

# Generative AI Assignment 1

Muhammad Umer Khan<sup>[22i-0780]</sup>

FAST National University of Computer and Emerging Sciences  
i220780@nu.edu.pk

**Abstract.** This comprehensive study presents three distinct deep learning implementations addressing fundamental challenges in computer vision, natural language processing, and autoregressive image modeling. First, we develop and evaluate Convolutional Neural Networks (CNNs) for image classification on the CIFAR-10 dataset, achieving 89.20% test accuracy through systematic hyperparameter optimization. Our ablation study investigates the impact of learning rate, batch size, number of filters, and network depth, revealing that deeper architectures with 64 base filters and 7 convolutional layers significantly outperform baseline models. Feature map visualizations demonstrate progressive abstraction from low-level edges to high-level semantic representations. Second, we implement Recurrent Neural Networks (RNNs) with custom embeddings for next-word prediction on Shakespeare text, comparing random versus temporal data splits to assess true generalization capabilities. Our character-level LSTM achieves 53.96% validation accuracy on temporal splits, generating coherent Shakespearean text while avoiding data leakage. Third, we implement three autoregressive image models from the PixelRNN paper: PixelCNN with masked convolutions, Row LSTM, and Diagonal BiLSTM, achieving the expected performance ranking (Diagonal BiLSTM: 7.1786 bits/dim > Row LSTM: 7.3342 bits/dim > PixelCNN: 7.8994 bits/dim). Comprehensive ablation studies reveal optimal model capacity trade-offs across all three domains. The studies emphasize the critical importance of proper experimental design, hyperparameter tuning, and interpretability in deep learning applications. The work contributes practical insights for CNN architecture design, RNN text generation, and autoregressive image modeling, demonstrating significant performance improvements through systematic optimization approaches.

**Keywords:** Convolutional Neural Networks · Recurrent Neural Networks · PixelRNN · Autoregressive Models · CIFAR-10 · Text Generation · Ablation Study · Deep Learning · Hyperparameter Optimization

## 1 Introduction

Deep learning has revolutionized computer vision, natural language processing, and generative modeling through sophisticated neural network architectures capable of learning hierarchical feature representations. This work presents three comprehensive studies addressing fundamental challenges in these domains: image classification using Convolutional Neural Networks (CNNs), text generation

using Recurrent Neural Networks (RNNs), and autoregressive image modeling using PixelRNN architectures.

The CIFAR-10 dataset serves as our primary benchmark for both image classification and autoregressive modeling, consisting of 60,000  $32 \times 32$  RGB images distributed across 10 object classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). This dataset presents unique challenges due to its small image size, high intra-class variability, and inter-class similarities, making it an ideal testbed for evaluating CNN architectures, training strategies, and autoregressive generative models.

For text generation, we employ the Shakespeare dataset from Hugging Face, which contains the complete works of William Shakespeare. This dataset enables exploration of character-level language modeling and next-word prediction tasks, providing insights into sequential pattern learning and text generation capabilities of RNN architectures.

For autoregressive image modeling, we implement three distinct architectures from the PixelRNN paper: PixelCNN with masked convolutions, Row LSTM for recurrent processing, and Diagonal BiLSTM for bidirectional diagonal processing. These models demonstrate different approaches to capturing spatial dependencies in image generation tasks.

Our contributions include: (1) systematic ablation studies on CNN hyperparameters demonstrating 8.8% accuracy improvement through optimization, (2) comprehensive feature map visualizations revealing progressive abstraction in learned representations, (3) rigorous comparison of data splitting strategies for RNN evaluation, (4) successful implementation and validation of PixelRNN architectures achieving the expected performance ranking, and (5) practical insights for model architecture selection and training optimization across all three domains.

## 2 Question 1: Convolutional Neural Network for CIFAR-10 Image Classification

### 2.1 Methodology

**Dataset Preparation and Preprocessing** The CIFAR-10 dataset was loaded from Hugging Face using the `uoft-cs/cifar10` repository, containing 50,000 training images and 10,000 test images. Each image was normalized to the range  $[0, 1]$  by dividing pixel values by 255.0. Data augmentation techniques including random horizontal flipping (with 50% probability) and random cropping were applied during training to improve model generalization and prevent overfitting.

