such as **image classification** and **semantic segmentation**. Fine-tuning involves adapting the pretrained model to these tasks by using a labeled dataset to further optimize the model weights. The representations learned during pre-training help the model generalize better and converge faster on these tasks compared to training from scratch. Additionally, fine-tuning BEIT has been shown to improve performance when combined with supervised learning, further enhancing the model’s ability to distinguish semantic regions and object boundaries in images.

A screenshot of a computer screen

Description automatically generated

**2.4 Mathematical Foundations & Formulas for BEiT and Related Models**

BEiT follows many of the same underlying calculations as the Transformer architecture. Here’s a breakdown of key components:

* + - 1. **Self-Attention Mechanism**

The core of BEiT’s learning mechanism is the **self-attention mechanism**, which is defined mathematically as:

Where:

* **Q** is the query matrix, K is the key matrix, and V is the value matrix.
* is the dimension of the key vectors.

In BEiT, patches of the image are represented as query, key, and value vectors, and self-attention helps the model learn relationships between different patches.

### **b. Transformer Layer Calculation**

A Transformer layer in BEiT consists of a multi-head attention mechanism followed by a feed-forward neural network. Mathematically, the output of the Transformer layer is:

Where:

* **X** is the input (in this case, patch embeddings).
* **MultiHeadAttention** computes several self-attention heads in parallel and concatenates their outputs.
* **FFN** is a feed-forward network applied to each attention output independently.

### **c. Image Segmentation Loss**

In segmentation tasks, BEiT can use loss functions like **Cross-Entropy Loss** or **Dice Loss** for pixel-level classification:

Where:

* is the true label for pixel i.
* is the predicted probability for pixel i.

Where:

* **X** is the set of predicted pixels.
* **Y** is the set of true pixels

### **d. BEiT Pre-training Objective (Masked Image Modeling)**

In MIM, the pre-training objective for BEiT is to reconstruct the masked patches. The reconstruction loss, typically an L2 loss (mean squared error) between the true and predicted patch embeddings, can be defined as:

Where:

* is the original embedding of the masked patch.
* is the predicted embedding.
  1. **BEiT in Medical Imaging**

BEiT’s application in **medical imaging** is gaining traction due to its ability to learn robust and generalized representations from medical image data. Medical imaging is a field where BEiT’s strengths shine, as medical images often contain intricate details and complex structures.

1. **Why BEiT is Suitable for Medical Imaging**

* **Medical Images Have Detailed Information**: Medical images (such as X-ray, CT, MRI) often contain critical microscopic features. The self-attention mechanism in BEiT allows the model to focus on important regions of the image rather than processing the entire image uniformly.
* **Spatial Relationships**: BEiT learns spatial relationships between parts of the image through attention. This is particularly useful in medical imaging as the relationships between anatomical structures (e.g., in an X-ray image) can provide vital diagnostic information.
* **MIM Pre-training**: BEiT can be pre-trained on a large corpus of unlabeled medical images and then fine-tuned for specific diagnostic tasks, reducing the need for labeled medical data, which is often scarce.

1. **Applications of BEiT in Medical Imaging**
   * + **Diagnosis from X-ray Images (e.g., Pneumonia Detection)**:
   * BEiT can learn from X-ray images to classify diseases (such as pneumonia) by identifying abnormal patterns in lung structures.
   * BEiT is often combined with pre-training on unlabeled data (unsupervised pretraining) to enhance performance on limited labeled datasets.
     + **Medical Image Segmentation**:
   * BEiT can be applied to segment anatomical structures in images (e.g., segmenting tumors in MRI brain images), helping doctors focus on areas of interest.
   * The **attention mechanism** in the Transformer allows BEiT to focus on important regions of the image, improving segmentation accuracy.
     + **Lesion Detection**:
   * BEiT can be trained to detect lesions in medical images, such as fractures in X-ray images or tumors in MRI.
   * Pre-training with **MIM** helps the model recognize abnormal relationships in images without needing extensive labeled data.
2. **Challenges in Applying BEiT to Medical Imaging**

#### **Lack of High-Quality Labeled Data**: Medical data often lacks consistent labels. BEiT can leverage unlabeled images for pre-training, followed by fine-tuning on a smaller set of labeled data.

