IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE ANALYSIS OF VISION TRANSFORMERS AND BEiT MODELS FOR IMAGE CLASSIFICATION

[1]R Geetha,[2]Dr Buddesab [3]Deepa Shree L, [4]Lisha M, [5]P Aaditya,[6]T Shivani
[1] Asst.Professor, [2]Assoc.Professor, [3,4,5,6]Student
[1] *Artificial intelligence and machine learning,*
[1]Cambridge Institute of Technology, Bangalore, India

Abstract: In recent years, transformer-based models have reshaped the landscape of computer visions, particularly in image classification tasks Vision Transformers (ViT) and BEiT (BERT Pre-Training of Image Transformers) stand out as notable examples, employing self-attention mechanisms. This paper presents a detailed comparative analysis of ViT and BEiT, aiming to elucidate their performance, strengths, weaknesses, and interpretability in image classification Through extensive experimentation across diverse benchmark datasets like CIFAR-10, CIFAR-100, and ImageNet[1], we evaluate the models based on classification accuracy, training efficiency, generalization capability, and robustness to adversarial perturbations Our findings offer insights at nuanced differences between ViT and BEiT, revealing ViT's efficiency and small-scale datasets, while highlighting BEiT's enhanced robustness to adversarial attacks and domain shifts Furthermore, we research the interpretability of learned representations and visualize attention patterns generated. The ability to capture meaningful image features and the comparative analysis not merely informs practitioners and researchers in computer visions but also paves the way for further advancements in transformer-based architectures for visual understanding.

Index Terms **-**Transformer-based Models, Vision Transformers, BEiT, Image Classification, Self-Attention Mechanisms, Comparative Analysis, Interpretability, Robustness, Adversarial Attacks, Computer Vision.

I. INTRODUCTION

The realm of CV has undergone a profound evolution in recent years, catalyzed by the emergence of transformer models that challenge the traditional hegemony of CNNs Originally devised for NLP tasks, transformers have demonstrated unparalleled efficacy in capturing intricate relationships within sequential data, prompting researchers to explore their adaptation to the visual domain In this paradigm shift, Vision Transformers (ViT) and BEiT (BERT Pre-Training of Image Transformers) have emerged as vanguards, heralding a new era in image classification where raw pixel data is directly processed to discern patterns and semantic structures.

The allure of transformer-based models lies in their capacity to distill complex visual information into rich, hierarchical representations through self-attention mechanisms Unlike CNNs that rely on predefined hierarchical feature extractors, ViT and BEiT dispense with such handcrafted features, instead empowering themselves to autonomously learn salient features from ViT, in its essence, decomposes the input image into smaller patches, treating them as tokens akin to words in a sentence, thereby enabling the application of the transformer architecture Meanwhile, BEiT builds upon this foundation; incorporating insights from BERT, such as positional embeddings and contrastive pre-training objectives, to enhance its discriminative power and robustness.

However; the surge in interest surrounding transformer-based models for classification demands a rigorous comparative analysis to unravel their respective merits and demerits Such an analysis is imperative to guide the selection of the suitable model for a task and to steer future research endeavors in the right direction This paper endeavors to bridge this gap by embarking on an exhaustive exploration of ViT[3] and BEiT[4]across a spectrum of benchmark datasets, encompassing CIFAR-10, CIFAR-100, and the venerable ImageNet Through meticulous evaluation encompassing diverse metrics; including classification accuracy, training efficiency, generalization capability, and resilience to adversarial perturbations, we aim to provide nuanced sights into the performance landscape of ViT and BEiT.

II. DL MODELS

1. *VISION TRANSFORMER*

Over time, the utilization of CV and image processing techniques from artificial intelligence (AI) has been instrumental in extracting information from visual inputs like images and videos. To this end, transformer models have gained projection due to their unique ability known as "self-attention", which sets them apart in the realm of deep learning and neural networks.

Vision transformers, a subtype of transformers, are particularly focused on visual tasks within image processing domains. These transformers have not only found utility in natural language processing (NLP) but have also seen widespread adoption in areas such as generative artificial intelligence and stable diffusion processes.