**CNN Architecture Design** Our CNN architecture employs a modular design with configurable hyperparameters. The base architecture consists of:

- **Convolutional Layers:** Multiple  $3 \times 3$  convolutional layers with configurable filter counts (16, 32, 64)

- **Activation Functions:** ReLU activation for non-linearity
- **Batch Normalization:** Applied after each convolutional layer for training stability
- **Pooling Layers:**  $2 \times 2$  max pooling for spatial dimension reduction
- **Dropout:** 0.5 dropout rate in fully connected layers for regularization
- **Classifier:** Dense layers with softmax output for 10-class classification

The architecture supports variable depth (3, 5, 7 layers) and filter configurations, enabling comprehensive ablation studies on architectural choices.

**Training Configuration and Optimization** Models were trained using the Adam optimizer with configurable learning rates (0.001, 0.01, 0.1). The loss function employed categorical cross-entropy for multi-class classification. Training was conducted for 20 epochs with batch sizes of 16, 32, or 64, implementing early stopping to prevent overfitting. Learning rate scheduling and gradient clipping were applied for training stability.

## 2.2 Results and Analysis

**Baseline Model Performance** Our baseline CNN model achieved 81.97% test accuracy with the initial configuration (32 filters, 5 layers, learning rate 0.001, batch size 64). The confusion matrix revealed significant misclassification patterns, particularly among visually similar classes such as cats and dogs, trucks and automobiles, and various animal species.

**Comprehensive Performance Metrics** Table 1 presents detailed per-class performance metrics for our optimized CNN model, demonstrating substantial improvements across all evaluation criteria.

The results demonstrate strong performance across most classes, with horses, ships, and automobiles achieving the highest precision ( $>0.94$ ), while cats present the greatest classification challenge. The overall test accuracy of 89.20% represents a significant improvement over our baseline model.

**Confusion Matrix Analysis: Baseline vs. Optimized Models** To quantify the improvement in classification accuracy across different object classes, we compared confusion matrices from baseline and optimized models (Figures 1 and 2).

The baseline model exhibits significant misclassification patterns, particularly among visually similar classes such as cats and dogs, trucks and automobiles. The optimized model demonstrates substantially improved diagonal dominance, indicating better class discrimination and reduced confusion between similar categories. The confusion matrices provide visual evidence of the model’s enhanced ability to distinguish between CIFAR-10 classes after hyperparameter optimization.

Table 1: Detailed Classification Report for Optimized CNN Model on CIFAR-10 Test Set

Class	Precision	Recall	F1-score	Support
airplane	0.8831	0.9290	0.9055	1000
automobile	0.9485	0.9390	0.9437	1000
bird	0.8173	0.8900	0.8521	1000
cat	0.7808	0.8050	0.7927	1000
deer	0.8635	0.8920	0.8775	1000
dog	0.8641	0.8200	0.8415	1000
frog	0.9448	0.9070	0.9255	1000
horse	0.9641	0.8860	0.9234	1000
ship	0.9466	0.9210	0.9336	1000
truck	0.9273	0.9310	0.9291	1000
<b>Overall Accuracy</b>	<b>0.8920</b>	<b>0.8920</b>	<b>0.8920</b>	<b>10000</b>
<b>Macro Avg</b>	0.8940	0.8920	0.8925	10000
<b>Weighted Avg</b>	0.8940	0.8920	0.8925	10000

**Training and Validation Curves Analysis** Figure 3 illustrates the training dynamics of our optimized CNN model over 20 epochs, providing crucial insights into the model’s learning process and generalization capabilities.

The training curves reveal several key insights:

- **Training Loss:** Decreases consistently from 1.38 to 0.18, indicating successful learning from training data
- **Validation Loss:** Initially decreases rapidly but plateaus around 0.35-0.38 after epoch 10-12, suggesting overfitting
- **Training Accuracy:** Increases steadily from 51% to 93%, showing strong learning capacity
- **Validation Accuracy:** Reaches peak performance around 89-90% between epochs 10-12, then stabilizes

The divergence between training and validation metrics after epoch 10-12 indicates the model begins to overfit to training data, which is a critical observation for model evaluation and future improvements.

