Grp17Net: Multi-Object Image Classification via Convolutional Neural Networks

Jaydeep Radadiya
Department of Computer Science
Texas A&M University
College Station, USA
jdr@tamu.edu

Can Polat
Electrical & Computer Engineering
Texas A&M University
College Station, USA
can.polat@tamu.edu

Ashutosh Punyani
Department of Computer Science
Texas A&M University
College Station, USA
ashutoshpunyani99@tamu.edu

Sharwari Udaykumar Shah
Department of Information and Operations Management
Texas A&M University
College Station, USA
sharwari.shah@tamu.edu

Tim Lyons
Department of Rangeland, Wildlife, and Fisheries Management
Texas A&M University
College Station, USA
timothy.lyons@tam.edu

***Abstract*—This project presents Grp17Net, a convolutional neural network (CNN) architecture, for multi-class image classification using the Fashion MNIST dataset as a benchmark. The study incorporates advanced techniques such as principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) for dimensionality reduction to improve CNN performance. The proposed model, inspired by VGGNet, integrates convolutional, batch normalization, and max-pooling layers, optimized via the Adam optimizer and hyperparameter tuning using Optuna. Data preprocessing, including normalization and augmentation, enhanced the model's robustness, achieving a validation accuracy of 91% with an F1-score of 0.89.**

This work further explores zero-shot image-to-text classification using the CLIP model, identifying challenges in distinguishing visually similar categories due to overlapping features. Misclassification analysis and embedding visualizations highlight areas for improvement, including the integration of attention mechanisms and advanced feature extraction methods. Future work will focus on leveraging transformers, graph neural networks, and parametric UMAP for enhanced classification performance and generalization across complex datasets. This project underscores the synergy of CNNs with dimensionality reduction and multimodal techniques in advancing image classification.

I. Introduction

Image classification is a foundational task in computer vision, aiming to assign labels to images based on their content. Over the years, various techniques have been developed to enhance the performance and efficiency of classification systems. Early approaches such as principal component analysis (PCA) leveraged dimensionality reduction to transform high-dimensional data into lower-dimensional spaces while preserving essential variance, facilitating efficient feature extraction and noise mitigation [1]. Building on these principles, clustering methods like nonlinear approximation and projection further refined classification pipelines by capturing nonlinear relationships through iterative data projections [2].

The advent of deep learning, particularly the introduction of convolutional neural networks (CNNs), revolutionized image classification by enabling models to learn hierarchical representations directly from data. These architectures have since driven remarkable improvements in accuracy and generalization across various applications, including object detection, semantic segmentation, and image generation [3], [4].

In recent years, emerging paradigms such as multi-modal zero-shot classification have pushed the boundaries of image classification by aligning image and text embeddings to enable generalization to novel categories without direct supervision. This approach leverages cross-modal learning to tackle challenges associated with unseen data, offering new opportunities for handling complex and multimodal datasets [5].

II. Background and Related Work

A. Convolutional neural networks

CNNs, consist of neurons with weights that can be learned from data. Each neuron within a layer receives inputs from other neurons and performs a dot product. The fully connected, *i.e.* - dense layer, contains the loss function of the algorithm and they incorporate the use of a nonlinearity function [6], [7]. CNNs consist of multiple layers, including convolutional layers that detect regional patterns in an image, pooling layers that reduce image dimensionality, and fully connected layers, which work together to extract and classify features from images [6], [7]. These features provide CNNs with the ability to automatically learn hierarchical features from raw pixel data [6], [8].

B. Principal component analysis

PCA is a nonparametric statistical technique used for dimensionality reduction [9], [10]. PCA dimensionality reduction is accomplished by transforming large, high-dimensional, multicollinear data into small, lower-dimensional uncorrelated data space that are expressed as linear combinations of the original input variables. These linear combinations retain the complete information of the original data as well as the maximum variance between principal components [9], [11], [12]. For machine learning algorithms, like CNNs, PCA removes the multicollinearity between input variables that reduces the data dimensionality, eliminates redundant information in the data set, identifies the essential input variables, and reduces computational expense that lead to more accurate predictive models [10]–[12].

C. Unifold approximation and projection clustering

Uniform manifold approximation and Projection (UMAP) is a nonparametric graph-based dimensionality reduction algorithm [13]. UMAP preserves both the local and global structure of a dataset through a two phase process. A k-nearest neighbor weighted graphical representation of a data set is first computed and then is optimized using Stochastic Gradient Decent (SGD) at low-dimensional embedding of the graph. [13], [14]. This allows UMAP to be particularly effective for visualizing high-dimensional data [9], [15], [16].