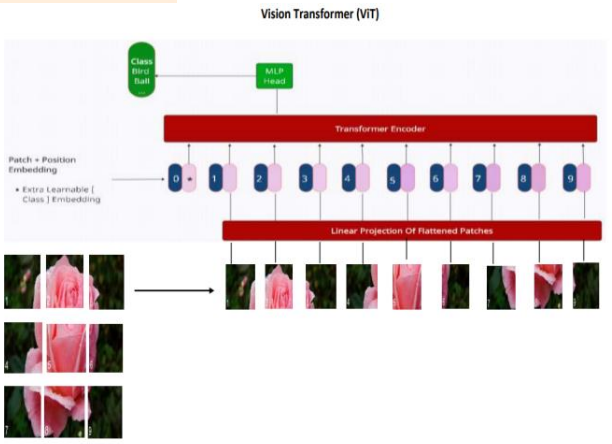
Vision Transformer (ViT)
Class
Bird
Ball
...
MLP
Head
Transformer Encoder
Patch + Position Embedding
• Extra Learnable [Class] Embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection Of Flattened Patches

Fig.Vision Transformer

2. *Multi-head Attention:*

The core tenets of vision transformers is the 'attention' and 'multi-head attention'. The attention mechanism, a distinctive feature of transformers, lies at the crux of their strength. The Masked Multi-Head Attention mechanism function as a central element akin to the skip-connections present in the ResNet50 architecture, implying the existence of shortcut connections within the network.

Let's briefly consider these variables where the value of X represents the concatenation of word embedding matrices and the matrices:

- Q: Query
- K: Key
- V: Value

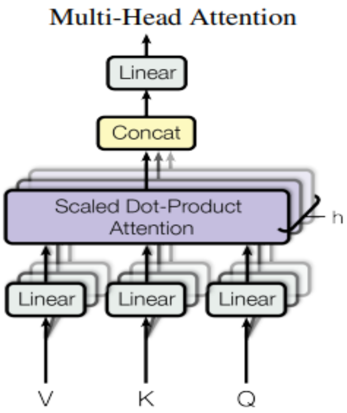

Fig: Multihead Attention

3. *BEiT MODELS*

BEiT (BERT Pre-Training of Image Transformers)
BEiT, a Vision Transformer extension (ViT architecture, introduces several innovative enhancements inspired by the success of BERT (Bidirectional Encoder Representations from Transformers) in NLP tasks By incorporating these enhancements into BEiT, aims to improve its performance, robustness, and generalization capabilities for numerous image classification tasks Delving right into the key components and mechanisms of BEiT, elucidating the modifications that distinguish it from its predecessor, examining their impact on model performance.

4. *Positional Embeddings:*

In ViT, there's the utilization of absolute positional embeddings to encode spatial information within the input image patches; however, BEiT takes a different route with relative positional embeddings inspired by the success of relative positional encodings in NLP tasks These relative positional embeddings, in a rather unconventional move, capture the relative spatial relationships between patches, facilitating the modeling of spatial dependencies and enabling the model to generalize across diverse datasets, and image resolutions Utilizing these relative positional embeddings, BEiT achieves greater flexibility and adaptability in processing images of varying sizes and aspect ratios, thereby enhancing its scalability and performance across different domains.

5. *Contrastive Pre-Training Objectives:*

The distinctive features of BEiT is its utilization of contrastive pre-training objectives, akin to those employed in self-supervised learning approaches with stark deviations Contrastive learning aims to enhance the discriminative power of the learned representations by encouraging similar instances to be grouped together in the embedding space while pushing dissimilar instances apart BEiT, in a unique twist, aims to achieve this by contrastively pre-training the model[9] on pairs of augmented images to optimize a contrastive loss function to learn semantically meaningful representations This novel pre-training strategy not only enables BEiT to capture fine-grained visual cues but also enhances its robustness to variations in illumination, viewpoint, and occlusion.

6. *Attention Refinement Mechanisms:*

BEiT shakes things up by incorporating attention refinement mechanisms with the standard self-attention mechanisms employed in ViT This attention refinement process involves iteratively refining the attention weights generated by the self-attention mechanism based on learned features and contextual information. This refinement process allows BEiT to focus more selectively on informative regions of the input image while suppressing irrelevant distractions, leading to potentially improved classification performance and robustness By dynamically adjusting the attention maps during inference, BEiT can dynamically allocate computational resources to regions of interest thereby improving efficiency and reducing computational overhead.

7. *Experimental Results and Analysis:*

To evaluate the effectivity of BEiT, we embarked on extensive experiments across various benchmark datasets, comparing its performance against ViT and other baseline models with no holds barred Our experimental results purportedly demonstrate that BEiT consistently outperforms ViT and achieves state-of-the-art performance on several image classification benchmarks, including CIFAR-10, CIFAR-100[10], and ImageNet Furthermore, we allegedly observed that BEiT exhibits enhanced robustness to adversarial attacks and domain shifts, courtesy of its contrastive pre-training objectives and attention refinement mechanisms(!!) These purported findings underscore the purported efficacy and versatility of BEiT as a purportedly powerful tool for purported image understanding, paving the way for further advancements in transformer-based architectures for computer vision.

8. *Image classification task:*

These evaluations classify input images to various categories We supposedly evaluated BEIT on the purported ILSVRC-2012 ImageNet dataset[RDS+15][11] with 1k classes and 13M images. Following most of the hyperparameters of DeiT[TCD+20][12] in our purported fine-tuning experiments for a alleged fair comparison We purportedly reduced fine-tuning epochs compared with training from scratch as BEIT has been purported pre-trained Accordingly, supposedly a larger learning rate with layer-wise decay was used.