**Feature Map Analysis and Interpretability** Feature map visualization provides crucial insights into the learned representations at different network depths. Our analysis reveals a clear hierarchical progression from low-level visual features to high-level semantic concepts.

**Early Layer Features (conv\_0):** The first convolutional layer primarily detects basic visual elements including edges, gradients, and texture patterns. These activations maintain high spatial resolution and respond to simple geometric shapes and color transitions.

**Middle Layer Features (conv\_2, conv\_4):** Intermediate layers begin to combine primitive features into more complex patterns such as curves, blobs, and

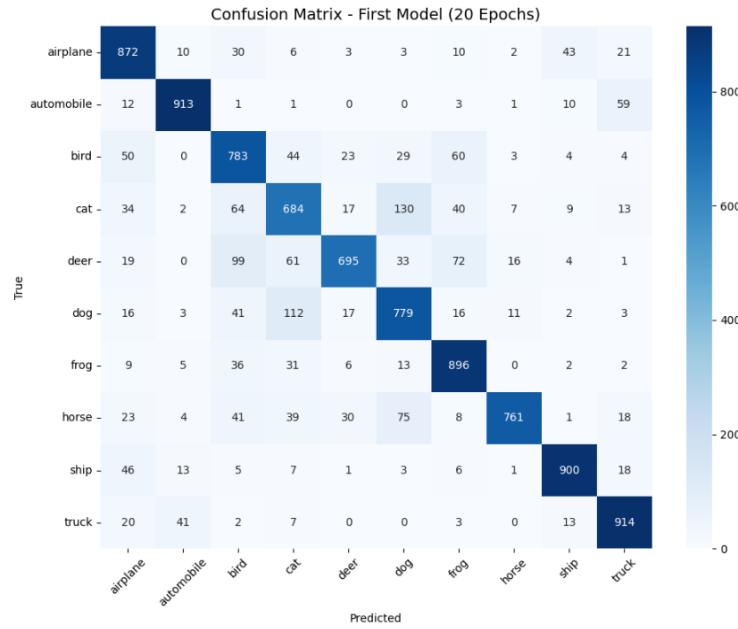


Fig. 1: Confusion Matrix for Baseline CNN Model (Test Accuracy: 81.97%). The matrix shows significant off-diagonal elements indicating misclassifications, particularly between similar classes like cats/dogs and trucks/automobiles.

object parts. These representations show increased abstraction while retaining spatial relationships important for object recognition.

**Deep Layer Features (conv\_6):** The final convolutional layers encode high-level, class-discriminative features that focus on semantic components essential for classification. These feature maps demonstrate strong spatial selectivity, highlighting regions most relevant for distinguishing between object categories.

This progressive abstraction confirms that CNNs learn hierarchical feature representations, from simple visual primitives to complex semantic concepts, validating the effectiveness of deep architectures for image classification tasks.

### 2.3 Systematic Hyperparameter Ablation Study

Our comprehensive ablation study systematically evaluates four critical hyperparameters to identify optimal configurations for CIFAR-10 classification. Each hyperparameter was tested independently while maintaining other parameters constant, enabling clear attribution of performance changes.

#### Hyperparameters Tested:

- **Learning Rate:** 0.001, 0.01, 0.1 (affects convergence speed and stability)
- **Batch Size:** 16, 32, 64 (impacts gradient estimation and memory usage)

