MennaAly at NakbaArchiveClassifier Shared Task: Transfer Learning with ResNet for Historical Image Classification

Menna Aly
Hamad Bin Khalifa University
Doha, Qatar
meal89637@hbku.edu.qa

Abstract

This paper describes our submission to the NakbaArchiveClassifier shared task at Nakba-NLP 2026, co-located with LREC 2026. The task consists of binary image classification, where a model must classify historical images into one of two categories: destruction or not_destruction. We adopt a transfer learning approach based on pretrained residual networks, fine-tuned on the provided training data. To mitigate class imbalance, we incorporate weighted cross-entropy loss during optimization. In the development phase, our ResNet18 model achieved a peak macro F1-score of 0.8137 on the validation set. For the final phase, we trained on the combined training and validation data (1,599 labeled images) and generated predictions for the hidden test set of 402 images. Our final submission achieved a macro F1-score of 0.83228 with an accuracy of 0.84577 on the official evaluation set. These results demonstrate that compact residual architectures can achieve strong performance in historical image classification under limited-data conditions without complex architectural modifications.

Keywords: binary image classification, transfer learning, residual networks, convolutional neural networks, historical image analysis

1. Introduction

The NakbaArchiveClassifier shared task aims to automatically classify historical images related to the Nakba archive into two categories: *destruction* and *not_destruction* (Abrahams et al., 2026). Such classification supports digital archiving, searchability, and historical analysis. However, the task presents challenges including limited labeled data, visual variability, and class imbalance.

In this work, we adopt a transfer learning strategy based on pretrained convolutional neural networks. Instead of training from scratch, we fine-tune ImageNet-pretrained ResNet models, which generalize well across domains (Deng et al., 2009; He et al., 2016). Our goal is not to design a complex architecture but to evaluate how far a carefully tuned baseline can go under shared task constraints.

This work provides three main contributions: (1) a transfer learning baseline using residual networks, (2) a comparison of model depth and training duration under low-resource conditions, and (3) empirical observations about overfitting behavior in historical image classification.

2. Related Work

Deep convolutional neural networks have become the standard approach for image classification. Early breakthroughs such as AlexNet demonstrated the effectiveness of deep learning for large-scale visual recognition (Krizhevsky et al., 2012). Large annotated datasets such as ImageNet have further enabled pretrained models to learn generalizable representations that transfer effectively across downstream tasks (Deng et al., 2009).

Residual Networks (ResNets) introduced by He et al. (2016) allow training deeper architectures through residual connections that mitigate vanishing gradients. Transfer learning has also been widely studied as a strong approach in scenarios with limited labeled data (Pan and Yang, 2010). In historical and archival image analysis, where annotated data is often scarce, fine-tuning pretrained convolutional neural networks (CNNs) offers a reliable and computationally efficient baseline.

3. Task and Dataset

The development dataset consists of 1,400 training images and 199 validation images. The final evaluation dataset includes 402 test images with hidden labels (Abrahams et al., 2026). The task requires classifying each image into one of two categories: *destruction* or *not_destruction*.

The development data is imbalanced, with 906 *not_destruction* samples and 494 *destruction* samples in the training split. This imbalance is addressed during training using class-weighted loss, as described in Section 4.3.

4. Methodology

4.1. Preprocessing

All images were resized to 224×224 pixels to match the input requirements of ResNet architectures. We applied standard ImageNet normalization us-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 226–228

ing mean values $[0.485, 0.456, 0.406]$ and standard deviations $[0.229, 0.224, 0.225]$.

No additional data augmentation was applied during the development phase, allowing us to evaluate the effectiveness of transfer learning without augmentation-based variability.

4.2. Model Architecture

We adopt pretrained convolutional neural networks based on the ResNet architecture (He et al., 2016). Specifically, we experiment with ResNet18 and ResNet34 models initialized with ImageNet weights.

The original fully connected layer, which outputs 1000 classes, was replaced with a linear layer of size 2 to perform binary classification. Transfer learning was selected due to the limited dataset size, which makes training deep networks from scratch impractical. Fine-tuning pretrained weights enables the model to leverage general visual representations learned from large-scale datasets.

4.3. Handling Class Imbalance

The development dataset contains more *not_destruction* samples than *destruction* samples. To mitigate bias toward the majority class, we compute class weights inversely proportional to class frequencies and incorporate them into a weighted cross-entropy loss function. This encourages stronger penalization of misclassification errors on the minority class.

4.4. Training Setup

Models were trained using the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 32. During the development phase, training was conducted for five epochs.

Performance was evaluated using macro F1-score and accuracy. All experiments were implemented in PyTorch and executed in a Google Colab environment.

5. Experiments and Results

During the development phase, we trained ResNet18 for five epochs and evaluated performance on the validation split using accuracy and macro F1-score.

The best macro F1-score (0.8137) was achieved at epoch 3. Performance slightly decreased afterward, indicating mild overfitting. This configuration achieved competitive performance in the development phase.

For the final phase, we combined the training and validation splits (1,599 images total) and trained both ResNet18 and ResNet34 models. We observed that shorter training schedules improved generalization performance on the hidden test set.

Epoch	Train Loss	Val Acc	Val Macro F1
1	0.4881	0.7990	0.7824
2	0.2037	0.7638	0.7558
3	0.0797	0.8342	**0.8137**
4	0.0569	0.8291	0.8087
5	0.0491	0.7839	0.7653

Table 1: Development phase validation performance of ResNet18.

Model	Epochs	Final Macro F1	Final Accuracy
ResNet18	2	**0.83228**	0.84577
ResNet34	6	0.82000	—

Table 2: Final phase performance comparison.

The best-performing configuration was ResNet18 trained for two epochs, achieving a macro F1-score of 0.83228 and accuracy of 0.84577. Although ResNet34 achieved low training loss, it did not outperform the smaller ResNet18 architecture, suggesting mild overfitting when training deeper models on the limited dataset.

6. Discussion

The results indicate that compact residual networks achieve competitive performance in low-resource historical image classification tasks. ResNet18 converged quickly and demonstrated strong generalization, particularly when trained for a limited number of epochs.

Longer training schedules and deeper architectures did not consistently improve macro F1 performance, highlighting the importance of controlling overfitting in small datasets. These findings suggest that carefully tuned transfer learning baselines can be highly effective without complex architectural modifications. The results also suggest that shorter training schedules can function as an implicit regularization mechanism in low-data settings.

7. Limitations

This work did not incorporate data augmentation, learning rate scheduling, or systematic hyperparameter search. Additionally, cross-validation was not performed. Future work could explore augmentation strategies, more extensive hyperparameter tuning, and ensemble approaches.

8. Conclusion

We presented a transfer learning-based system using ResNet architectures for the NakbaArchive-Classifier shared task. Our approach relied on pre-trained residual networks, weighted loss functions to address class imbalance, and minimal architectural modifications.

ResNet18 trained for two epochs achieved the best final performance, with a macro F1-score of 0.83228 on the hidden test set. The results demonstrate that lightweight CNN fine-tuning strategies can provide strong baselines for historical image classification under limited-data conditions.

9. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The NakbaArchive-Classifier Shared Task on Nakba Image Classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.