Enhancing Contextual Understanding with Multimodal Siamese Networks Using Contrastive Loss and Text Embeddings

Andro Aprila Adiputra
Computer Science and Engineering
Pusan National University
Busan, South Korea
androaprila@pusan.ac.kr

Ahmada Yusril Kadiptya
Computer Science and Engineering
Pusan National University
Busan, South Korea
yusril@pusan.ac.kr

Thi-Thu-Huong Le
Blockchain Platform Research Center
Pusan National University
Busan, South Korea
lehuong7885@gmail.com

JunYoung Son
Computer Science and Engineering
Pusan National University
Busan, South Korea
jysonpaperinfo@gmail.com

Howon Kim
Computer Science and Engineering
Pusan National University
Busan, South Korea
howonkim@pusan.ac.kr

***Abstract*—Deep learning has achieved significant advances in image representation learning, yet it remains constrained by challenges such as imbalanced datasets and limited contextual understanding of paired data. To address these issues, we propose a novel multimodal approach that integrates Contrastive Siamese Neural Networks with text embeddings generated using vision language models (VLMs) especially Pixtral. Our method aims to enhance contextual alignment between paired images by combining image embeddings and text embeddings derived from language models such as BERT or RoBERTa. Inspired by the architecture of CLIP, which synchronizes image and text encoders, our approach adapts contrastive learning to focus specifically on image embeddings while leveraging text embeddings to enrich the context. This multimodal framework is evaluated on both imbalanced and balanced datasets to determine its robustness and effectiveness. Key contributions include analyzing the role of generated text in providing context to images and demonstrating the potential of Siamese networks in multimodal settings. The experimental results highlight the advantages of our approach in improving contextual understanding and improving overall performance in balanced and imbalanced dataset settings.**

***Index Terms*—Siamese Neural Network, Multimodal Learning, Contrastive Loss, Text Embedding, Vision-Language Models (VLMs), Contextual Representation Learning, Image-Text Alignment**

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Convergence security core talent training business(Pusan National University) support program(RS-2022-II221201) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(RS-2020-II201797) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). And this reserach was also supported by Korea Planning & Evaluation Institute of Industrial Technology(KEIT) grant funded by the Korea government(MOTIE) (No.RS-2024-00407022, Enhancement of cloud BMS for xEV based on multi-safety sensor for SDV).
Corresponding author: JunYoung Son (jysonpaperinfo.com)

I. Introduction

Deep learning has become the cornerstone of modern advancements in Artificial Intelligence (AI), achieving remarkable success across diverse applications. Convolutional Neural Networks (CNNs) have played a key role in image-related tasks by robustly extracting patterns and meaningful representations from visual data [1]. The breakthrough in CNNs gained widespread attention with the ImageNet Challenge [2], which highlighted the importance of large-scale datasets and established critical benchmarks to evaluate image classification models.

AlexNet [3] marked a significant milestone in tackling large-scale image classification challenges, demonstrating the robustness of CNNs. Building on this success, VGG16 [4] introduced a deeper, carefully crafted CNN architecture that improved the receptive field and effectively captured non-linear patterns. However, deeper CNN architectures often encountered performance degradation due to vanishing gradients. The introduction of Residual Networks (ResNet) [5] addressed this challenge by facilitating the flow of gradients during back-propagation, enabling more effective training of deep networks.

Despite these advancements, deep learning models remain heavily dependent on large-scale datasets and are sensitive to class imbalance issues. Collecting diverse and robust datasets requires substantial resources and careful consideration of constraints to ensure the model's generalization capabilities across various conditions. To address the challenges posed by imbalanced class distributions, the Siamese Neural Network (Siamese) [6] was proposed. Siamese networks, often coupled with CNNs, can learn feature-rich image representations for tasks like one-shot learning. Various enhancements have been

made to Siamese networks, primarily focusing on emphasizing differences between paired inputs, with Contrastive Loss being a notable example.

