

Doodle Recognition using ConvLSTM2D: A Machine Learning Approach

Raj Kanchan¹, Supriya Lokhande², Shrinivas Patil³ and Vijayalaxmi Kanade⁴

¹⁻⁴Artificial Intelligence and Data Science Department, PVG's COET & GKP(W)IoM, Pune, Maharashtra, India
Email: {rajkanchan54, vijayalaxmi.jeure, supriya0179, patilshrinivas078}@gmail.com

Abstract—Doodle recognition has emerged as a key research area in machine learning, aiming to classify hand-drawn sketches by leveraging spatiotemporal data. ConvLSTM2D models, which integrate convolutional layers with LSTM units, are highly effective in capturing both spatial features and the temporal sequence of strokes. This enables the model to understand how doodles evolve over time, improving recognition accuracy. These advancements not only support applications in education, gaming, and art but also promote accessibility by enabling expressive visual communication, especially for individuals with hearing impairments.

Index Terms— Doodle Recognition, CNN, LSTM, ConvLSTM2D, Deep Learning and Quick! Draw Dataset.

I. INTRODUCTION

Doodle recognition is an exciting field within machine learning that leverages advanced neural network architectures to interpret and classify hand-drawn sketches efficiently [1], [2]. The use of ConvLSTM2D models allows for the integration of both spatial and temporal information, making them particularly well-suited for understanding the dynamic nature of doodles as they are drawn [3],[4]. This capability not only enhances the accuracy of recognition but also opens new possibilities for applications in areas such as education, art analysis, and interactive gaming [5]. By combining convolutional layers with LSTM units, ConvLSTM2D can effectively capture the sequence of drawing strokes, enabling the model to learn patterns that represent various doodle styles and techniques [3],[6].

Moreover, the implications of advancing doodle recognition technology extend beyond mere classification; they also encompass accessibility and inclusivity in communication. For instance, as visual representation becomes increasingly vital for individuals who are deaf or hard of hearing, effective doodle classification can significantly enhance their ability to convey urgent needs through sketches [5]. This not only fosters a more inclusive environment but also encourages creativity among users, allowing them to express complex ideas without relying solely on verbal language. The use of Google's Quick, Draw! dataset is instrumental in advancing doodle recognition through machine learning [7]. This dataset contains millions of hand-drawn sketches across various categories, providing a rich resource for training models like ConvLSTM2D. The Quick, Draw! dataset not only offers a diverse range of doodle styles and techniques but also includes the temporal aspect of how sketches are drawn, which is crucial for models that capture the sequence of strokes [1]. This enables the ConvLSTM2D architecture to effectively learn from both the spatial features of the drawings and the order in which they are created.

II. BACKGROUND

A. Background of Doodle Recognition

Doodle recognition has emerged as a significant area of research within the field of machine learning, focusing on the interpretation and classification of hand-drawn sketches. This area leverages advanced neural network architectures, particularly those that can process sequential data, such as Convolutional Long Short-Term Memory (ConvLSTM) networks. These models are adept at capturing both spatial and temporal information, making them particularly effective for recognizing doodles as they are drawn [3]. The integration of convolutional layers with LSTM units allows ConvLSTM2D models to learn the sequence of drawing strokes, which is essential for understanding the dynamic nature of doodles. This capability enhances the accuracy of recognition and opens up new possibilities for applications in various fields, including education, art analysis, and interactive gaming [5],[4].

One of the most notable datasets used in doodle recognition research is Google's Quick, Draw! dataset, which comprises millions of hand-drawn sketches across diverse categories. The dataset enables researchers to achieve better accuracy in recognizing and classifying sketches by learning from both the spatial features and the order of strokes [1]. Recent works have demonstrated the effectiveness of machine learning algorithms in doodle recognition. For example, researchers have explored the use of hybrid models that combine traditional machine learning techniques with deep learning approaches, achieving significant improvements in recognition accuracy [2],[9]. Additionally, advancements in real-time doodle recognition systems have enabled instant feedback to users, highlighting the practical applications of this research [10],[6].

B. Overview of ConvLSTM2D

ConvLSTM2D is a powerful neural network architecture that integrates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) units, specifically designed to process spatiotemporal data. This model is particularly effective in tasks where both spatial and temporal information are crucial, such as video analysis, weather forecasting, and doodle recognition [3]. The architecture of ConvLSTM2D consists of convolutional layers that capture spatial features from input data, combined with LSTM units that manage the temporal dynamics. This dual capability allows the model to learn not only the static features of an image or frame but also the sequence and timing of events, which is essential for understanding how objects evolve over time [3].