D. Multi-modal zero-shot classification

The introduction of Word2vec model in 2013 illustrated the importance of word embeddings for classification tasks and drove improved methods to enhance zero-shot, *i.e.* - generalizing to unseen data sets, text classification through visual language models such as contrastive language–image pretraining (CLIP) [17], [18]. CLIP is a pre-trained multimodal visual language model that learns rich visual concepts from raw text through natural language supervision. The development of "text-to-text" as a standard input-output interface for natural language processing (NLP) has enabled zero shot transfer to downstream data sets [17]. This architecture has greatly enhanced image-to-text classification, especially as separate data sets are merged from multiple sources [19], [20]. Thus, the CLIP model has proved successful in both zero-sample visual and text tasks [17], [18], [21]–[23].

III. Method

A. Model architecture

Model is based upon well studied CNN architecture, VGGNet [24], which consists of multiple convolutional layers $\mathcal{C}_i$ followed by batch normalization layers [25] $\mathcal{B}_i$, ReLU activations $\sigma(x) = \max(0, x)$, and max-pooling layers $\mathcal{P}_i$ to extract meaningful features from the images [26]. Batch normalization is an contribution to the original VGGNet. The model processes the input x as:

$$\mathbf{f} = \mathcal{P}_n(\sigma(\mathcal{B}_n(\mathcal{C}_n(\ldots\mathcal{P}_1(\sigma(\mathcal{B}_1(\mathcal{C}_1(x))))))))$$

where $\mathcal{C}_i$ represents the convolutional layers, $\mathcal{B}_i$ denotes batch normalization [25], $\sigma(x) = \max(0, x)$ is the ReLU activation, and $\mathcal{P}_i$ refers to max-pooling layers [26]. The features $\mathbf{f}$ are then passed through fully connected layers and transformed into class probabilities $\mathbf{p}$ using a softmax function:

$$\mathbf{p}_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)},$$

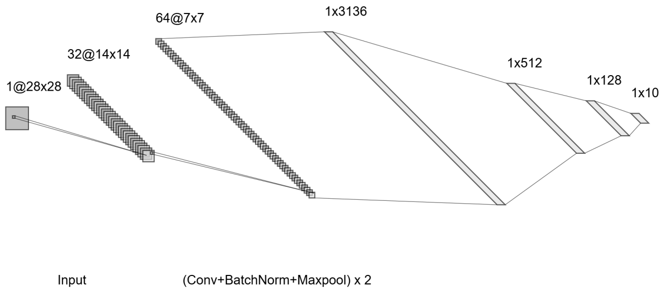


Fig. 1: Representation of the proposed CNN Architecture. Input is a single channel 28×28 image while the output is a 512 dimensional vector passed through softmax

where z_k is the output for class k and K is the total number of classes.

The training procedure utilized a Cross-Entropy Loss function, and the Adam optimizer was employed for efficient convergence. The training process was further enhanced by hyperparameter optimization using Optuna, which tuned learning rates and other training-related parameters.

IV. Experiments

A. Dataset and implementation

Dataset. The fashion MNIST dataset is a widely used benchmark that replaces the original MNIST dataset of handwritten digits [27]. It comprises 70,000 images, each labeled with one of ten categories: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. Each image is 28x28 pixels, providing a consistent input size for various algorithms. The dataset's uniform structure and standardized format make it ideal for training and testing machine learning models.

Implementation. Grayscale images from the fashion MNIST dataset were preprocessed with several transformations which included normalization to scale the pixel values to the range [-1, 1], random rotation (up to 10 degrees), and random horizontal flipping with a 50% probability to enhance the generalizability of the model. These augmentations increased the diversity of the training set and helped prevent overfitting. For training and validation, we used an 80-20 split on the training data, where 20% was randomly selected for validation purposes. The model was trained for a maximum of 15 epochs, with early stopping employed to prevent overfitting based on the validation loss.

Hyperparameter selection. Optuna [28], a hyperparameter optimization framework, was utilized to identify the best configuration. The key hyperparameters optimized were the learning rate, batch size, and the number of units in the fully connected layers. Each trial in Optuna adjusted the hyperparameters to maximize validation accuracy.

