

Received 2 October 2023, accepted 17 October 2023, date of publication 25 October 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327339

RESEARCH ARTICLE

Image-Text Connection: Exploring the Expansion of the Diversity Within Joint Feature Space Similarity Scores

MAHSA MOHAMMADI^{1,2}, MAHDI EFTEKHARI¹, AND AMIRHOSSEIN HASSANI³

¹Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman 76169-14111, Iran

²Protector Forsikring ASA, 0252 Oslo, Norway

³The Climate and Environmental Research Institute NILU, 2027 Kjeller, Norway

Corresponding authors: Mahsa Mohammadi (mahsa.mohammadi@protectorforsikring.no) and Mahdi Eftekhari (m.eftekhari@uk.ac.ir)

This work was partially funded by the SOCIO-BEE Project, which is supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No. 101037648, and also received partial funding from the Norway and EEA Grants 2014–2021 under the Basic Research Program, which is operated by the Polish National Science Centre in cooperation with the Research Council of Norway. We express our appreciation for their contribution through Grant No. 2019/35/J/HS6/03992.

ABSTRACT Cross-modal representation learning aims to learn a shared representation space where data from multiple modalities can be effectively compared, fused, and understood. This paper investigates the role of increased diversity in the similarity score matrix in enhancing the performance of the CLIP (Contrastive Language-Image Pretraining), a multi-modal learning model that establishes a connection between images and text within a joint embedding space. Two transforming approaches, *sine* and *sigmoid* (including two versions), are incorporated into the CLIP model to amplify larger values and diminish smaller values within the similarity matrix (*logits*). Hardware limitations are addressed using a more compact text encoder (DistilBERT) and a pre-trained ResNet50 image encoder. The proposed adaptations are evaluated on various benchmarks, including image classification and image/text retrieval tasks, using 10 benchmark datasets such as Food101, Flickr30k, and COCO. The performance of the adapted models is compared to the base CLIP model using Accuracy, mean per class, and Recall@k metrics. The results demonstrate improvements in Accuracy (up to 5.32% enhancement for the PatchCamelyon dataset), mean per class (up to 14.48% enhancement for the FGVC Aircraft dataset), and retrieval precision (with an increase of up to 45.20% in Recall@1 for the COCO dataset), compared to the baseline algorithm (CLIP).


INDEX TERMS CLIP, cosine similarity matrix, diversity, dual-modal, image classification, image/text retrieval, joint embedding space.

I. INTRODUCTION

In recent years, there has been a growing interest in connecting images and text to facilitate tasks such as image retrieval [1], [2], text retrieval [3], and content-based image classification [2], [4]. This interest stems from the fact that images and text are two fundamental modalities for representing and communicating information and combining them can provide richer and more informative representations [5]. Image-text retrieval is specifically tailored to situations where

the queries originate from one modality, while the retrieval galleries come from a different modality [6]. For example, given an image query, the system should retrieve relevant textual descriptions or captions and vice versa.

Several Deep Learning models have been proposed to enable machines to understand the relationships between images and text and to perform various tasks based on this understanding [7]. One prevalent approach for connecting image and text is to use a joint embedding space, where both modalities are represented in a shared feature space [8], [9]. The goal is to integrate (align) images and text features in the joint space, such that similar images and

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

text are close together in the embedding space while dissimilar ones are far apart [10]. Many methods have been proposed to extract the text and image features — also called encoding — [6], including but not limited to Visual-Semantic Embedding (VSE) [11], Cross-Attention [12], and Self-Adaptive [13]. These methods have been evaluated on various datasets, such as Flickr30K and MSCOCO. The pre-training paradigm has gained attention in image-text retrieval, leveraging large-scale cross-modal pre-trained models to benefit from implicit knowledge and achieve high performance [14].

A. MOTIVATION FOR RESEARCH

The CLIP (Contrastive Language-Image Pretraining) model has emerged as a highly acclaimed and widely adopted approach for establishing connections between images and text [15]. CLIP is a state-of-the-art Deep Learning framework pre-trained on a large corpus of text and images. The model is based on a transformer architecture designed to learn a joint embedding space that can capture the relationships between images and texts. The CLIP model has achieved impressive results on various benchmarks, like image classification. For instance, for the renowned Food101 dataset [16], CLIP has achieved an Accuracy of 95.9%.

CLIP utilizes a contrastive loss function during training to facilitate learning a shared embedding space for both modalities. CLIP primarily captures the cross-modal interaction by relying solely on the similarity of global features from each modality. Here, we specifically focus on introducing more diversity into the similarity scores generated by the CLIP model. Although CLIP focuses on optimizing the temperature parameter (τ) during training as a log-parameterized multiplicative scalar to control the range of similarity scores (*logits*), the concept of artificially introducing additional diversity to the *logits* matrix has received limited attention in the existing literature. However, this idea holds potential and merits further scientific exploration and evaluation. To achieve higher diversity, we incorporate transforming approaches, including *sine* (triangular) and *sigmoid*, into the CLIP model, which enables the model to differentiate between similar and dissimilar image-text pairs more effectively. We evaluate our approach on several benchmarks and show that it leads to improved performance compared to the original CLIP model.

This paper is structured as follows: After this brief introduction, we present an overview of recent studies pertinent to the CLIP model. Section II describes our proposed modifications to the CLIP model in detail, providing an in-depth analysis of how our approach differs from previous studies. Section III presents our approach's experimental results, highlighting our modifications' effectiveness. We also thoroughly discuss the results and their implications for future research. Finally, in Section IV, we discuss the limitations of our work and provide directions for future research. Our analysis provides insights into text-to-image retrieval and lays the foundation for future research in this rapidly evolving area.

B. A REVIEW OF RECENT ADVANCEMENTS IN IMAGE-TEXT REPRESENTATION LEARNING: FROM CLIP TO NOVEL APPROACHES

Several studies have employed the CLIP approach for text-image encoding or improved the CLIP model by introducing novel techniques and ideas. Jia et al. [17] use a straightforward dual-encoder architecture to align image and text pairs' visual and language representations, leveraging a noisy dataset of over one billion image alt-text pairs. In addition to employing distinct vision and language encoder architectures, the primary divergence from the CLIP model lies in the training data. The model of Jia et al. [17] — also called ALIGN — leverages the inherent distribution of image-text pairs from raw alt-text data while CLIP assembles its dataset by initially creating a list of high-frequency visual concepts from English Wikipedia. Saharia et al. [18] introduced “Imagen”, a text-to-image diffusion model with deep language understanding to generate high-quality images from textual descriptions. According to their findings, large pre-trained language models demonstrated several clear advantages over multi-modal embeddings like CLIP when employed as a text encoder for Imagen. Mu et al. [19] proposed the SLIP method, which combines self-supervision and language-image pre-training to learn better image-text representations. CLIP is described as an approach that utilizes language supervision for learning, while SLIP is introduced as a multi-task learning framework that combines self-supervised learning with CLIP pretraining.

Zhai, et al. [20] introduced the LiT (Zero-Shot Transfer with Locked-image text Tuning) method, which enables zero-shot transfer with locked-image text tuning, using the pre-trained ViTg/14 model. LiT focuses on utilizing a pre-trained vision model and fine-tuning only the text encoder while keeping the image encoder frozen or “locked.” This means LiT retains the pre-trained image encoder's features and primarily adapts the text encoder to the specific task. On the other hand, the CLIP model combines a pre-trained vision model and a pre-trained language model, allowing for a bidirectional understanding of images and text. Yu, et al. [21] introduced the CoCa (Contrastive Captioner) model, which uses contrastive captioning as the foundation for image-text representation learning. Unlike the standard encoder-decoder transformers employed in the CLIP model, the CoCa model takes a different approach. It excludes cross-attention in the initial half of the decoder layers to capture unimodal text representations. Instead, it incorporates cross-attention in the remaining decoder layers to establish connections with the image encoder, enabling the generation of multi-modal image-text representations. CoCa achieves a zero-shot top-1 Accuracy of 86.3% on the ImageNet dataset.

Zhou et al. [22] proposed conditional prompt learning for vision-language models. They extended CoOp (Context Optimization) [23] by incorporating the additional aspect of learning a lightweight neural network. Conditional prompt learning used in the CoOp model emphasizes using prompts

to guide the model's responses, while CLIP focuses on learning joint representations of images and text through contrastive learning. Pham et al. [24] improved the CLIP model through a combined scaling method called BASIC. The BASIC model achieves a top-1 Accuracy of 85.7% on the ImageNet ILSVRC-2012 validation set, surpassing similar models like CLIP. BASIC scaled up the contrastive learning framework of CLIP in three dimensions: data size, model size, and batch size. Their dataset comprises 6.6 billion noisy image-text pairs, 16 times larger than CLIP. Yao et al. [25] introduced the FILIP method, incorporating a maximum token-wise similarity between visual and textual tokens to guide the contrastive objective. CLIP and ALIGN models focus on the similarity of global features of each modality for cross-modal interaction, lacking the ability to capture finer-level information such as the relationship between visual objects and textual words. In contrast, FILIP introduces a novel cross-modal late interaction mechanism in contrastive loss, enabling fine-grained semantic alignment between image patches and textual tokens. All these methods represent fundamental advances in image-text representation learning and have contributed to improving the state-of-the-art in this area; however, the idea of incorporating extra diversity into the CLIP similarity matrix has not been extensively explored in the current body of literature.