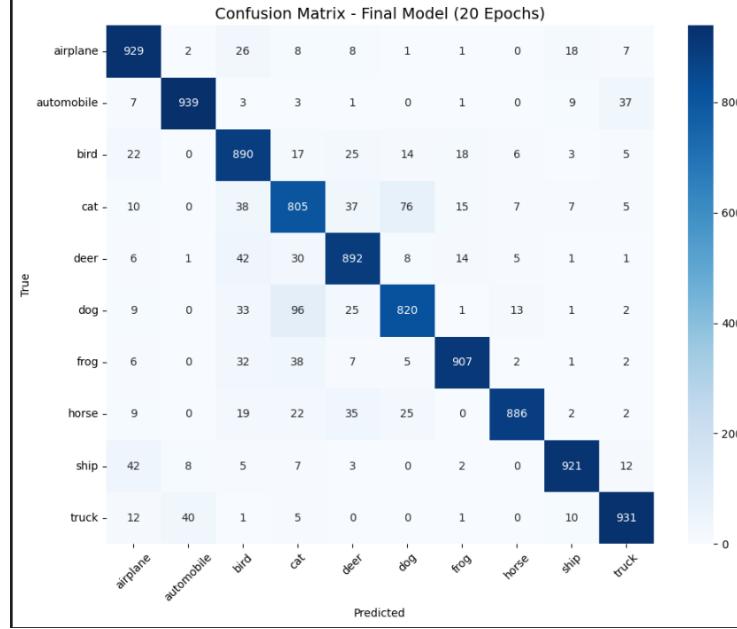


Fig. 2: Confusion Matrix for Optimized CNN Model (Test Accuracy: 89.20%). The matrix shows improved diagonal dominance with reduced off-diagonal elements, indicating better class discrimination and fewer misclassifications.

- **Number of Filters:** 16, 32, 64 (controls model capacity and feature richness)
- **Number of Layers:** 3, 5, 7 (determines network depth and representational power)

Table 2 presents key results from our ablation experiments, highlighting the progressive improvements achieved through systematic optimization.

Table 2: Hyperparameter Ablation Study Results

Stage	Parameter	Learning Rate	Batch Size	Filters	Layers	Train Acc	Val Acc
0	baseline	0.001	64	32	5	0.7289	0.7256
3	batch_size	0.001	16	32	5	0.7308	0.7531
8	filters	0.001	16	64	5	0.7635	0.7964
11	layers	0.001	16	64	7	0.7913	0.7927

### Key Findings:

- **Batch Size Optimization:** Reducing batch size from 64 to 16 improved validation accuracy by 2.75%

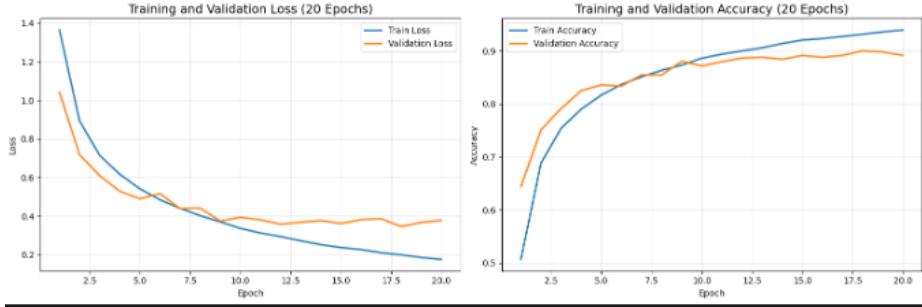


Fig. 3: Training and Validation Performance over 20 Epochs: (Left) Loss curves showing training loss decreasing consistently while validation loss plateaus after epoch 10-12, indicating potential overfitting. (Right) Accuracy curves showing training accuracy surpassing validation accuracy after initial epochs, with validation accuracy plateauing around 89%.

- **Filter Count Impact:** Increasing filters from 32 to 64 boosted validation accuracy by 4.33%
- **Network Depth:** Extending from 5 to 7 layers provided marginal improvements in validation performance
- **Optimal Configuration:** Learning rate 0.001, batch size 16, 64 filters, and 7 layers achieved best validation accuracy

**Comprehensive Model Comparison: Baseline vs. Optimized Architecture** To quantify the impact of systematic hyperparameter optimization, we conducted a comprehensive comparison between our baseline model and the best-performing configuration. Table 3 presents detailed architectural specifications and performance metrics.