B. Results

The optimized model achieved a validation accuracy of 91%, with a corresponding F1-score of 0.89. The confusion

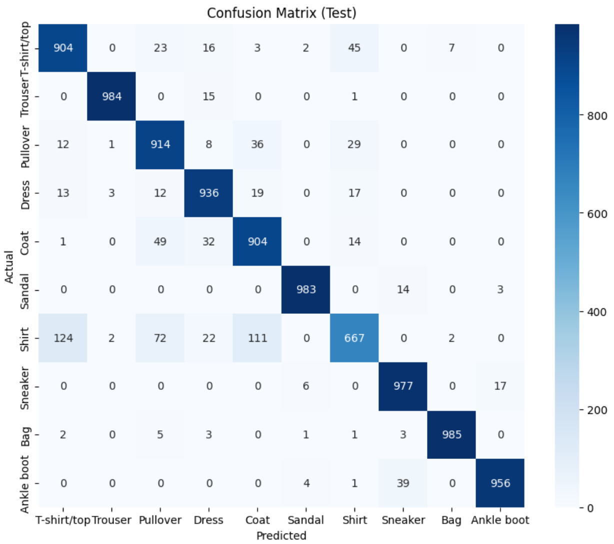
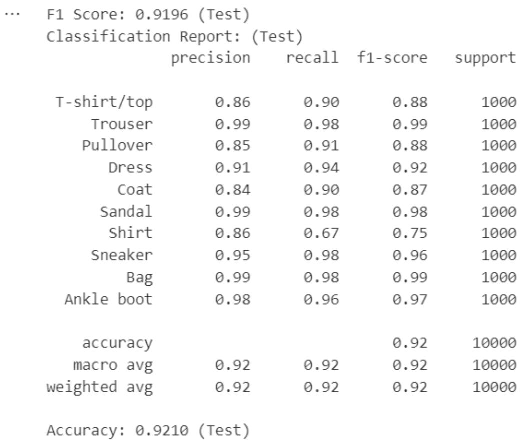


Fig. 2: Confusion matrix for test data.



```
... F1 Score: 0.9196 (Test)
    Classification Report: (Test)
                  precision    recall  f1-score   support

     T-shirt/top       0.86      0.90      0.88      1000
         Trouser       0.99      0.98      0.99      1000
        Pullover       0.85      0.91      0.88      1000
           Dress       0.91      0.94      0.92      1000
            Coat       0.84      0.90      0.87      1000
          Sandal       0.99      0.98      0.98      1000
           Shirt       0.86      0.67      0.75      1000
         Sneaker       0.95      0.98      0.96      1000
             Bag       0.99      0.98      0.99      1000
      Ankle boot       0.98      0.96      0.97      1000

        accuracy                           0.92     10000
       macro avg       0.92      0.92      0.92     10000
    weighted avg       0.92      0.92      0.92     10000

    Accuracy: 0.9210 (Test)
```

Fig. 3: Classification report for test data.

matrix, shared as Figure 2 revealed that the model was effective at distinguishing between most of the classes, with occasional misclassifications between similar items like shirts and t-shirts.

Furthermore, Figure 3 shares the classification report showing precision, recall, and F1-scores for each class. The highest accuracy was observed for the classes “Ankle Boot” and “Trouser”, while the model struggled slightly to distinguish “Shirt” from “T-shirt” due to their visual similarities.

Moreover, analysis of the misclassified samples in Figure 4 revealed that the majority of the errors occurred between visually similar classes, such as “Shirt” and “T-shirt” or “Pullover” and “Coat.” These misclassifications can be attributed to the overlapping visual features, such as similar textures and shapes. Furthermore, a detailed analysis of misclassfied samples are share in Table I.

Figure 5 presents the UMAP visualization of the model’s embeddings. The plot shows that the embedded data clusters of different classes are generally well-separated. This indicates that the model was able to learn meaningful representations for each clothing item. However, overlap was again observed between similar classes, such as ”Shirt” and ”T-shirt”, which corresponds with the observed misclassification errors.

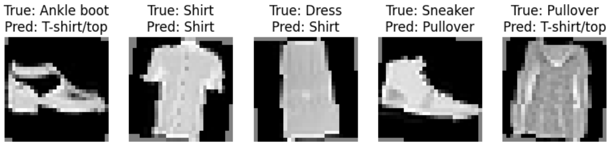


Fig. 4: Model predictions for different items including misclassification.

True Label	Predicted Label	Count
T-shirt/Top	Shirt	98
Shirt	Coat	95
Shirt	T-shirt/Top	93
Pullover	Coat	69
Pullover	Shirt	54
Shirt	Pullover	49
Coat	Shirt	42
Dress	Coat	32
Shirt	Dress	29
Ankle boot	Sneaker	27

TABLE I: Most frequent misclassifications of test data.