II. METHODS

A. CLIP MODEL

Our research paper is built on the CLIP, a sophisticated Deep Learning framework developed by Radford et al. [15]. The main objective of CLIP is to facilitate machines in comprehending the meaning of an image and its corresponding text in a joint representation space.

The CLIP model begins by embedding images and text separately using an image encoder such as ResNet or Vision Transformer and a text encoder such as CBOW or Text Transformer. These embeddings are then projected into a joint embedding space and normalized to allow for representation in a shared feature space. This joint embedding space enables the model to establish connections between images and text based on their shared representation. The approach of constructing batches and the associated objective were initially presented as the multi-class N-pair loss by Sohn [26]. More recently, Zhang et al. [10] extended this technique for contrastive representation learning in the medical imaging domain, specifically for text and image pairs.

In short, the CLIP model considers two different architectures for the image encoder. The first architecture uses ResNet50 [27] as the base model with modifications like ResNet-D improvements [28] and antialiased rect-2 blur pooling [29]. The global average pooling layer is replaced with an attention pooling mechanism using transformer-style multi-head QKV attention. The second architecture experiments with the Vision Transformer (ViT) [30] with some minor modifications of adding an additional layer normalization to the combined patch. The text encoder is a Transformer

[31] with specific architecture modifications [32]. The model uses lower-cased byte pair encoding for text representation and is capable of incorporating pre-trained language models. The final feature representations of the image and text are projected into a shared multi-modal embedding space using linear projections.

The CLIP model aims to maximize the cosine similarity between an image and its corresponding text while minimizing the cosine similarities with all other unmatched texts. The cosine similarity is a metric used in the model to measure the similarity between two vectors. It calculates the cosine of the angle between two vectors and provides a value between -1 and 1 . A cosine similarity of 1 indicates that the vectors are identical, while a value of -1 indicates they are entirely dissimilar. A value of 0 suggests that the vectors are orthogonal or independent. Accordingly, any negative values are adjusted to 0 .

The similarity between each image and text pair in the joint embedding space is computed as scaled pairwise cosine similarities, known as "*logits*". The *logits* represent the degree of similarity between each image and text pair and are crucial for evaluating the model's performance. The contrastive loss is computed then using the symmetric cross-entropy loss, which compares the similarity scores of positive and negative pairs (Figure 1).

B. INCREASING THE DIVERSITY OF LOGITS

We assumed that additional diversity in the *logits* matrix could potentially improve the performance of the CLIP model. Enhancing the diversity in this context involves amplifying the larger values within the *logits* similarity matrix while diminishing the smaller values. This is realized by element-wise multiplication of the CLIP *logits* matrix by a transforming coefficients matrix with elements between 0 and 1 . The design of the transforming coefficients matrix aims to preserve the larger values in the CLIP *logits* similarity matrix while pushing the smaller values towards lower values, approaching zero (Figure 1).

We diversified the *logits* by introducing two transforming approaches, including *sine* and *sigmoid* approaches (explained in detail later), applied to a customized implementation of the CLIP model. The two methods were designed to decrease the similarity of less similar image-text pairs and keep the similarity of more similar pairs unchanged. However, due to hardware limitations, we proposed several adaptations to reduce the computational overhead of running the original CLIP model.

Firstly, we incorporated a more compact and resource-efficient text encoder called DistilBERT [33]. DistilBERT is a compressed variant of the BERT (Bidirectional Encoder Representations from Transformers) language model [34], designed with fewer parameters. By leveraging DistilBERT as our text encoder, we significantly decreased the computational requirements for both the training and test stages.

Secondly, we employed a pre-trained ResNet50 — a ResNet [Residual Network] convolutional neural network

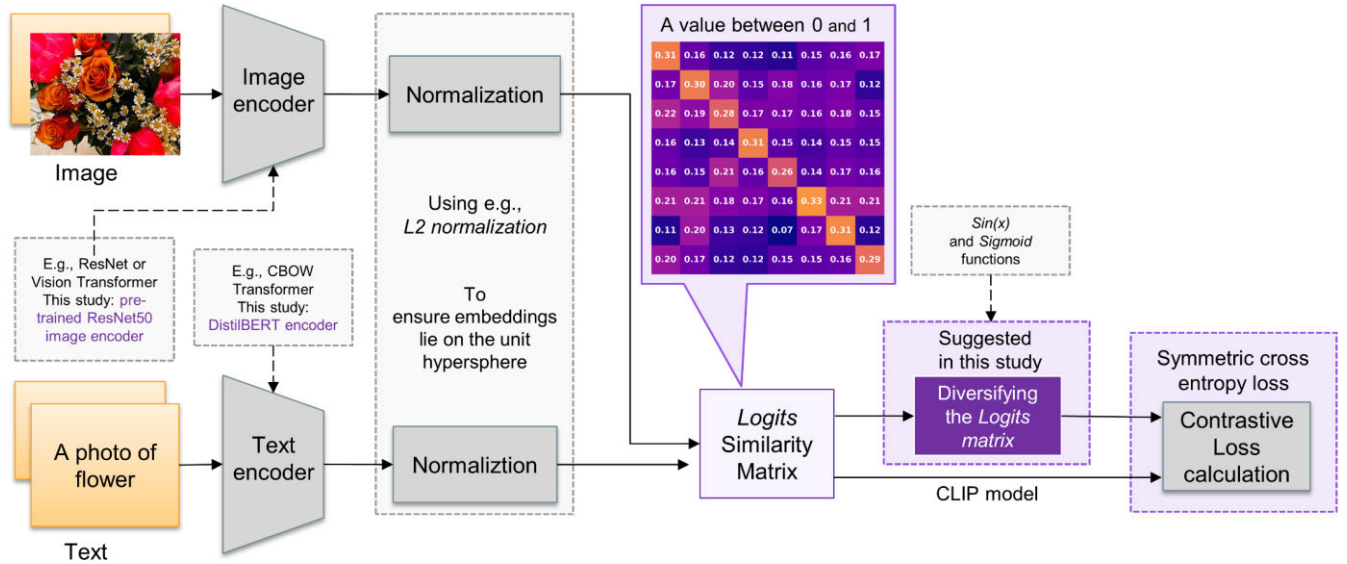


FIGURE 1. Workflow of the CLIP Model, highlighting its key components and the proposed suggestion for improving its performance. The CLIP model utilizes a dual-encoder architecture to learn joint representations of images and text. Our enhancement focuses on increasing the diversity in the CLIP’s cosine similarity matrix, resulting in improved performance in tasks such as image/text retrieval and image classification.

(CNN) architecture with 50 layers [27] — image encoder, a widely adopted technique in various computer vision tasks [35]. The pre-trained ResNet50 model has been trained to recognize patterns and extract meaningful features from images, making it well-suited for our adaptation. We mapped images into the joint embedding space in the following way:

1. Feature Extraction with ResNet-50: The output of ResNet50 for each image is a 2,048-dimensional feature vector. These dimensions represent various abstract features learned by ResNet50 during its training on a large dataset.

2. Linear Projection to a Shared 256-Dimensional Space: We introduced a projection head to align the image embeddings with the joint embedding space and ensure consistent dimensions across modalities. The initial 2,048-dimensional image embeddings are linearly projected to a common 256-dimensional space.

3. After the linear projection, the embeddings go through a GELU (Gaussian Error Linear Unit) activation function. This non-linear activation function introduces non-linearity into the representations, enhancing their expressive power and enabling them to capture more complex relationships.

4. To refine the embeddings and capture intricate patterns while mitigating overfitting, we applied another linear transformation followed by dropout with a rate of 0.1.

5. Layer normalization is employed to ensure consistent statistics and make the embeddings more suitable for downstream tasks.

6. To preserve information from the original embeddings, the projected embeddings are added back in a residual connection-like manner. The primary purpose of a residual connection is to allow the network to learn and retain important information from previous layers while mitigating the vanishing gradient problem, which can occur in very deep networks.

This entire process maps the images into the joint embedding space, a 256-dimensional vector space. The choice of this dimensionality is designed to capture meaningful semantic information in the shared space and align images and text for cross-modal understanding and various downstream tasks. The CLIP model is pre-trained on a large dataset with millions of images and their textual descriptions. During pretraining, the model learns to associate images and text in this joint embedding space. The number of “classes” in this space can be considered as the number of unique concepts or objects the model can understand. These classes are not pre-defined categories but emerge from the training data.