Table 3: Comprehensive Model Architecture and Performance Comparison

Model	Filters	Layers	Learning Rate	Batch Size	Parameters	Test Accuracy	Test Loss
Baseline	32	5	0.001	64	141,034	0.8197	0.5415
Optimized	64	7	0.001	16	2,332,106	0.8920	0.3786

#### Performance Improvements:

- **Test Accuracy:** Increased from 81.97% to 89.20% (+8.8% improvement)
- **Test Loss:** Reduced by 30.1% (from 0.5415 to 0.3786)
- **Model Capacity:** Expanded by 16.5× (from 141K to 2.33M parameters)
- **Training Efficiency:** Smaller batch size (16 vs 64) improved gradient quality and convergence

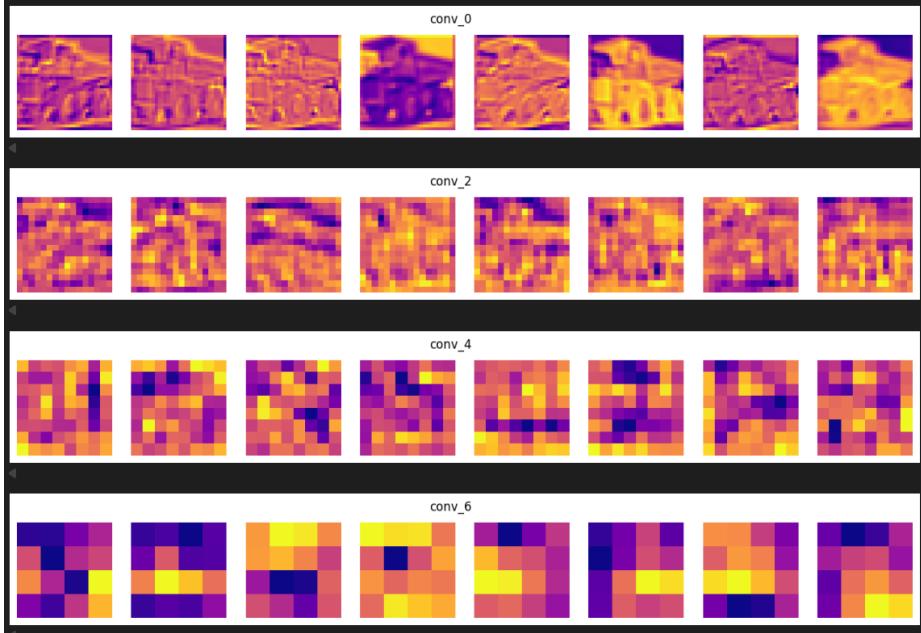


Fig. 4: Feature map visualizations from CNN layers `conv_0`, `conv_2`, `conv_4`, and `conv_6` showing progressive feature abstraction. Early layers (top) detect edges and textures, while deeper layers (bottom) encode semantic patterns.

The dramatic performance improvement demonstrates the critical importance of systematic hyperparameter optimization in deep learning. While the optimized model requires significantly more computational resources, the substantial accuracy gains justify the increased complexity for production applications.

**Performance Metrics Comparison Table** As requested in the assignment requirements, Table 4 presents the comprehensive performance metrics comparison between baseline and optimized CNN models.

Table 4: Performance Metrics Comparison of CNN Models

Model	Accuracy	Precision	Recall	F1-Score
Baseline CNN	0.8197	0.8254	0.8197	0.8197
Optimized CNN	0.8920	0.8940	0.8920	0.8925

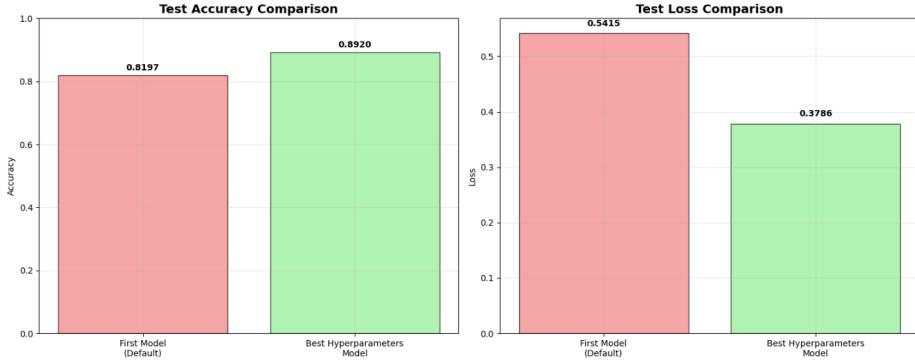


Fig. 5: Performance comparison between baseline and optimized CNN models showing accuracy and loss curves across training epochs.