Recently, multimodal approaches that integrate visual and textual data have demonstrated significant potential in enhancing representation learning. For instance, CLIP (Contrastive Language-Image Pre-training) [7] combines an image encoder and a text encoder, aligning their representations through contrastive loss to achieve state-of-the-art performance in cross-modal tasks. Inspired by this paradigm, our work extends the Siamese network architecture by incorporating text embeddings, generating textual descriptions for images using Vision-Language Models (VLMs) such as Pixtral or QwenVL. This multimodal approach aims to provide contextual understanding between paired images, thereby enhancing the robustness of feature learning.

Our approach differs from CLIP in its application of contrastive loss, as we utilize it specifically for image embeddings while integrating text embeddings to enrich the contextual understanding of paired images. Additionally, language models such as BERT and RoBERTa are explored to enhance text representation, ensuring high-quality embeddings for training.

The main contributions of this paper are as follows.

- We evaluated the robustness of Siamese Neural Networks combined with text embeddings on both imbalanced and balanced datasets.
- We analyze the effectiveness of optimally generated text in providing context to images and enhancing multimodal learning.

The remainder of the paper is structured as follows. Section II reviews related works in Siamese networks, multimodal learning, and their applications. Section III presents the methodology, including the model architecture and the training process. Section IV outlines the experimental setup and results, followed by a discussion in Section V. Finally, Section VI concludes the paper and outlines future research directions.

II. Related Works

A. Siamese Neural Network

Siamese Neural Networks (SNNs) have emerged as a robust architecture for tasks involving pairwise comparisons, particularly in one-shot learning and similarity-based tasks. The foundational work by Koch et al. [6] introduced shared weight networks to learn feature embeddings from paired inputs, effectively capturing similarity and dissimilarity between data points. This innovative design has proven to be highly effective in tasks such as image matching and verification. A key component of the SNN framework is the Contrastive Loss function [8], which optimizes the embedding space by penalizing dissimilar pairs, enhancing the network's ability to distinguish between similar and dissimilar inputs.

SNNs have demonstrated strong performance, especially in scenarios with limited labeled data or imbalanced class distributions, where conventional supervised learning methods often struggle. Extensions to the Siamese framework, such as the incorporation of triplet loss [9], further improve embedding discriminability by considering triplets of anchor, positive, and negative examples. Moreover, the adoption of deeper convolutional architectures like Residual Networks (ResNet) [5] has significantly enhanced the generalization capabilities of Siamese networks across diverse datasets.

B. Multimodal Classification

Multimodal learning, which integrates data from multiple modalities such as images and text, has gained considerable traction to improve task performance by leveraging complementary information. A landmark contribution in this domain is CLIP (Contrastive Language-Image Pre-training) [7], which aligns visual and textual representations using a contrastive loss function. CLIP employs an image encoder (e.g., Vision Transformers) and a text encoder (e.g., transformer-based language models) to jointly learn embeddings, enabling the model to perform zero-shot image classification by utilizing textual descriptions.

The integration of multimodal approaches has revolutionized cross-modal tasks such as image captioning [10], visual question answering (VQA) [11], and image-text retrieval [12]. Models like VisualBERT [13] take a step further by treating image regions as tokens within a transformer framework, learning joint representations of images and text. These methods leverage large-scale datasets containing paired image-text annotations, facilitating the development of robust embeddings that generalize effectively across various domains.

Recent advancements have also explored combining CNN-based image encoders with language models such as BERT or RoBERTa for tasks like image-text retrieval and captioning [14]. By unifying visual and textual information, these approaches enable models to comprehend the semantic context of images while retaining the ability to extract visual features. This fusion of modalities ensures more accurate, contextually aware predictions, paving the way for improvements in multimodal learning applications.

C. Vision-Language Models

Vision-Language Models (VLMs), such as Pixtral [15] and QwenVL [16], are specifically designed to generate textual descriptions for images and facilitate the alignment of multimodal data. These models utilize large-scale datasets containing paired images and text to learn robust joint embeddings that effectively capture the semantic relationships between the two modalities.

