Report

Author: Sonja Nabiesade
Word Count: 467 Words

This project applies a convolutional neural network (CNN) to a binary image classification task: distinguishing frames of the cartoon characters Tom and Jerry. A CNN is appropriate here because it is state-of-the-art for image classification and is designed to learn spatial hierarchies of features directly from pixels, avoiding manual feature engineering.[1]

The dataset contains 3,187 RGB images across two classes (“jerry”, “tom”), split 80/20 into training and validation sets. A simple two-class problem is well suited to a compact CNN: the focus is on understanding the full pipeline (pre-processing, architecture, training and evaluation) rather than scaling to many classes. Images were resized to 128×128 and normalised to [0,1] to standardise input size and improve numerical stability during training.

The model architecture follows the standard CNN pattern described in the literature: repeated blocks of convolution plus max-pooling, followed by fully connected layers for classification.[2] Three Conv2D layers with 32, 64 and 128 filters and 3×3 kernels allow the network to progressively learn low-level edges and textures up to higher-level character features, while MaxPooling2D reduces spatial resolution and computation. A Dense(128) layer acts as a decision layer over the extracted features, and Dropout(0.5) is included as a regularisation mechanism to reduce overfitting by randomly disabling neurons during training.[3] The final Dense(1, sigmoid) output with a binary cross-entropy loss matches the binary nature of the task and is the standard formulation for two-class classification problems.[4] The Adam optimiser is chosen as it typically converges quickly and robustly across vision tasks without heavy tuning.

Data augmentation (random horizontal flips, small rotations and zooms) is applied to the training images. This artificially increases dataset diversity and is widely documented to improve generalisation and reduce overfitting in CNNs by reducing reliance on specific poses or viewpoints.[5] In this context, flipping and small geometric changes are realistic transformations for cartoon frames and preserve class identity.

---

[1] X. Zhao, Y. Li, and Z. Wang, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 3, pp. 1–31, 2024, doi: 10.1007/s10462-024-10721-6.

[2] X. Zhao, Y. Li, and Z. Wang, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 3, pp. 1–31, 2024, doi: 10.1007/s10462-024-10721-6.

[3] X. Zhao, Y. Li, and Z. Wang, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 3, pp. 1–31, 2024, doi: 10.1007/s10462-024-10721-6.

[4] W. Rawat and Z. Wang, “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review,” *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017, doi: 10.1162/NECO_a_00990.:contentReference[oaicite:2]{index=2}

[5] IBM, “What is data augmentation?,” IBM Think, 2023. [Online]. Available: https://www.ibm.com/think/topics/data-augmentation. [Accessed: 26-Nov-2025].

The trained model achieved around 96% validation accuracy, with a confusion matrix showing low false positives and false negatives for both classes. Precision, recall and F1-scores were high for each class, indicating that performance is balanced rather than biased towards one character. Learning curves showed decreasing training and validation loss with closely tracking accuracy, suggesting that regularisation and augmentation were effective in controlling overfitting.

However, the approach has limitations. The dataset is relatively small and specific to a single cartoon style, so generalisation to other depictions of Tom and Jerry or different domains is uncertain. The model is also relatively shallow compared to modern architectures used on large-scale benchmarks such as ImageNet, which may limit performance on more complex tasks. Potential sources of error include ambiguous or occluded frames and class imbalance if one character appears more frequently. Future work could explore transfer learning from a pretrained CNN, more systematic augmentation, or calibration of predicted probabilities to better reflect uncertainty.

References

[1] X. Zhao *et al.*, “A review of convolutional neural networks in computer vision,” *Artif. Intell. Rev.*, 2024.SpringerLink

[2] F. Sultana, A. Sufian, and P. Dutta, “Advancements in image classification using convolutional neural network,” *arXiv preprint arXiv:1905.03288*, 2019.arXiv

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012.papers.nips.cc

[4] IBM, “What is data augmentation?” IBM Think, 2023.IBM

[5] T. M. Omoniyi *et al.*, “The effect of data augmentation on performance,” *Appl. Sci.*, 2025.MDPI