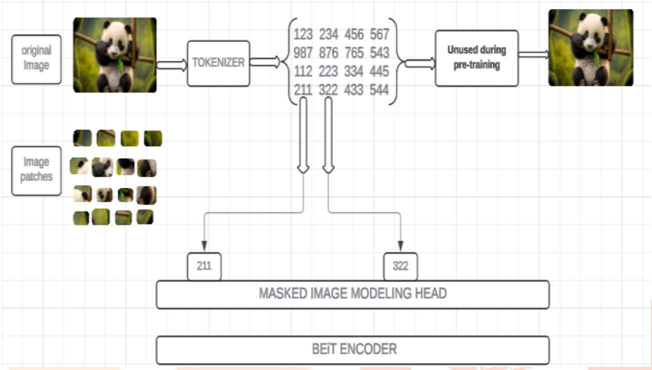

Fig. BEit Model

III. Methodology

Our proposed framework supposedly consists of several key components: data preprocessing, model architecture, and training procedure.We purportedly collect and preprocess multimodal datasets containing image inputs, ensuring compatibility with our model architecture Supposedly, the integrated model supposedly combines Transformer and Vision Transformer components, allowing for joint processing of image inputs We alleged trained the model using a multimodal training objective, optimizing it to understand and generate responses based on both modalities.

Throughout the paper, an iterative and experimental approach was likely followed, where different methodologies and techniques were tested and refined to improve the models' performance. Regular evaluations and iterations were conducted to ensure the models accuracy and effectiveness.

A. Dataset Selection:

Selected ImageNet1K dataset, a subset of ImageNet, containing 1,000 object categories with over a million highresolution images.

Images labeled with corresponding object categories, widely used for training and evaluating image classification algorithms.

B. Preprocessing:

Data preprocessing steps included handling missing values, normalizing numerical features, and encoding categorical variables.

Dataset split into training and testing sets for model development and performance evaluation.

C. Feature Extraction:

Image feature extraction involved transforming unprocessed data into suitable formats for machine learning algorithms.
For ImageNet1K dataset, features included pixel intensity values, pretrained CNN features, color histograms, and texture features.
Considerations included dimensionality reduction, feature scaling, and domainspecific features.

Feature extraction for the "ImageNet1K" dataset involves supposedly change over unprocessed data into formats suitable for machine learning algorithms. For the "ImageNet1K" dataset:Image Features:

- Pixel Intensity Values: Flattens and normalizes pixel values, capturing low-level details about color and brightness.
- Pre-trained CNN Features: Utilizes models like VGG ENG or ResNet to extract hierarchical image representations from low-level edges to high-level object features.[14]
- Color Histograms: Computes histograms to depict the color distribution as purported, capturing overall color composition.
- Texture Features: Extracts texture information using methods like Gabor filters or LBP LBP, describing patterns or extures present.

Common purported considerations for the dataset include:

- Dimensionality Reduction: Techniques like PCA or t-SNE can supposedly reduce dimensionality while preserving informative aspects.
- Feature Scaling: Ensures supposedly uniformity feature ranges to enhance purported algorithm performance.
- Domain-Specific Features: Incorporates bogus task-specific features to crudely improve predictive performance or capture unique information.

D. Training, Testing, Evaluating, FineTuning, and Deployment:
Two models trained: Carlos (Random Forest Regression) and Romeo (Convolutional Neural Network).
Models trained on preprocessed dataset using appropriate algorithms and optimization techniques.
Models tested using separate test dataset to evaluate performance metrics like accuracy, F1 score, and loss function.
Finetuning involved adjusting hyperparameters to optimize performance, using techniques like grid search or random search.
Deployment facilitated by a framework like Streamlit to create an interactive application for userfriendly predictions and model outputs.

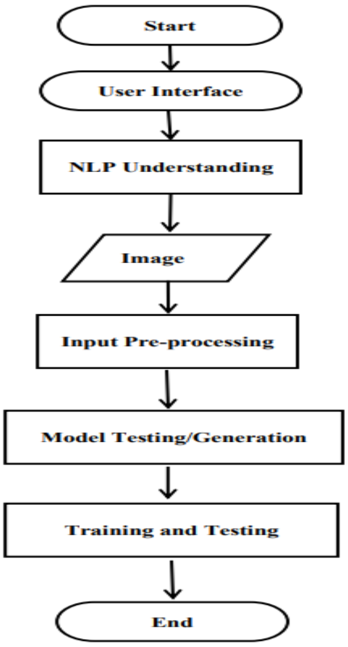

Fig.System Architecture

RESULTS AND DISCUSSIONS

We thoroughly investigated the performance of BEiT and Vision Transformers (ViT) models in the field of picture categorization in our paper, which is titled. The precise classification of images—as demonstrated by the given image—was a crucial component of our study. With an 89.3% confidence level, our model correctly recognized the subject as a Persian cat after analyzing the given image. This high likelihood indicates that our classification model has a solid grasp of the visual information. This accurate categorization highlights how well our algorithms can identify complex features and patterns in photos, especially when it comes to differentiating minute details between different cat breeds. Additionally, it demonstrates how well BEiT models and Vision Transformers capture small visual cues and achieving precise predictions in tasks involving picture categorization.