By leveraging these joint embeddings, VLMs have demonstrated significant potential across a range of applications, including image captioning, visual question answering (VQA), and cross-modal retrieval. The high-quality text representations generated by these models enhance the contextual understanding of visual data, making them an invaluable asset in multimodal learning tasks. Their ability to bridge the gap between vision and language further strengthens their applicability to real-world scenarios, where integrating

multimodal information is crucial for achieving state-of-the-art performance.

III. Multimodal Siamese Neural Network

This section details our proposed approach for leveraging multimodal data in a Siamese Neural Network (SNN) framework as shown in Figure 1. The method integrates both image and text embeddings to enhance contextual understanding and address challenges related to imbalanced datasets and limited labeled data. Key components of the methodology include image embedding, text embedding, contrastive learning with textual context, and the generation of training data using Vision-Language Models (VLMs).

A. Image and Text Embedding

Image embedding involves transforming visual data into dense vector representations that capture the most salient features of the input. In the proposed framework, Siamese networks leverage CNN-based architectures to extract these embeddings from input images. Contrastive loss is applied during training to optimize the embedding space, ensuring that similar images are drawn closer together while dissimilar images are pushed further apart. This approach facilitates the learning of robust visual representations, particularly in challenging scenarios such as imbalanced datasets, where traditional methods often struggle.

Text embedding, on the other hand, involves converting textual descriptions into numerical vector representations that encapsulate their semantic meaning. In this framework, the text embeddings are designed to complement the image embeddings by enriching the overall contextual understanding of the data. The integration of text embeddings ensures that the framework leverages both visual and semantic features, enhancing its ability to generalize across diverse tasks.

For image embeddings, this research employs a variety of CNN architectures, including VGG16 [4], ResNet50, ResNet101 [5], and InceptionV3 [17]. These architectures are selected for their proven ability to extract robust visual features from diverse datasets.

For text embeddings, transformer-based models such as BERT [18], ALBERT [19], RoBERTa [20], and DeBERTa [21] are utilized. These models are chosen for their capacity to generate rich semantic representations of textual data, which complement the visual features extracted from images. By combining the strengths of these models, the proposed framework achieves a more comprehensive and context-aware representation, enabling improved performance in multimodal tasks.

B. Contrastive Siamese with Text as Context

We adapt a training pipeline inspired by CLIP and extend it into a Siamese framework with automated text descriptions generated using Vision-Language Models (VLMs). In standard Siamese networks employing contrastive loss, the model processes pairs of inputs (anchor and positive/negative samples) and calculates the Euclidean distance between them to measure similarity. However, in our proposed method (illustrated in Fig. 1), we introduce additional distance calculations before the inputs pass through the Siamese module. This enhancement incorporates contextual information by calculating similarities between images and their descriptive texts as well as between text pairs. These additional calculations create new pairings (e.g., anchor or positive/negative image with its corresponding text), enriching the overall representation.

To align the distributions of image and text embeddings, we integrate image and text embedding models into an additional module. This module, as shown in Fig. 1, uses the same layer configurations for both modalities. It translates the extracted features and contexts into a unified representation. To further mitigate extreme variations between the two modalities, a Tanh activation function is employed along with an MLP layer to control and normalize the distributions.

The extracted embeddings from the backbone models (image and text) are then used to calculate similarity and distance metrics. While standard contrastive loss training computes a single Euclidean distance for image pairs, our approach combines both Euclidean distance (for same-modality pairs such as image-to-image or text-to-text) and cosine similarity (for cross-modality pairs such as image-to-text).

$$d\left(X_{image_1}, X_{image_2}\right) = \sqrt{\sum_{i=1}^{n} \left(X_{image_{2_i}} - X_{image_{1_i}}\right)^2} \tag{1}$$

Here, X_{image_1} denotes the anchor image, and X_{image_2} denotes the positive/negative image. This equation measures the Euclidean distance to quantify the similarity or dissimilarity between the two images, a standard approach in Siamese contrastive learning.

$$d\left(X_{text_1}, X_{text_2}\right) = \sqrt{\sum_{i=1}^{n} \left(X_{text_{2_i}} - X_{text_{1_i}}\right)^2} \tag{2}$$