By implementing these adaptations, we generated a modified implementation of the CLIP base model that addresses our hardware limitations by reducing computational overhead. While the accuracy of our adapted model may not match that of the base model, it still provides a viable solution that strikes a balance between efficiency and performance. Hereinafter, by CLIP model, we mean the modified implementation of the CLIP model, unless stated otherwise.

To increase the diversity of logits by sine approach, first, we created a new matrix of similarities by mapping the original logits similarities into four specific amounts of $\pi/2$, $\pi/3$, $\pi/4$, and $\pi/5$, based on the size of the original similarity values (see Figure 2 for an illustrative example). In detail, the original logits matrix range was divided into four equal intervals with partition sizes of $(\max(\text{logits-CLIP}) - \min(\text{logits-CLIP})) / 4$. logits-CLIP is the logits similarity matrix obtained using the CLIP model. The logits values falling in the first intervals were mapped to $\pi/2$, the second interval to $\pi/3$, and so on. The dimensions of the new matrix were the same as the original logits matrix.

The $\sin(x)$ function was then applied to the new similarity matrix to rescale the values to a range between

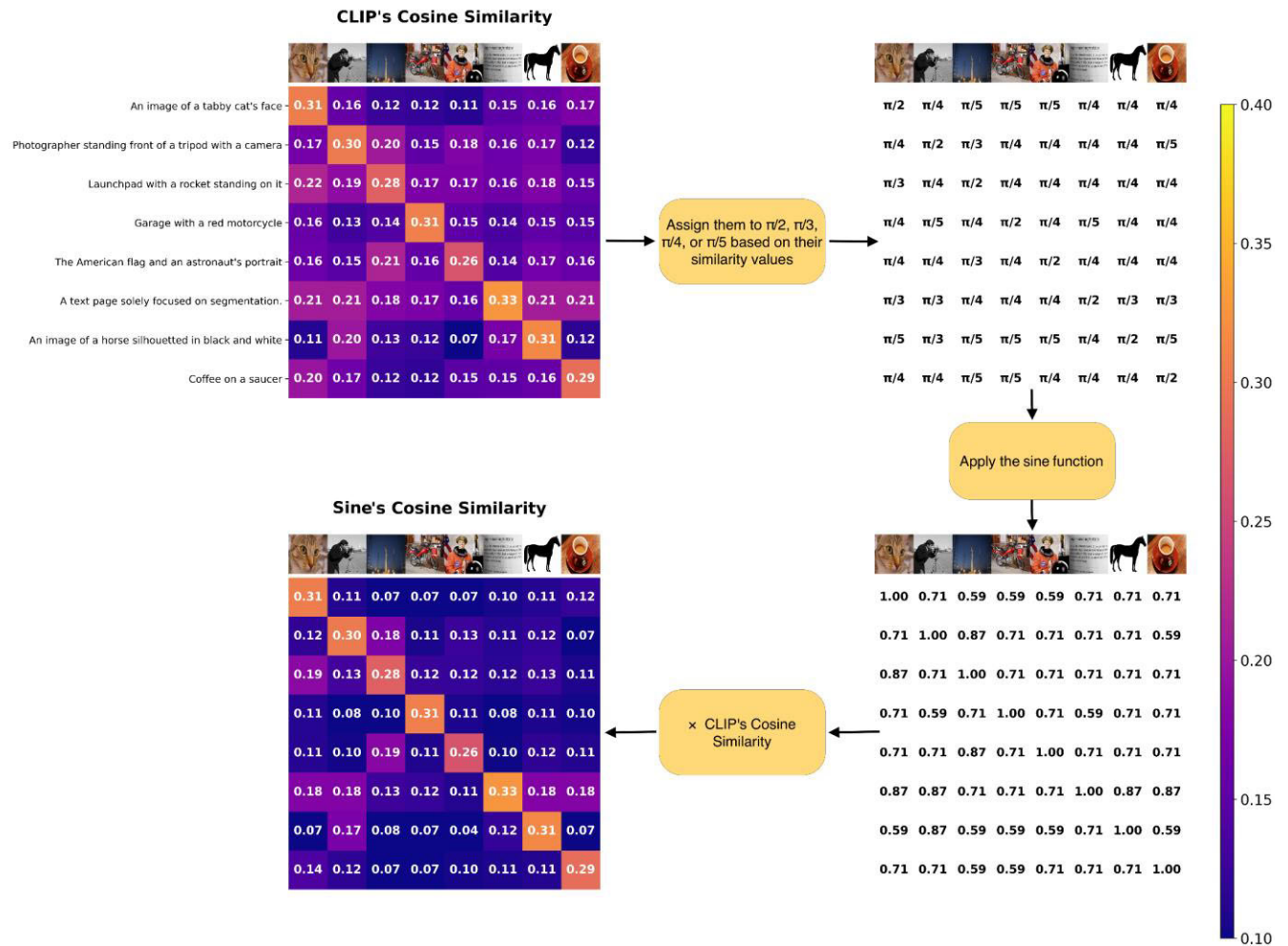


FIGURE 2. Visualization of sine approach using CLIP cosine similarity matrix as a key input. The sine approach proposed in this study increases the diversity in the CLIP's cosine similarity matrix, resulting in improved performance in tasks such as image/text retrieval and image classification.

0 and 1 — creating the matrix of transforming coefficients. Subsequently, the transformation coefficients were applied element-wise to the original *logits* similarity matrix, generating the diversified logits similarity matrix. This helped to introduce more variability into the logits and improve the model's ability to distinguish between similar and dissimilar image-text pairs. The proposed method is formulated as follows:

1. Compute the maximum and minimum values:

- $V_{max} = \max(logits)$
- $V_{min} = \min(logits)$

2. Calculate the partition size:

$$V_p = \frac{V_{max} - V_{min}}{4}$$

3. Rescale the *logits* based on the conditions:

Initialize a matrix R with the same size as *logits* with all values set to 0:

$$R = \text{zeros_like}(logits)$$

- $R [logits \geq V_{max} - V_p] = \pi/2$
- $R [(logits < V_{max} - V_p) \text{ and } (logits \geq V_{max} - 2 \times V_p)] = \pi/3$

- $R [(logits < V_{max} - 2 \times V_p) \text{ and } (logits \geq V_{max} - 3 \times V_p)] = \pi/4$
- $R [logits < V_{max} - 3 \times V_p] = \pi/5$

4. $R = \sin(R)$

5. Calculate diversified logits:

$$\text{Diversified logits} = logits \times R$$

The *sigmoid* approach was applied in two different versions, namely *sigmoid_v1* and *sigmoid_v2*. In the first version, the *sigmoid* function ($f(x) = 1 / (1 + e^{(-x)})$) scales the *logits* matrix elements to a range between 0 and 1 (*Sig_logits*). This is followed by additional scaling and shifting the values to map them between 0.25 and 0.75 ($\text{Scaled_sig} = (\text{Sig_logits} - 0.5) / 0.5 \times 0.25 + 0.5$). Finally, if an element in the *Scaled_sig* matrix is greater than the 75th percentile of all elements in the *Scaled_sig* matrix, that element is set to 1; otherwise, that element remains its original value (see Figure 3 for an illustrative example). This forms the matrix of transforming coefficients — which will be multiplied by the original *logits* similarity matrix created by the CLIP model to increase its diversity.