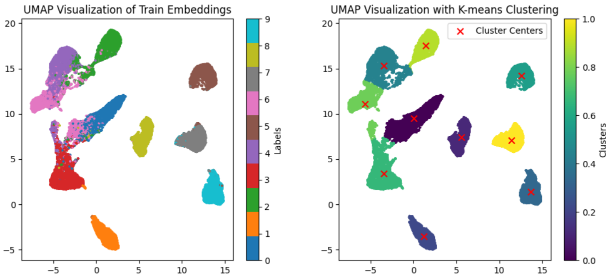


Fig. 5: UMAP visualizations of (left): train embeddings and (right): k-means clustering.

Hyperparameter tuning using Optuna led to an optimal learning rate of 0.001 and a batch size of 64, that significantly improved model convergence and generalization. Additionally, data augmentation played a crucial role in preventing overfitting (reflected in Figure 6), as observed by a 3% increase in validation accuracy compared to the non-augmented training.

Overall, the results indicate that the proposed model architecture, combined with appropriate hyperparameter tuning, and data augmentation techniques was effective in classifying the Fashion MNIST dataset with a high degree of accuracy.

C. Image-to-text classification via CLIP

Results from the CLIP classification showed worse outcomes compared to the original CNN results shared in Section IV-B. For images that were visually similar their associated text label was misclassified, such as when coats were misclassified with pullovers 802 times by the model. Moreover, the

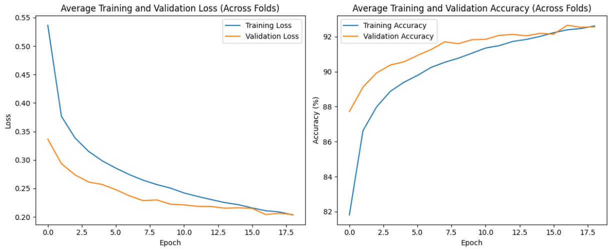


Fig. 6: Training and Validation average loss and accuracy over epochs across 5 folds.

accuracy and f1-score of the CLIP confusion matrix, 0.629 and 0.598, respectively, support the need for improving model pre-training. As it discussed by Shen et al. [18] when coupled with task-specific fine-tuning or combining CLIP with vision and language model pre-training and transferring to downstream tasks there is marked improvement in classification tasks.

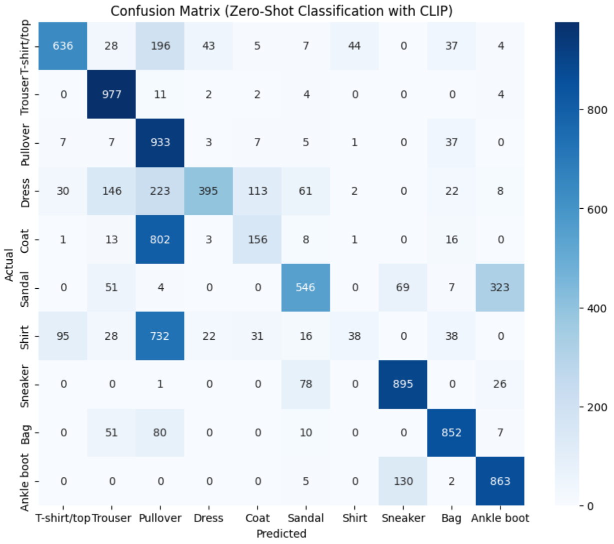


Fig. 7: Confusion matrix for CLIP based classification.

V. Conclusion and Future Work

Summary of findings. In this study, we proposed a convolutional neural network model for classifying images from the fashion MNIST dataset. Through extensive data preprocessing, hyperparameter optimization, and data augmentation techniques, a high validation accuracy of 91% was achieved. The experiments demonstrated the importance of hyperparameter tuning and data augmentation to imporve model performance and generalization.

Challenges and limitations. The analysis of misclassified samples highlights the challenge of distinguishing between visually similar classes. This suggests that additional techniques, such as advanced feature extraction or attention mechanisms [29], should be explored to further improve classification performance.

Future work. For this project, focus on incorporating advanced techniques, like residual networks [8], transformers [30], graph neural networks, [31] or transfer learning approaches to further improve classification performance by utilizing larger and more complicated datasets. Moreover, future work could also incorporate additional feature extraction methods or leverage more advanced architectures, such as attention mechanisms, to improve the model's ability to distinguish between these similar items [29].

Additionally, other advanced techniques could focus on updating the preprocessing methods used for this project. For example, using the optimization step in UMAP as a parametric algorithm to optimize the objective function of the CNN. This would capture the structure within the data set over the neural network weights allowing the CNN to learn the parametric relationship between data and embedding [13]. This should provide more meaningful representations of the data set [9], [15], [16]. An additional benefit parametric UMAP is more efficient computation by removing redundant data from the data set before the CNN algorithm is implemented [10], [12], [13].