In the context of doodle recognition, ConvLSTM2D excels at interpreting hand-drawn sketches by capturing the sequence of drawing strokes. This is vital for recognizing the nuances in how a doodle is created, as the order and speed of strokes can significantly influence the final interpretation [5],[4]. The model's ability to process sequential data makes it particularly suited for applications where the temporal aspect of input is as important as the spatial features. The effectiveness of ConvLSTM2D has been validated through various studies, showcasing its application in real-time systems that provide immediate feedback to users and enhance the overall accuracy of recognition tasks [10],[6]. By leveraging extensive datasets like Google's Quick, Draw!, which includes millions of hand-drawn sketches, researchers can further improve the performance of ConvLSTM2D models, allowing for better classification and understanding of diverse doodle styles [1],[2],[9].

III. PREVIOUS WORKS

A. Background of Doodle Recognition

1) CNN based approach:

Convolutional Neural Networks (CNNs) have been widely used for doodle recognition due to their effectiveness in capturing spatial features. Reference [2] introduced a novel sketch recognition model based on CNNs, demonstrating the model's ability to identify patterns in hand-drawn sketches with high accuracy. This research highlighted the importance of network depth and optimized convolutional layers in achieving robust recognition results.

2) LSTMs and Attention:

LSTMs, often paired with attention mechanisms, have also shown promising results in doodle recognition. Reference [4] proposed a sketch recognition model using LSTMs with an attention mechanism and a Minimum Cost Flow Algorithm, achieving improved accuracy by selectively focusing on the most relevant parts of the sketches. This approach effectively handled complex and overlapping strokes by emphasizing critical features during the recognition process.

3) *Hybrid Approaches:*

Hybrid models combining CNNs with other architectures have further enhanced doodle recognition. Reference [5] presented a combined Deep CNN-LSTM network for sketch recognition, demonstrating the effectiveness of merging spatial and sequential feature extraction. Their study emphasized the benefits of utilizing both local and global features for accurate classification.

4) *Real Time Systems:*

Real-time systems have also been a focus of recent research. Reference [10] developed an LSTM-based character recognition system for text extraction from natural scenes. Although not directly applied to doodle recognition, the real-time nature of their system showcases the potential for interactive applications. Additionally, [9] applied CNN-based methods for recognizing handwritten mathematical expressions, demonstrating the adaptability of CNNs in understanding free-form sketches.

5) *Quick! Draw Dataset:*

Large-scale datasets like Google's Quick, Draw! have been instrumental in training and evaluating doodle recognition models. Reference [1] conducted a meta-analysis of various deep learning models applied to doodle recognition using this dataset, providing insights into the effectiveness of different architectures. Their study underlined the importance of data diversity and augmentation techniques in enhancing model performance.

B. Role of ConvLSTM2D in Image Processing

ConvLSTM2D has been extensively used for video analysis tasks such as action recognition, motion prediction, and video anomaly detection. Reference [3] demonstrated the effectiveness of ConvLSTM2D in analyzing video frames by leveraging both spatial features and temporal dependencies. This architecture has also been widely applied in applications requiring scene understanding and behavior prediction. In medical image processing, ConvLSTM2D has shown great promise in tasks like disease diagnosis and organ segmentation. Researchers have employed ConvLSTM2D networks to analyze sequential MRI and CT scan images, providing accurate predictions through temporal information modeling. For instance, [3] applied ConvLSTM2D to classify time-series medical images, improving diagnostic accuracy. Another notable application of ConvLSTM2D is in weather prediction, where it processes satellite images to forecast precipitation patterns [8]. The network's ability to model temporal changes in atmospheric conditions has made it a valuable tool in meteorological studies.

IV. METHODOLOGY

A. Data Collection and Preprocessing

1) *Data Collection*

Google's Quick, Draw! dataset is a large-scale collection of hand-drawn sketches, consisting of over 50 million drawings across 345 categories. The dataset was originally created as part of an interactive game that collected user-drawn doodles from a global audience. This dataset is publicly available and provides both the final static image and the sequential stroke data, making it suitable for tasks involving spatiotemporal analysis. For this study, the dataset was accessed using Google's provided download links, selecting categories that align with the research objectives. Only a subset of categories, representing common objects, was used to ensure balanced representation across different classes.