Here, X_{text_1} represents the text description of the anchor image, while X_{text_2} corresponds to the text description of the positive/negative image. Calculating the distance between text embeddings emphasizes important descriptive features for both images.

$$similarity\left(X_{image_1}, X_{text_1}\right) = \frac{X_{image_1} \cdot X_{text_1}}{\|X_{image_1}\| \, \|X_{text_1}\|} \tag{3}$$

This cosine similarity equation calculates the correlation between an anchor image and its descriptive text, emphasizing the alignment of features and context.

$$similarity\left(X_{image_2}, X_{text_2}\right) = \frac{X_{image_2} \cdot X_{text_2}}{\|X_{image_2}\| \, \|X_{text_2}\|} \tag{4}$$

The similarity calculation for the positive/negative image and its descriptive text helps determine the contextual and feature alignment relative to the anchor.

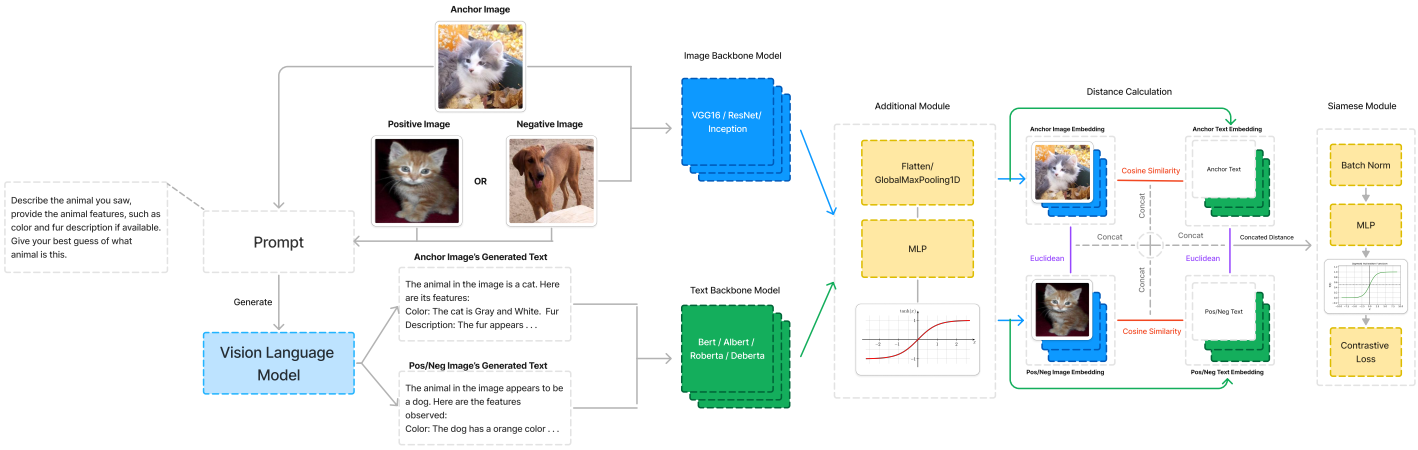

Fig. 1. The Proposed Model Design and Training Pipeline: In the additional module, only DeBERTa and RoBERTa incorporate GlobalMaxPooling1D for feature aggregation. The pipeline illustrates the integration of image and text encoders, the application of contrastive loss for robust representation learning, and the generation of textual descriptions to provide contextual information.

$$Loss = (Y)\,(Y_{pred})^2 + (1-Y)\,\{\max(0, m - Y_{pred})\}^2 \quad (5)$$

Here, Y represents the ground truth for the pair, Y_{pred} is the model's predicted result, and m is the margin that separates dissimilar pairs. Contrastive loss minimizes the distance between similar features while penalizing dissimilar pairs based on the ground truth.

Finally, all calculated distances and similarities are concatenated and passed into the Siamese module. Within the module, batch normalization is applied to regularize the feature distributions before the data is passed to the classifier layer. The contrastive loss function refines the feature representations by reducing distances for similar pairs and increasing distances for dissimilar pairs, ensuring effective learning of meaningful embeddings.