Finally, pairing the CLIP model with other vision and language models such as visual question answering or image captioning would improve the zero-shot image-to-text classifications attempted in this project. Additional improvements to CLIP could follow the CLIP-Decoder developed by Ali et al. [19] where text and image representations are merged onto the same dimension and use CLIP's alignment loss to align them. This would enhance the multi-modal representation learning and result in better synergy between vision and language modalities.

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] S. Tasoulis, N. G. Pavlidis, and T. Roos, "Nonlinear dimensionality reduction for clustering," *Pattern Recognition*, vol. 107, p. 107508, 2020.

[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[4] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[5] 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*, pp. 8748–8763, PMLR, 2021.

[6] M. R. Karim, P. Pujari, and V. K. Ayyadevara, *Practical Convolutional Neural Networks: Implement Advanced Deep Learning Models Using Python*. Packt Publishing, May 2018.

[7] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, p. 99, 2024.

[8] 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*, pp. 770–778, 2016.

[9] A. Dalmia and S. Sia, "Clustering with umap: Why and how connectivity matters," *arXiv preprint arXiv:2108.05525*, 2021.

[10] M. S. Sulaiman, M. M. Abood, S. K. Sinnakaudan, M. R. Shukor, G. Q. You, and X. Z. Chung, "Assessing and solving multicollinearity in sediment transport prediction models using principal component analysis," *ISH Journal of Hydraulic Engineering*, vol. 27, no. sup1, pp. 343–353, 2021.

[11] Shang, Xueyi and Li, Xibing and Morales-Esteban, A. and Chen, Guanghui, "Improving microseismic event and quarry blast classification using Artificial Neural Networks based on Principal Component Analysis," *Soil Dynamics and Earthquake Engineering*, vol. 99, pp. 142–149, 2017.

[12] C. Fan, N. Zhang, B. Jiang, and W. V. Liu, "Using deep neural networks coupled with principal component analysis for ore production forecasting at open-pit mines," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 16, no. 3, pp. 727–740, 2024.

[13] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric umap embeddings for representation and semisupervised learning," *Neural Computation*, vol. 33, pp. 2881–2907, 10 2021.

[14] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[15] Allaoui, Mebarka and Kherfi, Mohammed Lamine and Cheriet, bdelhakim, "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study," in *Image and Signal Processing* (El Moataz, Abderrahim and Mammass, Driss and Mansouri, Alamin and Nouboud, Fathallah, ed.), (Cham), pp. 317–325, Springer International Publishing, 2020.

[16] M. Islam and J. Fleischer, "Manifold-aligned neighbor embedding," *arXiv preprint arXiv:2205.11257*, 2022.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[18] Shen, Sheng and Li, Liunian Harold and Tan, Hao and Bansal, Mohit and Rohrbach, Anna and Chang, Kai-Wei and Yao, Zhewei and Keutzer, Kurt , "How Much Can CLIP Benefit Vision-and-Language Tasks?," 2021.

[19] M. Ali and S. Khan, "Clip-decoder : Zeroshot multilabel classification using multimodal clip aligned representation," 2024.

[20] P. Wang, D. Li, X. Hu, Y. Wang, and Y. Zhang, "Clipmulti: Explore the performance of multimodal enhanced clip for zero-shot text classification," *Computer Speech & Language*, vol. 90, p. 101748, 2025.

[21] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2022.

[22] A. Gera, A. Halfon, E. Shnarch, Y. Perlitz, L. Ein-Dor, and N. Slonim, "Zero-shot text classification with self-training," 2022.

[23] Qin, Libo and Wang, Weiyun and Chen, Qiguang and Che, Wanxiang, "CLIPText: A New Paradigm for Zero-shot Text Classification," in *Findings of the Association for Computational Linguistics: ACL 2023* (Rogers, Anna and Boyd-Graber, Jordan and Okazaki, Naoaki, ed.), pp. 1077–1088, Association for Computational Linguistics, July 2023.

[24] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," *arXiv preprint arXiv:1606.01781*, 2016.

[25] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," *Advances in neural information processing systems*, vol. 31, 2018.

[26] H. Ide and T. Kurita, "Improvement of learning for cnn with relu activation by sparse regularization," in *2017 international joint conference on neural networks (IJCNN)*, pp. 2684–2691, IEEE, 2017.

[27] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[29] J. Luo and D. Hu, "An image classification method based on adaptive attention mechanism and feature extraction network," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, p. 4305594, 2023.

[30] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.