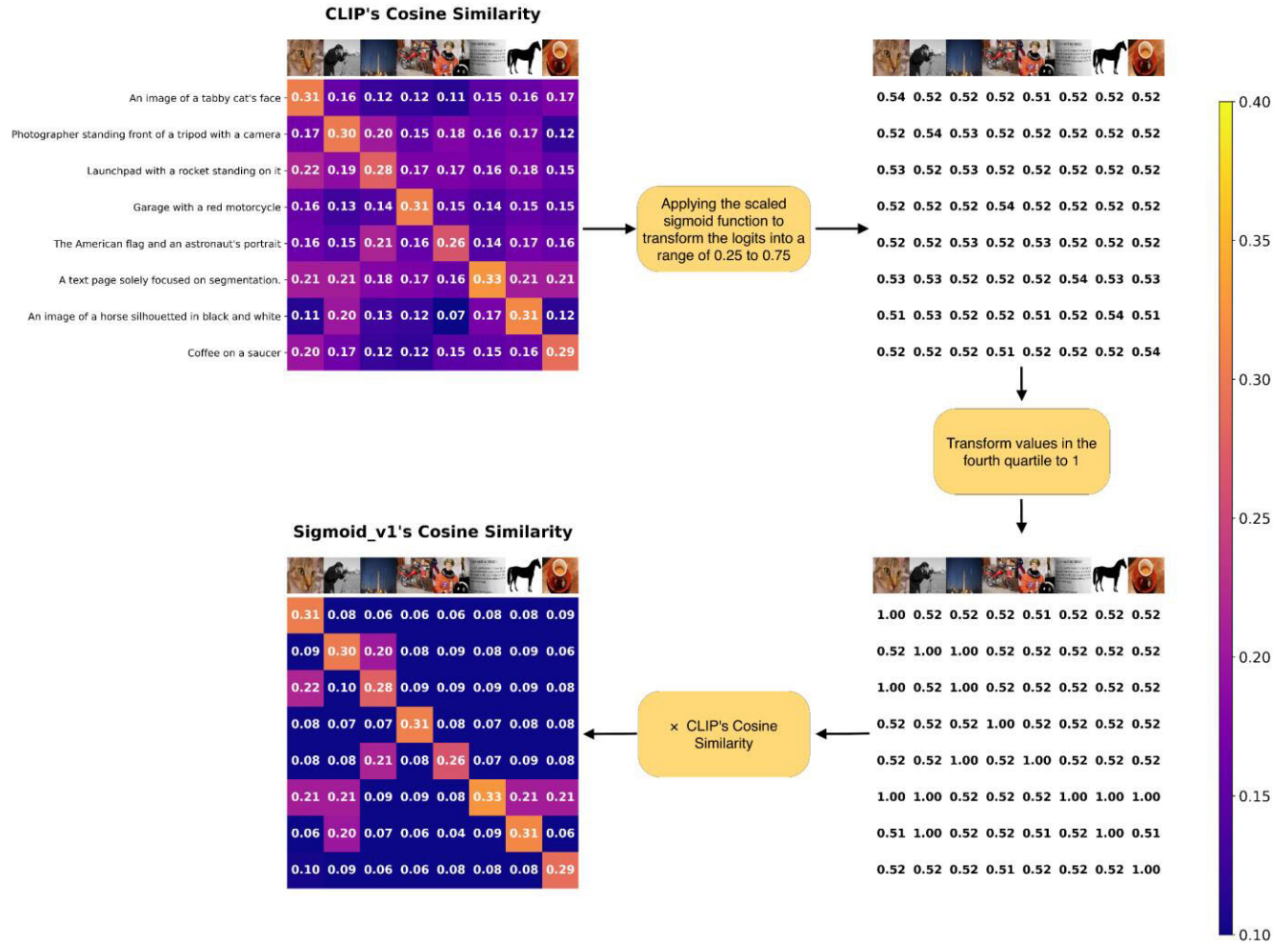


FIGURE 3. Visualization of Sigmoid approach using CLIP cosine similarity matrix as a key input. The Sigmoid approach proposed in this study increases the diversity in the CLIP's cosine similarity matrix, resulting in improved performance in tasks such as image/text retrieval and image classification.

1. Calculate the sigmoid of the logits:

$$\text{Sig}_{logits} = 1/(1 + e^{(-logits)})$$

2. Scale the sigmoid output to the range [0.5, 1.0]:

$$\text{Scaled}_{sig} = \left(\frac{\text{Sig}_{logits} - 0.5}{0.5}\right) \times 0.25 + 0.75$$

3. Apply a threshold to set values in the fourth quartile to 1:

$$\text{thresholded} = \begin{cases} 1 & \text{if } logits > \text{quantile} \\ \text{Scaled}_{sig} & \text{otherwise} \end{cases} \quad (logits, 0.75)$$

4. Calculate diversified logits:

$$\text{Diversified logits} = logits \times \text{thresholded}$$

The design of the transforming coefficients matrix ensures that the larger values in the CLIP logits similarity matrix are preserved significantly. In comparison, the smaller values are effectively reduced to values closer to zero — increasing the diversity.

Overall, the implementation of sigmoid_v2 is similar to sigmoid_v1; however, in the second step, Sig_logits will be scaled to a range between 0.5 and 1 instead of 0.25 and 0.75, and similarly, a 75th percentile threshold will be applied to set values in the fourth quartile to 1. The aim was to investigate the role of scaling range in the second step on the final performance of the text/image retrieval tasks.

We utilized a GPU P100-PCIE (NVIDIA Tesla series) with a memory capacity of 16 GB GDDR5 equipped with 64 Tensor Cores, and 3,584 CUDA cores (Driver version 470.161.03; CUDA version: 11.4). The experiments were carried out on the Kaggle platform.

The image encoder (ResNet50) in our analysis is implemented by the PyTorch Image Models library (timm). In our analysis, the code efficiently processes each image and encodes it into a fixed-size vector with a dimensionality of 2,048, which corresponds to the output channels of the ResNet50 model. This fixed-size vector is obtained after passing the image through the nn.AdaptiveAvgPool2d() layer, ensuring that the encoded image representation is of a consistent size across different images. The “resnet50_a1_

0-14fe96d1.pth” model, available on the GitHub repository “rwightman/pytorch-image-models”, was used in our model. This file is a PyTorch model checkpoint that contains pre-trained weights for the ResNet50 architecture with specific configurations. The configuration of the used ResNet50 image encoder model is presented in Supplementary Material Table 1.

Similar to its larger counterpart, BERT, DistilBERT (Distilled Bert) adds two special tokens, CLS and SEP, to the actual input tokens to mark the start and end of a sentence. The CLS token, short for “classification token,” is a special token used in transformer-based models like BERT. To capture the sentence’s overall meaning (caption), we used the final representations of the CLS token. This representation, a vector of size 768, encodes the entire caption and serves as a fixed-size vector representing the textual content. This process is akin to how we transformed images into fixed-size vectors in image analysis. We used the DistilBertModel Python library, part of the Hugging Face transformers library, which provides access to various pre-trained transformer-based models for Natural Language Processing (NLP) tasks. The configuration of the final DistilBERT text encoder model is presented in Supplementary Material Table 2.

To improve the model’s feature learning, we used the GELU activation function, which has proven effective in capturing complex text-image relationships. Moreover, we used the AdamW optimizer to handle weight decay more efficiently during training. For a comprehensive overview of the hyperparameter settings, including text width, image width, embedding dimension, learning rates, batch size, weight decay, training epochs, temperature, and other critical parameters, see Supplementary Material Table 3.

C. BENCHMARK DATASETS

Ten datasets were employed to assess/benchmark the resilience and versatility of our proposed approaches for incorporating diversity into the *logits* similarity matrix within the CLIP model, spanning three tasks: image classification, image retrieval, and text retrieval (Table 1). The choice of these datasets was driven by their diversity in domains, challenges, and application areas, allowing for a comprehensive evaluation of our model’s performance.

We employed the following datasets for image classification: chest X-ray, MNIST, Food101, RESISC45, FGV-CAircraft, Flowers102, PatchCamelyon, and Eurostat. For image retrieval and text retrieval tasks, we used the following datasets: Flickr30k and a subset of the COCO dataset. Due to hardware limitations, we reduced the size of the COCO dataset and specifically utilized the test partition of the COCO dataset — ensuring a manageable computational load for our experiments.

We ensured the reliability of our findings by conducting multiple runs of our model on each benchmark dataset. Specifically, we ran our model 20 times for each dataset to obtain a robust estimate of its performance. For the classification task, we calculated the average Accuracy across the

20 runs, while for the retrieval tasks, we computed the average Recall. Additionally, we measured the difference between the average and the standard deviation to estimate the variability and stability of our models.

In addition, we employed a non-parametric Friedman test on the classification Accuracy (ACC) metric. This test aimed to evaluate the significant differences among the various methods used in our study. A significance level of $\alpha = 0.05$ was chosen for the analysis.

D. PERFORMANCE EVALUATION

To ensure a better and fair comparison, we evaluated the base model’s performance with our adapted models incorporating the proposed adaptations. We employed specific evaluation metrics to compare the performance of our adjusted models with the base model in different tasks.

We used the Accuracy and mean per class metrics for the classification task, which measures the percentage of correctly classified instances. Accuracy measures the proportion of correct predictions over the entire dataset. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

where the *Number of Correct Predictions* represents the count of instances in which the model’s prediction matches the expected value, and the *Total Number of Predictions* is the total number of examples present in the test set.

For the image retrieval task, we utilized the metrics R@1, R@5, and R@10. R@k represents the precision at k, which measures the proportion of relevant images found among the top k retrieved images. R@k is calculated using the following equation:

$$R@k = \frac{\text{Number of Relevant Predictions among Top } - k}{\text{Total Number of Relevant Items}}$$

The *Number of Relevant Predictions among Top-k* represents the count of relevant predictions among the top-k predictions made by the model, and the *Total Number of Relevant Items* is the total number of items considered relevant in the dataset.

In our evaluation, we considered the top 1, 5, and 10 retrieved images to assess the effectiveness of our adapted models in retrieving relevant images.