C. Text Generation and Training

We evaluate the model using two types of datasets: the first is a face recognition dataset [22], which exhibits imbalanced class settings, and the second is a classic classification dataset for animal classification [23].

The face recognition dataset consists of 1,323 distinct classes, each with unique facial features, and no data augmentation is applied. The class distribution is illustrated in Fig. 2, focusing on the top 10 classes with more than 15 images and those with fewer. The dataset is highly imbalanced; some classes contain over 20 images, while others have as few as 2 images.

The second dataset, the animal classification dataset, includes three classes (cat, dog, and snake), with an equal distribution of data across the classes. Each class contains 1,500 images, resulting in a balanced dataset.

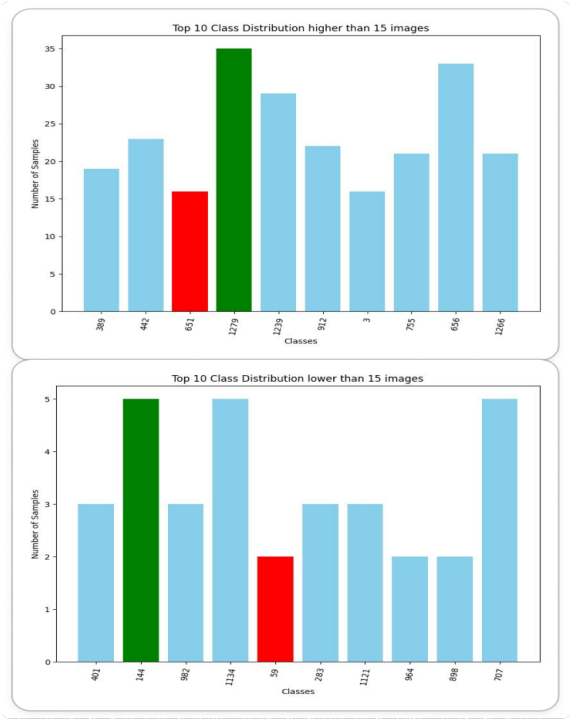

Fig. 2. Face Recognition Training Distribution

Before training, we split each dataset using stratified k-fold cross-validation to ensure balanced distributions between the training and test sets. Text descriptions for the images are generated using VLMs, specifically Pixtral.

The model is trained using the SGD optimizer with a learning rate scheduler, starting at 0.1 and decreasing progressively to 0.0001. The number of epochs varies depending on the

dataset: the face recognition dataset is trained for 20 epochs, while the animal classification dataset requires 30 to 40 epochs for better convergence.

For the image embedding model, we utilize pre-trained ImageNet weights and freeze these weights during training. For the text embedding model, we use pre-trained base models and only DeBERTa using the base small pre-trained model. Finally, the Siamese model is trained on GPUs, specifically using 2×A100.

IV. Experiment

In this section, we analyze the performance of each image and text backbone combination. Additionally, we compare the original Siamese contrastive learning model with different image backbones as a baseline. A random threshold close to 0.5 from the precision-recall curve is used to assess model consistency and improvements compared to the original Siamese model.

The results in Table I demonstrate that integrating pre-trained text embeddings (Bert, Albert, Deberta, Roberta) significantly enhances performance. On the imbalanced Face Recognition dataset, ResNet101 + Bert achieves the highest accuracy (99.96%), while residual-based networks show consistent performance across both datasets. In the balanced Animal dataset, several combinations achieve near-perfect accuracy, with ResNet50 + Albert attaining the highest F1-score (97.63%).

Interestingly, VGG16 with Deberta shows an 8% improvement in F1-score for the Face Recognition task, outperforming other VGG16 variants. However, in the Animal dataset, VGG16 lags behind residual networks, which leverage skip connections to reduce overfitting. InceptionV3 exhibits strong performance on the Animal dataset but struggles with Face Recognition, indicating a limitation in extracting deeper features.