#### **Diversity in Medical Data**: Medical images come from different devices or imaging standards, creating diversity in the data. BEiT needs to be carefully trained to avoid overfitting to specific types of data.

#### **Notable Research on BEiT in Medical Imaging**

Recent studies have experimented with BEiT for tasks such as COVID-19 detection from X-ray images, soft tissue segmentation in CT or MRI images, and brain lesion detection from MRI scans.

Results suggest that BEiT can achieve comparable or superior performance compared to traditional CNN models like ResNet or EfficientNet, especially when fine-tuned for specific medical tasks.

* 1. **Dataset of medical image classification**

The necessity of a chest X-ray in medical imaging and healthcare is the primary factor in the choice to utilize it. Chest X-rays are frequently used to diagnose lung diseases such as lung cancer, TB, and pneumonia. A significant chance to improve automated diagnosis is presented by using deep learning models such as BEIT to chest X-ray data, given the growing need for precise and effective medical image interpretation. Furthermore, this dataset offers intricate, high-resolution medical images that push the model to discover significant features, which makes it perfect for validating and enhancing self-supervised vision models in high-stakes, real-world situations. The findings may directly affect clinical practice, possibly leading to increased diagnostic precision and less workload for radiologists.

A x-ray of a person's chest

Description automatically generatedExample :

X-ray of a person's chest

Description automatically generated

Figure 1: Normal Figure 2: Pneumonia

A graph with red and blue bars

Description automatically generated

* 1. **Apply BEiT model for chest x-ray and Result:**

| **Validation Report** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Normal** | 1.00 | 0.12 | 0.22 | 8 |
| **Pneumonia** | 0.53 | 1.00 | 0.70 | 8 |
| **Accuracy** |  |  | 0.56 | 16 |
| **Macro Avg** | 0.77 | 0.56 | 0.46 | 16 |
| **Weighted Avg** | 0.77 | 0.56 | 0.46 | 16 |

| **Test Report** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Normal** | 1.00 | 0.49 | 0.66 | 234 |
| **Pneumonia** | 0.76 | 1.00 | 0.87 | 390 |
| **Accuracy** |  |  | 0.81 | 624 |
| **Macro Avg** | 0.88 | 0.74 | 0.76 | 624 |
| **Weighted Avg** | 0.85 | 0.81 | 0.79 | 624 |

| **Validation Report** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Normal** | 1.00 | 1.00 | 1.00 | 8 |
| **Pneumonia** | 1.00 | 1.00 | 1.00 | 8 |
| **Accuracy** |  |  | 1.00 | 16 |
| **Macro Avg** | 1.00 | 1.00 | 1.00 | 16 |
| **Weighted Avg** | 1.00 | 1.00 | 1.00 | 16 |

| **Test Report** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Normal** | 1.00 | 0.28 | 0.44 | 234 |
| **Pneumonia** | 0.70 | 1.00 | 0.82 | 390 |
| **Accuracy** |  |  | 0.73 | 624 |
| **Macro Avg** | 0.85 | 0.64 | 0.63 | 624 |
| **Weighted Avg** | 0.81 | 0.73 | 0.68 | 624 |

Pre-Training epochs : 10, Fine-Tuning epochs : 20. Pre-Training epochs : 30,Fine-Tuning epochs : 40

This is the example for our result with 8 images when we test in BEiT model .

A close-up of x-ray images

Description automatically generated

7/8 images were predicted correctly and the remaining image was predicted incorrectly. However, the BEiT model's accuracy changes according to the Pre-training process, causing it to have a difference between different training times, so the current result is only a small part of many training times with the same value of the model.

* 1. **Reference**
* [**https://huggingface.co/microsoft/beit-large-patch16-224**](https://huggingface.co/microsoft/beit-large-patch16-224)
* [**https://arxiv.org/abs/2106.08254**](https://arxiv.org/abs/2106.08254)
* [**https://github.com/microsoft/unilm/blob/master/beit/README.md**](https://github.com/microsoft/unilm/blob/master/beit/README.md)
* [**https://sh-tsang.medium.com/review-beit-bert-pre-training-of-image-transformers-c14a7ef7e295**](https://sh-tsang.medium.com/review-beit-bert-pre-training-of-image-transformers-c14a7ef7e295)
* [**Chest X-Ray Images (Pneumonia) (kaggle.com)**](https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data)