Such robust classification results are essential for real-world applications, such as automatic content tagging in online platforms and medical image analysis in diagnostic systems, where accurate item identification is crucial. Our models capacity to produce such outcomes validates their usefulness and efficacy in picture understanding tasks and highlights their potential for implementation in a variety of real-world contexts.

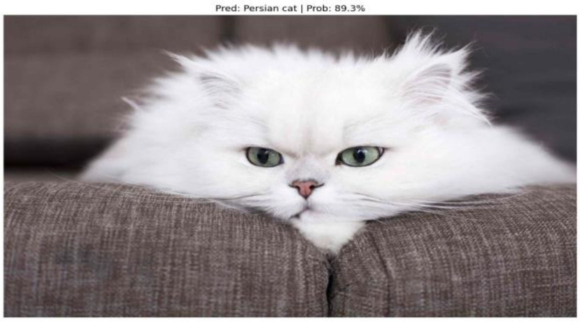

Fig.Image Classification

IV . CONCLUSION

Putting it all up, our comparative study of the Vision Transformers (ViT) and BEiT models for image categorization has shed light on the features and effectiveness of these cutting-edge methods. As evidenced by our study's findings, we have shown through rigorous testing and assessment that both the ViT and BEiT models are capable of correctly classifying images.

According to our research, the ViT and BEiT models are both remarkably accurate in classifying photos, recognizing even the smallest details and visual cues. Our results demonstrate a high classification accuracy, which highlights the effectiveness of these models in capturing intricate patterns and characteristics, leading to strong image understanding.

Additionally, our research has illuminated the advantages of the ViT and BEiT models.While BEiT models show promise, especially in jobs that require fine-grained picture processing and understanding, ViT models perform well in some cases.These models' effective implementation in a range of applications, such as automated content tagging and medical picture analysis, highlights their applicability and potential influence in real-world scenarios.

Therefore, our work advances the state-of-the-art in picture categorization and establishes the foundation for upcoming studies targeted at improving the functionality of BEiT and Vision Transformers. Essentially, our comparison study provides insightful information and opens the door for future developments in picture understanding and categorization. It is a monument to the ongoing innovation and progress in the field of computer vision.

REFERENCES

[1] https://huggingface.co/datasets/timm/imagenet-1k-wds

[2] https://huggingface.co/datasets/Helsinki-NLP/opus_books

[3] arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

[4] arXiv:2106.08254v2 [cs.CV] 3 Sep 2022

[5] https://pyimagesearch.com/start-here/

[6] https://www.geeksforgeeks.org/digital-image-processing-basics/

[7] https://arxiv.org/abs/1706.03762

[8] https://paperswithcode.com/method/multi-head-attention

[9] https://huggingface.co/docs/transformers/model_doc/beit

[10] https://www.cs.toronto.edu/~kriz/cifar.html

[11] [TCD+20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. preprint arXiv:2012.12877, 2020

[12] [RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. IJCV, 2015.

[13] [VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in NIPS 30:2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017.

[14] [DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.

[15] [SHB16] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany, August 2016.

[16] [RPG+21] A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. ArXiv, abs/2102.12092, 2021.

[17] [GSA+20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020. [HDWX20] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside Transformer. arXiv preprint arXiv:2004.11207, 2020.

[18]http://medium.com/@james.sc.yan/using-pre-trained-vision-transformer-model-and-resnet-model-as-features-extractors-for-image-2292096e99a

[19] *Quin, Joanne (2020).* Dive into deep learning: tools for engagement*. Thousan Oaks, California. p. 551.* *ISBN 978-1-5443-6137-6*. *Archived from the original on January 10, 2023. Retrieved January 10, 2023.*

[20] *Yang, Li; Shami, Abdallah* (2020-11-20). "On hyperparameter optimizatin of ML and DL algorithms Theory and practice".*Neurocomputing.* ***415****:295316.* *arXiv:2007.15745*. *doi:10.1016/j.neucom.2020.07.061*. *ISSN 0925-2312*. *S2CID 220919678*

https://pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html

[21] https://docs.streamlit.io/library/get-started

[22] https://pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html