Figure 3 highlights that all models achieve at least 64% precision, with residual networks showing high recall but being overconfident with negative samples. Conversely, InceptionV3 demonstrates higher precision but lower recall, favoring positive samples. Adding text embeddings improves performance consistency, emphasizing the advantage of incorporating textual context with image features to enhance decision boundaries.

In the animal dataset, as illustrated in Figure 4, the residual networks consistently achieve peak performance in both precision and recall. The VGG16 model exhibits a slight decrease in precision and recall, likely due to overfitting when compared to its original Siamese model. Interestingly, InceptionV3 combined with Albert shows a noticeable drop in performance, despite maintaining overall parity with other InceptionV3 variants for animal dataset classification.

V. Discussion

We evaluated various combinations of image backbone models with different text backbone models. The choice of image backbone is a significant factor in determining consistent

TABLE I
Image and Text Backbone Evaluations

Model Name	Face Recognition		Animal Dataset	
	Accuracy	**F1**	**Accuracy**	**F1**
VGG16	0.6807	0.6849	0.9996	0.9396
VGG16 + Bert	0.9376	0.7643	0.9766	0.8930
VGG16 + Albert	0.9952	0.7458	0.9926	0.8820
VGG16 + Deberta	0.8886	**0.7656**	0.9813	0.8740
VGG16 + Roberta	0.9621	0.7572	0.9729	0.8880
InceptionV3	0.6416	0.6500	0.9943	0.9086
InceptionV3 + Bert	0.9458	0.6729	0.9996	0.9753
InceptionV3 + Albert	0.9398	0.6689	0.9996	0.9756
InceptionV3 + Deberta	0.9793	0.6608	0.9996	0.9726
InceptionV3 + Roberta	0.9891	0.6855	**1.0**	0.9720
ResNet50	0.7296	0.6986	0.9986	0.9023
ResNet50 + Bert	0.9988	0.7214	0.9996	0.9739
ResNet50 + Albert	0.9970	0.7214	0.9993	**0.9763**
ResNet50 + Deberta	0.9968	0.7201	**1.0**	0.9673
ResNet50 + Roberta	0.9981	0.7206	0.9996	0.9739
ResNet101	0.6930	0.6840	0.9996	0.9123
ResNet101 + Bert	**0.9996**	0.7246	0.9996	0.9753
ResNet101 + Albert	0.9962	0.7210	0.9996	0.9756
ResNet101 + Deberta	0.9988	0.7224	0.9996	0.9726
ResNet101 + Roberta	0.9962	0.7192	**1.0**	0.9676

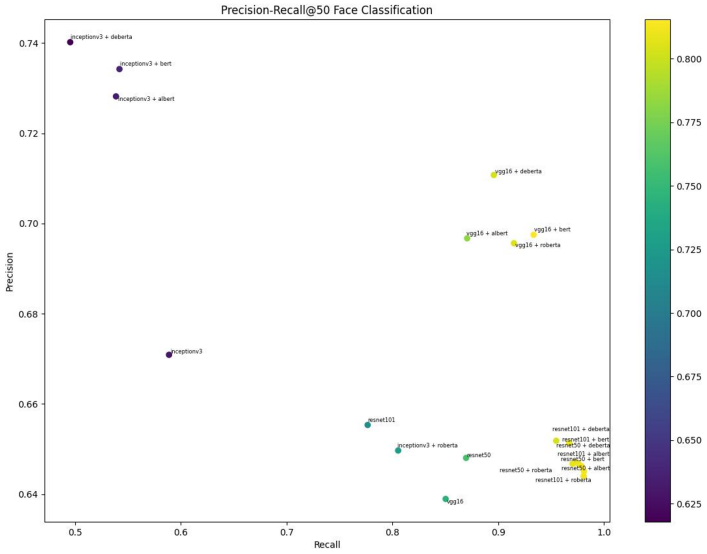

Fig. 3. Face Recognition Precision-Recall@50. Zoom in for details.

model performance. For instance, VGG16 outperforms other image backbones in face recognition but lags slightly behind in animal classification tasks, whereas residual-based networks demonstrate consistent performance across both datasets. We hypothesize that VGG16 tends to overfit and struggles to effectively link image features with the provided text embedding context. Conducting additional experiments on different datasets could provide broader insights into whether our method produces suboptimal outcomes with other image/text model combinations. Moreover, incorporating regularization techniques or altering activation functions may help mitigate overfitting.