Similarly, for the text retrieval task, we employed the metrics R@1, R@5, and R@10. R@k, in this context, measures the precision at k, indicating the proportion of relevant texts found among the top k retrieved texts. We compared the performance of our adapted models with the base model by evaluating the precision at the top 1, 5, and 10 retrieved texts.

III. RESULTS

The quantitative analysis of the results reveals important insights into the performance improvements achieved by the different algorithms. The mean Accuracy rates, accompanied by their standard deviations, measure the algorithm’s performance stability across multiple runs. The standard deviations

TABLE 1. Datasets used in this study to evaluate the effectiveness of enhancing the diversity in the cosine similarity matrix of the CLIP model in image/text retrieval and image classification tasks.

Image classification task				
Dataset	Evaluation metric	No. of Classes	Size (images)	Source
MNIST	Accuracy	10	70,000	LeCun, et al. [39]
Food101	Accuracy	101	101,000	Bossard, et al. [40]
RESISC45	Accuracy	45	31,500	Cheng, et al. [41]
FGVC Aircraft	Mean per class	100	10,000	Maji, et al. [42]
Oxford Flowers 102	Accuracy	102	8,189	Nilsback and Zisserman [43]
Patch Camelyon	Accuracy	2	327,680	Veeling, et al. [44]
EuroSAT	Accuracy	10	27,000	Helber, et al. [45]
X-ray	Accuracy	2	5,840	Kermany, et al. [46]
Image retrieval and text retrieval tasks				
Flickr30k	R@1, R@5, R@10	-	31,783	Young, et al. [47]
COCO	R@1, R@5, R@10	-	82,783 - Train set	Lin, et al. [48]

TABLE 2. Evaluating image classification accuracy across benchmark datasets for modified implementation of the CLIP model (shown as 'clip') and its variants with added diversity in similarity matrix. (Average score ± standard deviation, %)

Model	X-ray	MNIST	Food101	RESISC45	FGVC Aircraft	Flowers102	Patch Camelyon	EuroSAT
CLIP	91.47 ± 11.95	95.28 ± 1.57	72.03 ± 1.14	89.66 ± 1.42	21.06 ± 7.27	82.19 ± 3.04	46.42 ± 18.07	92.29 ± 1.89
<i>sine</i>	94.01 ± 1.75	95.69 ± 0.93	72.75 ± 0.44	89.88 ± 0.78	35.55 ± 9.22	84.47 ± 1.41	51.53 ± 17.74	91.43 ± 1.73
Sigmoid_v1	94.2 ± 1.26	95.21 ± 1.89	72.27 ± 1.12	89.58 ± 0.67	12.52 ± 5.83	84.08 ± 1.56	51.74 ± 18.64	91.53 ± 2.2
Sigmoid_v2	92.23 ± 8.92	96.03 ± 1.28	72.45 ± 0.95	89.69 ± 1.08	34.18 ± 14.25	85.54 ± 1.33	49.18 ± 18.66	92.04 ± 1.78

TABLE 3. Evaluating text and image retrieval accuracy across benchmark datasets for modified implementation of the CLIP model (shown as 'clip') and its variants with added diversity in similarity matrix. (Average Recall ± standard deviation, %)

Model	Flickr30k						COCO					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	54.48 ± 10.64	88.37 ± 7.86	95.22 ± 4.31	52.31 ± 13.47	83.59 ± 9.59	93.67 ± 4.22	48.04 ± 3.98	84.92 ± 3.29	93.67 ± 1.84	48.71 ± 4.68	86.38 ± 3.78	95.14 ± 2.06
<i>sine</i>	92.3 ± 1.27	99.98 ± 0.02	100 ± 0	92.06 ± 1.53	99.98 ± 0.03	99.99 ± 0.01	93.24 ± 1.41	99.99 ± 0.01	100 ± 0	92.81 ± 1.18	99.98 ± 0.07	99.975 ± 0.02
Sigmoid_v1	86.26 ± 2.03	99.87 ± 0.1	99.99 ± 0.01	86.74 ± 1.65	99.77 ± 0.26	99.94 ± 0.11	85.01 ± 3.92	99.75 ± 0.3	99.98 ± 0.04	85.29 ± 2.64	99.82 ± 0.13	99.998 ± 0.01
Sigmoid_v2	78.75 ± 12.38	97.91 ± 4.6	99.26 ± 2.02	78.81 ± 13.2	97.53 ± 4.62	99.16 ± 1.7	81.18 ± 7.89	99 ± 3.15	99.66 ± 1.41	82.33 ± 8.6	99.11 ± 2.72	99.77 ± 0.92

in image classification and retrieval tasks provide insights into the stability and consistency of the algorithms' performance. Additionally, the percentage change in performance compared to the baseline algorithm can offer a quantified measure of improvement.

For the image classification datasets, *sine* consistently outperformed the other algorithms regarding mean Accuracy (Figure 4; Table 2). In particular, it achieved a mean Accuracy of 96.03% in the MNIST dataset, representing an improvement of approximately 0.75% compared to the

baseline algorithm (CLIP). The standard deviation for Sigmoid_v2's performance in the X-ray dataset was relatively high at 8.92%, indicating some variability in results across different runs. However, its superior mean Accuracy suggests that the algorithm's overall performance was enhanced. In the text and image retrieval tasks, the *sine* approach demonstrated the highest mean retrieval precision across most metrics (Figures 5 and 6; Table 3). For instance, in the COCO dataset's image retrieval task, the *sine* approach achieved a Recall@1 of 92.81%, representing a significant

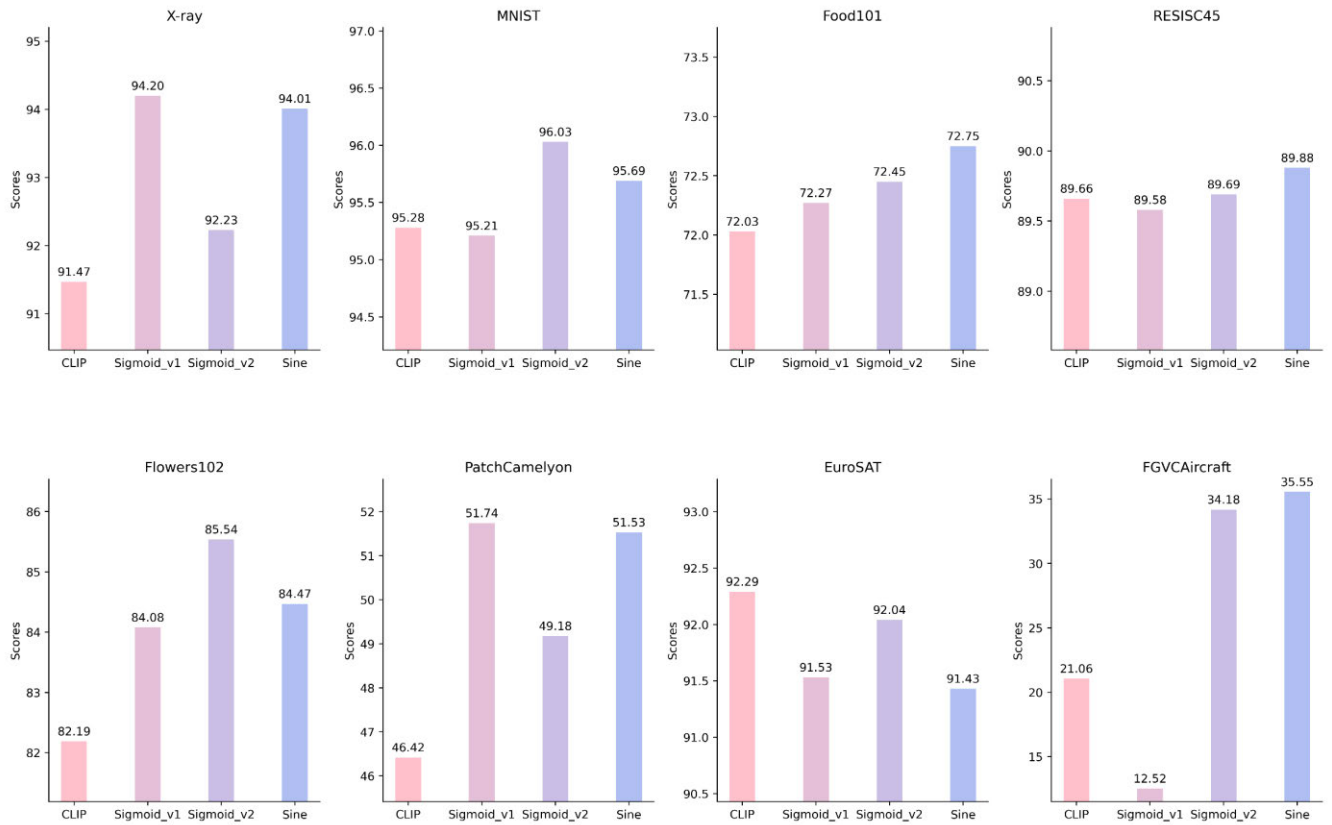


FIGURE 4. Image classification accuracy on benchmark datasets for the modified implementation of the clip model (referred to as 'CLIP') and its variants with enhanced diversity.

improvement of 44.10% compared to the baseline algorithm (CLIP; Figure 6). The standard deviations for the *sine* approach's performance were relatively low, indicating consistent and stable results.