2) *Data Preprocessing*

After the data collection, the following preprocessing steps were applied to clean and standardize the dataset:

- **Data Filtering:** Drawings that were incomplete, corrupted, or contained fewer than three strokes were removed.
- **Normalization:** All sketches were resized to a fixed image dimension (e.g., 64x64 pixels) to ensure consistency across the dataset.
- **Grayscale Conversion:** Since the dataset primarily consists of line drawings, color information was discarded, converting all images to grayscale.

Fig. 1 below shows some doodles while also highlighting the strokes with each colour representing a different stroke.

3) *Stroke Data Processing*

The sequential stroke data was processed to preserve temporal information for models like ConvLSTM2D. The preprocessing steps included:

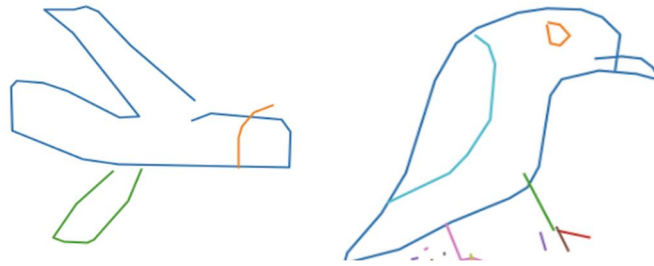


Figure 1. Sample doodle of an airplane and a bird from the training dataset

- **Stroke Padding:** Drawings with fewer strokes were padded with empty strokes to maintain a uniform sequence length.

Normalization: Stroke coordinates were normalized to the range $[0, 1]$ to prevent scale variations from influencing the model.

To prepare the stroke-based doodles from the Quick, Draw! dataset for ConvLSTM2D input, each drawing is converted into a sequence of image frames. These frames simulate the temporal order in which the doodle was drawn, with each frame showing the sketch as it progresses stroke by stroke. This sequential representation allows the model to capture both spatial features and the temporal evolution of the drawing.

4) Data Splitting

The dataset was split into training and validation sets using a ratio of 80:20. Class balancing was maintained to ensure fair representation of all categories.

Fig 2. below illustrates the construction of a "cup" doodle.

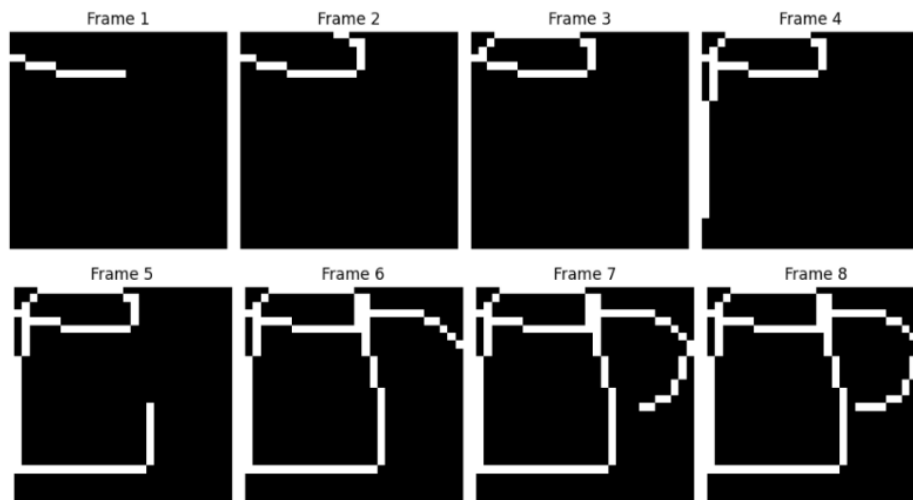


Figure 2. Frame-wise construction of a "cup" doodle across 8 time-steps

B. Model Architecture

The proposed model leverages ConvLSTM2D layers to effectively capture both spatial and temporal dependencies inherent in sequential image data. This makes it well-suited for tasks like doodle recognition, where each drawing is represented as a sequence of grayscale images. The model begins with an input layer that accepts a sequence of single-channel images of shape $(\text{sequence_length}, \text{image_size}, \text{image_size}, 1)$.