Our experiments reveal that adding text as contextual information enhances overall performance. However, variations

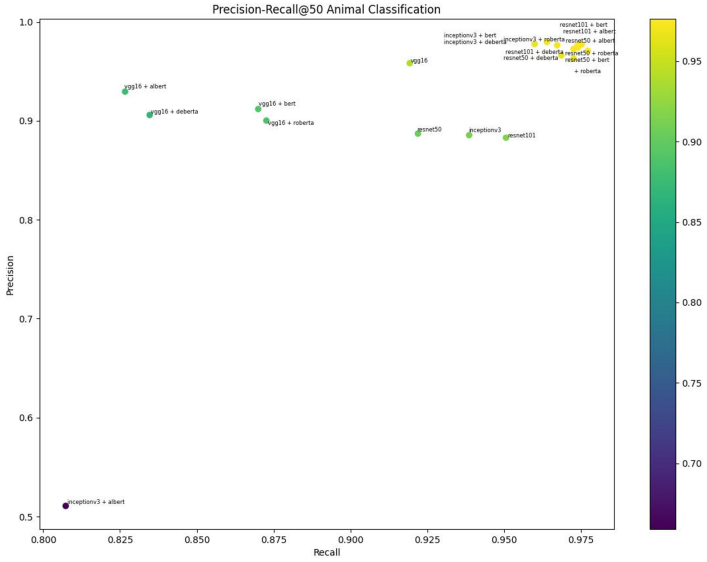

Fig. 4. Animal Classification Precision-Recall@50.
Zoom in for details.

among different text models yielded only marginal differences in results. This observation suggests that the text embedding models exhibit similar embedding convergence. Additionally, the use of the `tanh` activation function may constrain the embedding space, resulting in similar distance calculations across embeddings. Future work could explore whether larger pre-trained models might enhance performance or investigate alternative scaling techniques that avoid excessively limiting the text model's embedding space.

VI. Conclusion

In this paper, we demonstrated that a contrastive Siamese neural network can be significantly enhanced by aligning the image feature map embeddings with contextual text embeddings. Our approach improved overall performance by an 8% margin, increasing the F1-Score from 68% to 76% on the challenging face recognition task with extreme class imbalance, compared to the original Siamese neural network.

Through precision-recall threshold analysis, models such as VGG16 and residual networks (ResNet50 and ResNet101) exhibited substantial improvements over the baseline models. On an ideal dataset, while VGG16 lagged by approximately 2% compared to the original Siamese model, InceptionV3 and residual networks achieved improvements of up to 7%. Overall, the highest-performing original Siamese model achieved F1-Scores of 69.86% and 93.96% on the face recognition and animal classification tasks, respectively. In contrast, our enhanced Siamese model with text embeddings achieved 76.56% and 97.63%, reflecting significant improvements across both datasets.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). IEEE, 6 2009. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206848

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015, pp. 1–30.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[8] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[10] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[11] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[12] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems*, vol. 27, 2014.

[13] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "Visualbert: A simple and performant baseline for vision and language. arxiv 2019," *arXiv preprint arXiv:1908.03557*, vol. 2.

[14] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.

[15] P. Agrawal, S. Antoniak, E. B. Hanna, D. Chaplot, J. Chudnovsky, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob *et al.*, "Pixtral 12b," *arXiv preprint arXiv:2410.07073*, 2024.

[16] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[18] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Z. Lan, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[20] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[21] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[22] "Face recog celebrity face." [Online]. Available: https://www.kaggle.com/datasets/stoicstatic/face-recognition-dataset

[23] "Animal classification dataset." [Online]. Available: https://www.kaggle.com/datasets/borhanitrash/animal-image-classification-dataset