Comparatively, Sigmoid_v1 and Sigmoid_v2 demonstrated improved performance in various retrieval tasks. For instance, in the text retrieval task for the Flickr30k dataset, Sigmoid_v2 achieved a Recall@1 of 78.75%, representing an improvement of approximately 24.27% compared to CLIP. However, it should be noted that Sigmoid_v2 showed relatively high standard deviations, indicating some variability in performance across different runs. In contrast, the *sine* approach outperformed both Sigmoid_v1 and Sigmoid_v2, achieving a Recall@1 of 92.30% with a lower standard deviation.

Overall, the results indicate that applying *sine* as a diversified transformation led to significant Accuracy and retrieval precision (scores) improvements across multiple datasets and tasks. However, in some cases, the high standard deviations observed for Sigmoid_v2 suggest potential room for further optimization and stability enhancement.

The results of the Friedman test for image classification are presented in Table 4, which provides insights into the relative performance of the different algorithms. Based on the rankings obtained, it can be observed that *sine* achieved

TABLE 4. Average ranks obtained by each method in the friedman test. The best method (control method) is the *sine* method with the lowest ranking value.

Algorithm	Ranking
CLIP	3.25
Sigmoid v1	2.875
Sigmoid v2	2
<i>sine</i>	1.875

the best ranking (the lowest value of ranking), followed by Sigmoid_v2 and Sigmoid_v1, while CLIP obtained the worst ranking. The significant difference between the rankings suggests that applying diversified transformation in the algorithms led to improved Accuracy in the datasets under consideration. The Friedman statistic, distributed according to chi-square with 3 degrees of freedom, was 6.45, and the *p*-value computed by the Friedman test was 0.091655.

TABLE 5. Post-hoc comparison for alpha = 0.05 (Friedman).

<i>i</i>	Algorithm	$z = (R_0 - R_i)/SE$	<i>p</i>	Li
3	CLIP	2.130141	0.03316	0.008082
2	Sigmoid_v1	1.549193	0.121335	0.008082
1	Sigmoid_v2	0.193649	0.846451	0.05

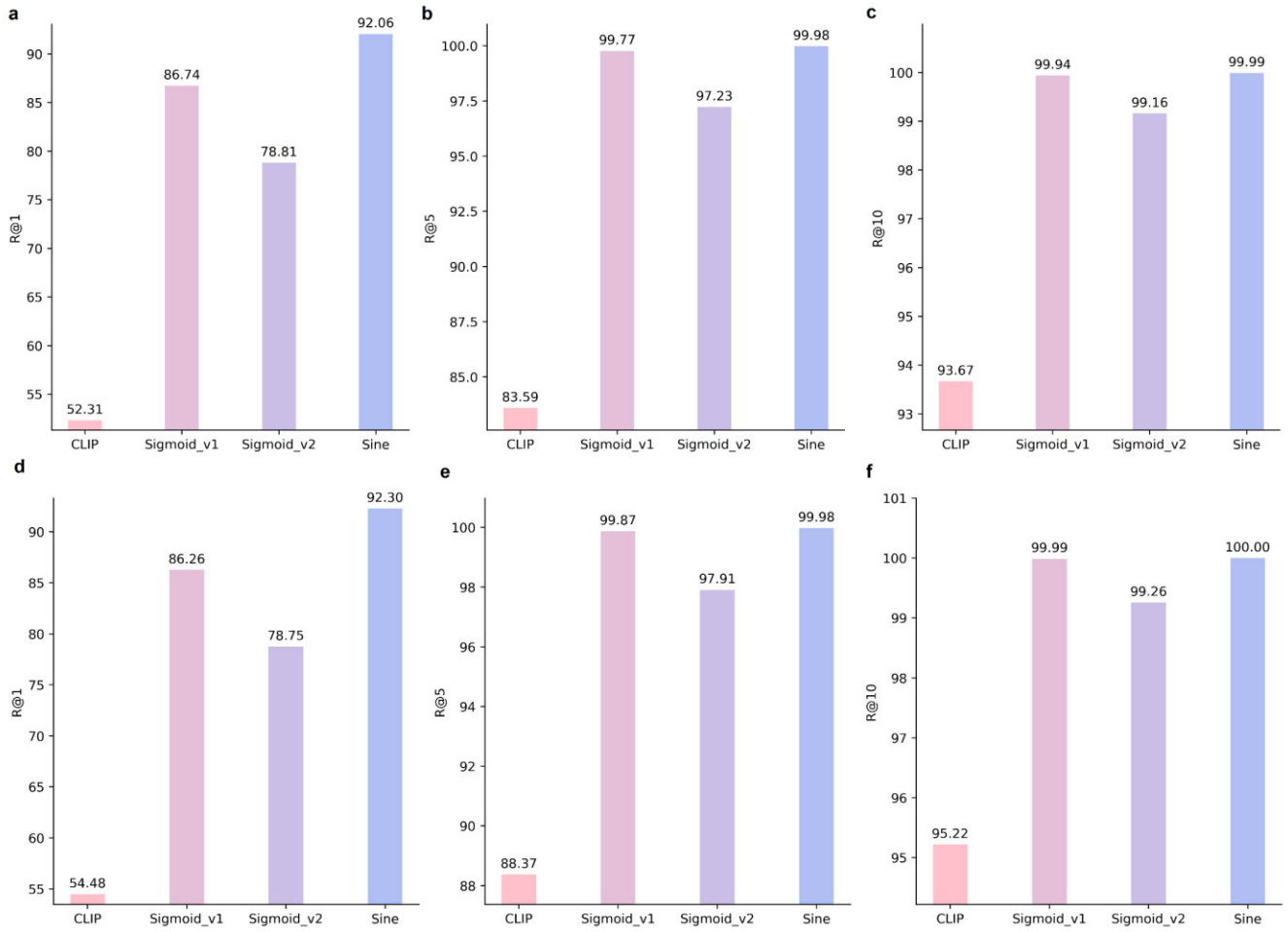


FIGURE 5. Text and image retrieval accuracy on the flickr30k benchmark dataset for the modified implementation of the clip model (referred to as ‘CLIP’) and its variants with enhanced diversity in the similarity matrix. **a, b, and c,** image retrieval accuracies. **d, e, and f,** text retrieval accuracies.

The first step of the Freidman test confirms significant differences between the compared methods. After this step, a post-hoc procedure is used to compare the control method (the best-ranked method) with the other methods as a pairwise comparison. The following Table 5 shows the results of Li’s post-hoc procedure while the *sine* method is chosen as the control method. Li’s approach rejects those hypotheses that have an unadjusted p-value ≤ 0.008082 . Therefore, *the sine* method significantly outperforms the sigmoid_v1 and CLIP methods, while there is no significant difference between the *sine* and sigmoid_v2 methods. In other words, *sine* and sigmoid_v2 statistically perform the same.

IV. DISCUSSION

A. CLIP MODEL WITH DIVERSITY-INDUCING FUNCTIONS

We observed improvements in the performance metrics for MNIST, Food101, RESISC45, FGVCaircraft, Flowers102, and PatchCamelyon, as well as image retrieval and text retrieval tasks on COCO and Flickr30k datasets, indicating the efficacy of our method in enhancing the CLIP model’s capabilities.

The improved results on datasets such as MNIST, Food101, RESISC45, FGVCaircraft, Flowers102, and PatchCamelyon can be attributed to the inherent characteristics of these datasets. MNIST, for instance, consists of handwritten digit images, which are relatively distinct from one another. Therefore, applying diversity-inducing functions helps accentuate the differences between the images, leading to more accurate matching with the corresponding text descriptions. Similarly, the Food101 dataset includes images of different types of food, which often possess unique visual features. By reducing the similarities among visually different food items, our approach facilitates better discrimination and improves the overall performance of this dataset.

The RESISC45 dataset, which contains images of various land cover categories, and the FGVCaircraft dataset, comprising images of different aircraft models, also benefit from our diversity-inducing functions. These datasets exhibit significant visual dissimilarities between their respective classes. By reducing the impact of shared visual features within a class, our method improves the discrimination capability of the CLIP model, leading to enhanced performance.