The first ConvLSTM2D layer consists of 64 filters with a kernel size of $(3,3)$, using 'same' padding to preserve spatial dimensions and `return_sequences = True` to maintain the temporal structure. This is followed by batch normalization to stabilize learning and MaxPooling3D with a pool size of $(1, 2, 2)$, reducing spatial dimensions while preserving the sequence. A second ConvLSTM2D layer with 96 filters and the same kernel size continues the temporal modeling, again followed by normalization and pooling. The final ConvLSTM2D layer uses 128 filters, with `return_sequences = False` to collapse the temporal dimension before classification. After a final batch normalization step, a flatten layer converts the multidimensional output into a vector. This is followed by two fully connected layers with 256 and 128 neurons respectively, each using ReLU activation and dropout for regularization. The final dense layer applies a softmax activation to produce class probabilities for the output.

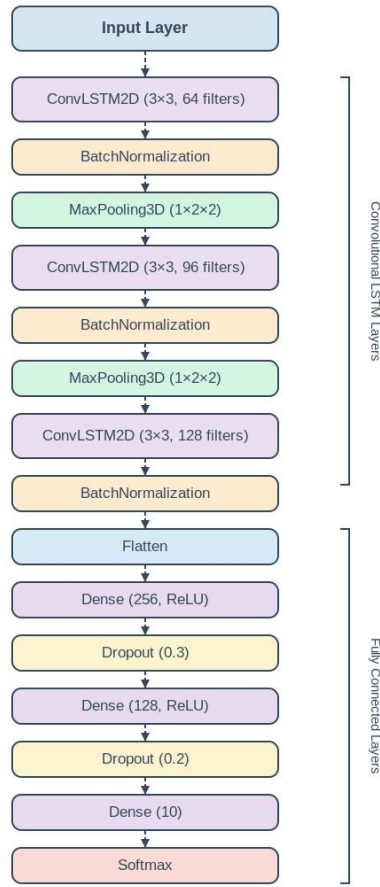


Figure 3. Architecture of the proposed ConvLSTM-based doodle recognition model

C. Training the model

The dataset used for training consists of doodle drawings in the form of stroke-based data from the Google Quick, Draw! Dataset. The dataset was preprocessed by normalizing the strokes using a custom normalization function that scaled the stroke coordinates to a range of $[0, 1]$. Each drawing was padded or truncated to ensure a consistent sequence length of 100 points. Labels were encoded using integer values representing each class. Model was compiled using the Adam optimizer with a learning rate of 0.001. Sparse categorical cross-entropy was used as the loss function since the problem is a multi-class classification task. Accuracy was used as the primary evaluation metric. The model was trained using a batch size of 32 and for a total of 9 epochs. Model checkpoints were used to save the best model based on validation accuracy. Fig 4. shows the model loss and model accuracy over number of epochs.

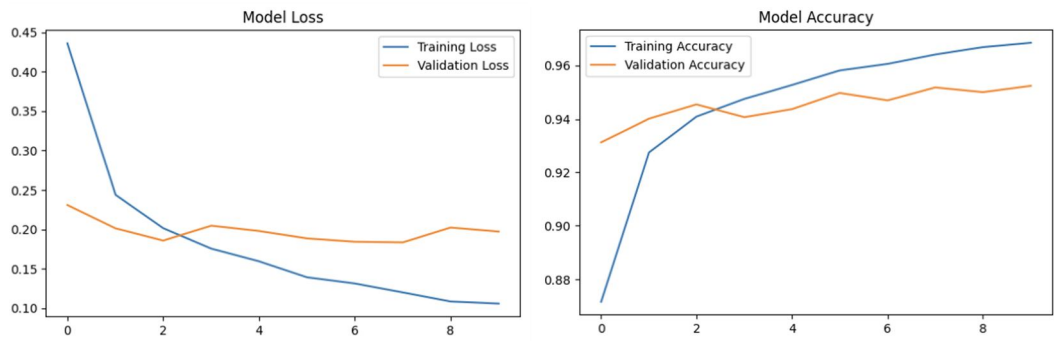


Figure 4. Model Loss and Model Accuracy across 9 epochs

V. RESULTS

To comprehensively evaluate the performance of the ConvLSTM2D-based doodle recognition model, we employed precision, recall, and F1-score as the evaluation metrics. Table I shows the F1-scores along with precision and recall of 10 classes that were trained over 10,000 images.