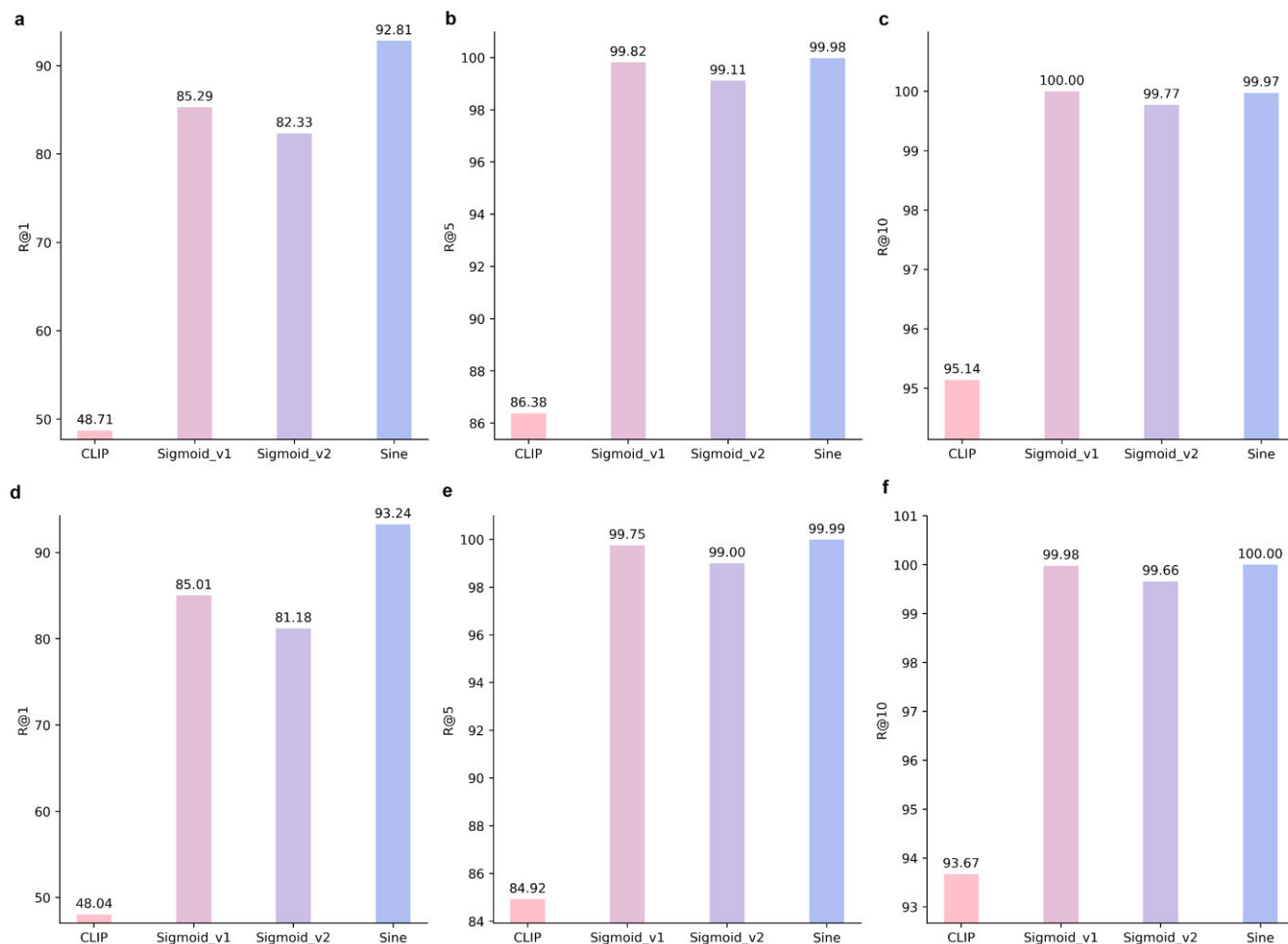


FIGURE 6. Text and image retrieval accuracy on the coco benchmark dataset for the modified implementation of the clip model (referred to as ‘CLIP’) and its variants with enhanced diversity in the similarity matrix. a, b, and c, image retrieval accuracies. d, e, and f, text retrieval accuracies.

Moreover, the Flowers102 dataset and PatchCamelyon dataset, both consisting of distinct categories of flowers and histopathology images, respectively, demonstrate improved results due to the presence of visually diverse classes. The diversity-inducing functions aid in capturing the subtle differences between these visually intricate classes, thus yielding more accurate matching between the images and text.

We observed improved performance in image retrieval and text retrieval tasks on the COCO dataset and Flickr30k dataset. These datasets contain many images with associated captions, making them suitable for evaluating the effectiveness of our diversity-inducing approach in multi-modal retrieval tasks. By incorporating the diversity-inducing functions into the CLIP model, we were able to enhance the discriminative power of the model in retrieving relevant images given textual queries and vice versa. This improvement suggests our approach can have practical implications in real-world applications such as image search engines and image captioning systems.

However, it is noteworthy that our approach did not yield comparable improvements on the EuroSAT dataset.

The EuroSAT dataset comprises satellite images of different land cover types, which might exhibit similarities in their visual characteristics, such as color and texture. These similarities may hinder the effectiveness of the diversity-inducing functions, as they tend to reduce the differences among visually similar images. Further investigation is required to understand the specific challenges associated with the EuroSAT dataset and explore alternative techniques to improve performance.

Up to this point, our evaluation has encompassed an array of models, including the modified implementation of the CLIP model, as well as its various variants with added diversity in similarity matrix. We also compare these models’ performance against other existing models, as mentioned in the introduction and literature review. It’s worth emphasizing that these models come with their unique assumptions, simplifications, encoding-decoding approaches, and computational resources. Additionally, they may employ distinct evaluation schemes. However, by undertaking this comparative approach, we gain a better understanding of how our models stack up against established ones.

TABLE 6. Evaluating text and image retrieval accuracy: modified implementation of the clip model (shown as 'clip-modified') and its variants with added diversity in similarity matrix, original clip model, and other models.

	Text Retrieval						Image Retrieval					
	Flickr30k			MSCOCO			Flickr30k			MSCOCO		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALIGN	88.6	98.7	99.7	58.6	83.0	89.7	75.7	93.8	96.8	45.6	69.8	78.6
FILIP	89.8	99.2	99.8	61.3	84.3	90.4	75.0	93.4	96.3	45.9	70.6	79.3
CoCa	92.5	99.5	99.9	66.3	86.2	91.8	80.4	95.7	97.7	51.2	74.2	82.0
CLIP	88.0	98.7	99.4	58.4	81.5	88.1	68.7	90.6	95.2	37.8	62.4	72.2
CLIP-Modified	54.5	88.4	95.2	48.0	84.0	93.7	52.3	83.6	93.7	48.7	86.4	95.1
Sine	92.3	100	100	93.2	100	100	92.1	100	100	92.8	100	100
Sigmoid v1	86.3	99.9	100	85.0	99.8	100	86.7	99.8	99.9	85.3	99.8	100
Sigmoid v2	78.8	97.9	99.3	81.2	99.0	99.67	78.8	97.5	99.2	82.3	99.1	99.8

Table 6 presents the results of evaluating various models (including other existing models in the literature) on text retrieval and image retrieval tasks for two datasets: Flickr30k and MSCOCO. We have conducted a comparison of our work with several other models: ALIGN [17], FILIP [25], CoCa [21], and the original CLIP model. To analyze and discuss the performance of “CLIP-Modified” (the CLIP version used here) in comparison to other models, we focus on the Recall at different levels (R@1, R@5, and R@10) for both tasks.

CLIP-Modified achieves relatively lower performance than other models, with R@1 values of 54.5% and 48.0% on Flickr30k and MSCOCO text retrieval, respectively. The performance gap becomes more evident at higher recall levels (R@5 and R@10), with CLIP-Modified consistently performing worse than other models across both datasets.

CLIP-Modified also underperforms on the image retrieval task, achieving R@1 values of 52.3% and 48.7% on Flickr30k and MSCOCO, respectively. It continues to exhibit lower performance at higher recall levels (R@5 and R@10) when compared to other models. However, after introducing diversity into the models, there has been a clear and significant improvement in their performance. These improvements have placed the models' accuracy within a range comparable to previous models. In some instances, they have even surpassed the accuracy of the original CLIP model. These enhancements indicate that increasing diversity in the similarity matrix is a successful approach to improve the performance of “CLIP Modified” in text and image retrieval tasks, making it more effective in finding the most relevant results for given queries.

B. LIMITATIONS AND FUTURE RESEARCH RECOMMENDATIONS

Due to the constraints of our hardware resources, we had to make certain modifications to the base model. For instance, we used a pre-trained ResNet50 model as an image encoder and DistilBERT as a text encoder. While these modifications allowed us to proceed with our experiments, it is possible that alternative architectures or models could yield different results. Exploring a wider range of model configurations could provide valuable insights.

Several avenues for future work can be explored based on the results and discussion presented in this paper. Firstly,

investigating the reasons behind the limited improvement observed on the EuroSAT dataset can provide insights into the challenges associated with visually similar images. Developing alternative diversity-inducing techniques tailored to such datasets might help address this limitation and further improve performance.

To add diversity to our similarity matrix, we employed simple transformation functions. These functions were designed to introduce diversity based on our understanding of the data, but there is room for improvement. Further research could focus on developing more sophisticated and data-specific diversity functions, which might lead to even better performance and richer representations. The choice of diversity-inducing functions can influence the overall performance, and further research is needed to identify the most effective functions for different dataset characteristics.

For example, the current study employs a 4-level quantization for Scaled_sig matrix. Further research should investigate the implications of using higher quantization levels, such as 6 or 8, on model accuracy and computational efficiency. In addition, exploring non-linear quantization techniques may potentially yield more accurate or efficient models. Techniques like logarithmic scaling should be evaluated.

Furthermore, it would be valuable to investigate the generalizability/transferability of our approach to other multi-modal models beyond CLIP. Testing the effectiveness of diversity-inducing techniques on models like VSE++ [36], Unicoder [37], or LXMERT (Language-Enabled Multi-modal Pretraining (see, e.g., Liu et al. [38])) could provide insights into the broader applicability and robustness of these techniques across different architectures.

Conducting a comprehensive analysis of the computational cost associated with the diversity-inducing functions and their impact on inference time can help evaluate the trade-off between performance gains and computational efficiency. This analysis can inform the practical deployment of the proposed approach in resource-constrained environments.

Lastly, exploring the transferability of the diversity-inducing techniques to other domains or datasets with unique characteristics, such as medical imaging or remote sensing, holds promise for expanding the applications of multi-modal models.

V. CONCLUSION

In this research paper, we explored an approach to improve the performance of the CLIP model by introducing more diversity into the matrix of similarity values, referred to as *logits*. Upon evaluation, we identified the need for increased diversity in the *logits* matrix to enhance the CLIP model approach. We integrated two functions, *sine* (triangular) and *sigmoid* (with two versions), into a modified CLIP model. By incorporating these functions, we introduced variations in the *logits*, which resulted in improved performance. This modification allows our model to capture a broader range of relationships and nuances between textual and visual representations. Through experimentation and analysis, we demonstrated the effectiveness of our approach in achieving enhanced performance metrics. *sine* consistently outperformed other algorithms in image classification, with a mean Accuracy of 96.03% in the MNIST dataset, representing an improvement of approximately 0.75% compared to the baseline algorithm. Sigmoid_v2 had a relatively high standard deviation in the X-ray dataset, indicating some variability in results. In text and image retrieval tasks, the *sine* approach demonstrated the highest mean retrieval precision, particularly in the COCO dataset's image retrieval task, achieving a Recall@1 of 92.81%, a significant improvement of 44.10% compared to the baseline algorithm. The standard deviations for the *sine* approach's performance were relatively low, indicating consistent and stable results. Our findings highlight the significance of incorporating diverse functions into the model's architecture to unlock its full potential. The insights gained from this study pave the way for further advancements in multi-modal understanding and its applications across different domains.

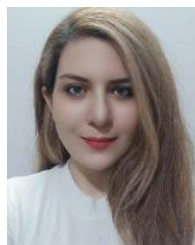
ACKNOWLEDGMENT

The authors would like to acknowledge the valuable support and resources provided by the Norwegian Institute for Air Research (NILU) in facilitating and conducting the research presented in this paper. Additionally, they are grateful for the financial support received from various sources that made this research possible.

REFERENCES

- [1] A. Latif, "Content-based image retrieval and feature extraction: A comprehensive review," *Math. Problems Eng.*, vol. 2019, Aug. 2019, Art. no. 9658350.
- [2] N. K. Rout, M. Atulkar, and M. K. Ahirwal, "A review on content-based image retrieval system: Present trends and future challenges," *Int. J. Comput. Vis. Robot.*, vol. 11, no. 5, pp. 461–485, 2021.
- [3] J. Liu, X. Chu, Y. Wang, and M. Wang, "Deep text retrieval models based on DNN, CNN, RNN and transformer: A review," in *Proc. IEEE 8th Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2022, pp. 391–400.
- [4] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.
- [5] J. Liu, C. Xu, and H. Lu, "Cross-media retrieval: State-of-the-art and open issues," *Int. J. Multimedia Intell. Secur.*, vol. 1, no. 1, pp. 33–52, 2010.
- [6] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, "Image-text retrieval: A survey on recent research and development," 2022, *arXiv:2203.14713*.
- [7] J. Chen, L. Zhang, C. Bai, and K. Kpalma, "Review of recent deep learning based methods for image-text retrieval," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 167–172.
- [8] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.
- [9] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese, "Text2Shape: Generating shapes from natural language by learning joint embeddings," in *Proc. Comput. Vis. (ACCV) 14th Asian Conf. Comput. Vis. Perth, SCT, Australia: Springer*, Dec. 2018, pp. 100–116.
- [10] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proc. Mach. Learn. Healthcare Conf.*, 2022, pp. 1–16.
- [11] A. Frome, G. S. Corrado, and J. Shlens, "A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–6.
- [12] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–9.
- [13] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic modality interaction modeling for image-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1104–1113.
- [14] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.
- [15] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. PMLR*, 2021, pp. 8748–8763.
- [16] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Comput. Vis.-ECCV 13th Eur. Conf.*, Zurich, Switzerland, vol. 13, Sep. 2014, pp. 446–461.
- [17] C. Jia, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
- [18] C. Saharia, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
- [19] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," in *Proc. Comput. Vis. ECCV 17th Eur. Conf. Tel Aviv, Israel: Springer*, Oct. 2022, pp. 529–544.
- [20] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18123–18133.
- [21] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.
- [22] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 16816–16825.
- [23] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
- [24] H. Pham, "Combined scaling for open-vocabulary image classification," 2021, *arXiv:2111.10050*.
- [25] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "FILIP: Fine-grained interactive language-image pre-training," 2021, *arXiv:2111.07783*.
- [26] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [28] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2019, pp. 558–567.
- [29] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.
- [30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

- [31] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [33] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
- [34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [35] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic image captioning based on ResNet50 and LSTM with soft attention," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–7, Oct. 2020.
- [36] F. Faghri, D. J. Fleet, J. Ryan Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.
- [37] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, and M. Zhou, "Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks," 2019, *arXiv:1909.00964*.
- [38] T. Liu, Z. Wu, W. Xiong, J. Chen, and Y.-G. Jiang, "Unified multimodal pre-training and prompt-based tuning for vision-language understanding and generation," 2021, *arXiv:2112.05587*.
- [39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Jul. 1998.
- [40] L. Bossard, M. Guillaumin, and L. Van Gool, *Food-101-Mining Discriminative Components With Random Forests*. Cham, Switzerland: Springer, 2014, pp. 446–461.
- [41] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [42] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.
- [43] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 1–8.
- [44] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in *Medical Image Computing and Computer Assisted Intervention-MICCAI*. Granada, Spain: Springer, Sep. 2018, pp. 210–218.
- [45] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.
- [46] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification," *Mendeley Data*, vol. 2, no. 2, p. 651, 2018.
- [47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.
- [48] T.-Y. Lin, "Microsoft COCO: Common objects in context," in *Proc. Comput. Vis. ECCV 13th Eur. Conf.*, Zurich, Switzerland, Sep. 2014, pp. 740–755.



MAHSA MOHAMMADI received the bachelor's degree in computer software engineering from Shiraz University, and the master's degree in artificial intelligence from the Shahid Bahonar University of Kerman. Her academic background equipped her with knowledge and skills in areas, such as natural language processing (NLP) and machine learning. She is currently an accomplished Software Developer with a wealth of experience in the field. She is also with Protector Forsikring ASA, Oslo, Norway, responsible for developing software solutions using technologies, such as Hibernate, Spring Boot, Java, and JavaScript.



MAHDI EFTEKHARI was born in Kerman, Iran, in 1978. He received the B.Sc. degree in computer engineering and the M.Sc. and Ph.D. degrees in artificial intelligence from the Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran, in September 2001, 2004, and 2008, respectively. He has been a Faculty Member with the Shahid Bahonar University of Kerman, Kerman, since 2008. He is currently a Full Professor with the Department of Computer Engineering. He is the author and coauthor of about 140 papers in cited journals and conferences. His research interests include deep learning, machine learning, fuzzy methods and systems, and the application of intelligent methods in bioinformatics.



AMIRHOSSEIN HASSANI received the Ph.D. degree. He is currently a Researcher with The Climate and Environmental Research Institute NILU. With a background in petroleum engineering and a strong passion for environmental sciences, his research focuses on developing data-driven tools to inform policymaking and enhance the resilience of societies and ecosystems to future environmental challenges. His expertise lies in the application of Earth system science data to address current and projected social, economic, and environmental challenges. He is also committed to leveraging the power of machine learning applications beyond environmental domains.

...