TABLE I. CLASSIFICATION REPORT OF THE CLASSES TRAINED

Class	Precision	Recall	F1-score
sun	98%	97%	97%
bicycle	98%	97%	97%
sock	97%	95%	95%
bird	88%	89%	89%
cup	94%	94%	94%
car	95%	96%	96%
airplane	91%	92%	92%
house	98%	98%	98%
tree	97%	97%	97%
t-shirt	97%	96%	97%

The accuracy was computed to assess the efficacy of the suggested method, indicating the extent to which the classifier can correctly classify the data. Reference [5] evaluates three architectures CNN, LSTM, and a combined CNN-LSTM network for sketch recognition. Our proposed ConvLSTM-based model outperforms these configurations by more effectively capturing both spatial and temporal patterns within stroke sequences. While both approaches leverage spatiotemporal modeling, our deeper ConvLSTM stack with structured pooling and regularization demonstrates superior performance. Specifically, we achieve higher F1-scores and overall accuracy, highlighting the effectiveness of our design in capturing complex stroke dynamics and improving generalization. A side-by-side comparison of key evaluation metrics is presented in Table II.

TABLE II. COMPARISON OF EVALUATION METRICS WITH OTHER HIGH PERFORMING MODELS

Model	Accuracy	F1 score
CNN	90.5%	90.4%
LSTM	90.9%	90.9%
CNN-LSTM	94.9%	94.9%
ConvLSTM	96.5%	96.5%

VI. CONCLUSIONS

The ConvLSTM2D-based model demonstrated strong capabilities in recognizing hand-drawn doodles by leveraging both spatial and temporal information. It was observed that drawings with a consistent and well-defined stroke order yielded higher classification accuracy. Conversely, categories with diverse or ambiguous stroke sequences showed reduced accuracy, highlighting the importance of temporal consistency in sketch-based recognition tasks. Future improvements could focus on expanding the number of classes to cover a wider range of doodle categories, enhancing the model's generalizability. Additionally, deploying a larger and deeper neural architecture may lead to improved performance, especially for more complex sketches. However, the current implementation was constrained by limited computational resources, and future experiments with more powerful hardware could enable the training of more sophisticated models for better accuracy and robustness.

REFERENCES

- [1] Sethi, P., Duhan, S., Gupta, S., & Jain, G. (2021). Meta Analysis of Deep Learning Models for Doodle Recognition. <https://doi.org/10.1109/CONIT51480.2021.9498373>

- [2] Kabakus, A. T. (2020). A Novel Sketch Recognition Model based on Convolutional Neural Networks. 1–6. <https://doi.org/10.1109/HORA49412.2020.9152911>
- [3] Lee, S.-I., Song, S., & Ko, D.-S. (2021). Performance Improvement for Automatic Classification of Partial Charge Types using ConvLSTM2D Technique. 19(4), 19–26. <https://doi.org/10.14801/JKIIT.2021.19.4.19>
- [4] Nguyen-Xuan, B. (2019). Sketch Recognition Using LSTM with Attention Mechanism and Minimum Cost Flow Algorithm. *International Journal of Contents*, 15(4), 8–15. <https://doi.org/10.5392/IJOC.2019.15.4.008>
- [5] El Mouna, L., Silkan, H., Koumetio Tekouabou, S. C., Hanyf, Y., tourad, M. C., & Nanne, M. F. (2024). A Combined Deep Cnn-Lstm Network For Sketch Recognition. *International Journal of Computing and Digital Systems*. <https://doi.org/10.12785/ijcds/160147>
- [6] Song, X. (2025). Exploration and Research of Convolutional Neural Networks in Image Recognition. *Applied and Computational Engineering*, 121(1), 58–65. <https://doi.org/10.54254/2755-2721/2025.19558>
- [7] Fernandez-Fernandez, R., Victores, J. G., Estevez, D., & Balaguer, C. (2019). Quick, stat!: A statistical analysis of the quick, draw! dataset. *arXiv preprint arXiv:1907.06417*
- [8] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28, 802–810.
- [9] Rowney, T., & Iliev, A. I. (2024). Recognition of Handwritten Mathematical Expressions Using Systems of Convolutional Neural Networks. *Serdica Journal of Computing*, 17(2), 107–116. <https://doi.org/10.55630/sjc.2023.17.107-116>
- [10] Soni, V., Chouksey, P., Tandan, S. R., Pimpalkar, A., Miri, R., & Nema, N. K. (2024). LSTM Based Character Recognition for Text Extraction in Real-Time Natural Scenes Images. 1–7. <https://doi.org/10.1109/icdsns62112.2024.10691042>.