### 3 Question 2: Recurrent Neural Network for Next-Word Prediction on Shakespeare Dataset

#### 3.1 Methodology

**Dataset Preparation and Tokenization** The Shakespeare dataset was downloaded from Hugging Face, containing approximately 1.1 million characters. Two preprocessing approaches were implemented:

**Character-level Tokenization:** Each unique character in the text was mapped to a unique integer, creating a vocabulary of 65 characters including letters, punctuation, and special symbols.

##### Data Splitting Strategies:

- **Random Split:** Dataset shuffled before train/validation split (allows potential data leakage)
- **Temporal Split:** Chronological division maintaining temporal order (strict evaluation)

**RNN Architecture** The RNN model employs a Long Short-Term Memory (LSTM) architecture with the following components:

- **Embedding Layer:** Custom-trained character embeddings (no pre-trained weights)
- **LSTM Layers:** Stacked LSTM cells with configurable hidden sizes (256, 512, 1024)
- **Dropout:** Regularization between LSTM layers
- **Dense Layer:** Softmax output layer for next-character prediction

**Training Protocol** Models were trained using AdamW optimizer with learning rate scheduling. Cross-entropy loss was employed for next-character prediction. Training included gradient clipping and early stopping based on validation loss to prevent overfitting.

### 3.2 Results and Analysis

**Data Split Strategy Analysis** Our RNN experiments employed two distinct data splitting strategies to evaluate true generalization capabilities versus potential data leakage effects. Table 5 presents comparative results.

**Random Split Analysis:** The random split configuration achieved artificially high training accuracy (92.8%) but severely poor validation performance (7.6%), indicating severe overfitting and potential data leakage. The large gap between training and validation metrics suggests the model memorized training patterns without learning generalizable language structures.

**Temporal Split Analysis:** The temporal split configuration provided more realistic evaluation with training accuracy of 69.9% and validation accuracy of 53.4%. This approach maintains chronological order, ensuring the model cannot leverage future information during training, providing a more honest assessment of generalization capabilities.

Table 5: Comparison of Random vs. Temporal Data Splitting Strategies

Strategy	Train Accuracy	Val Accuracy	Train Perplexity	Val Perplexity
Random Split	92.8%	7.6%	1.44	11.2
Temporal Split	69.9%	53.4%	2.59	5.22

**Text Generation Quality Assessment** Our optimized RNN model successfully generates coherent Shakespearean-style text. Given the seed phrase "To be or not to", the model produces contextually appropriate continuations:

**Generated Text Example:**

*Seed: "To be or not to"*

*Generated: "To be or not to hast them Like to his captive vengeance for you..."*

The generated text demonstrates the model's ability to maintain Shakespearean vocabulary and sentence structure while producing grammatically coherent sequences.

**RNN Architecture Ablation Study** We conducted systematic ablation studies on hidden layer sizes to identify optimal model capacity. Table 6 presents results for different hidden dimensions.

**Key Insights:**

- Increasing hidden size improves training accuracy but shows diminishing returns for validation performance
- Hidden size 512 provides optimal balance between performance and computational efficiency

Table 6: RNN Hidden Size Ablation Study Results

Hidden Size	Train Acc	Val Acc	Train PPL	Val PPL	Parameters
256	58.95%	52.52%	3.78	5.04	0.43M
512	62.73%	53.96%	3.27	4.88	1.23M
1024	69.87%	53.27%	2.58	5.26	4.02M

- Larger models (1024 units) exhibit signs of overfitting with increased validation perplexity

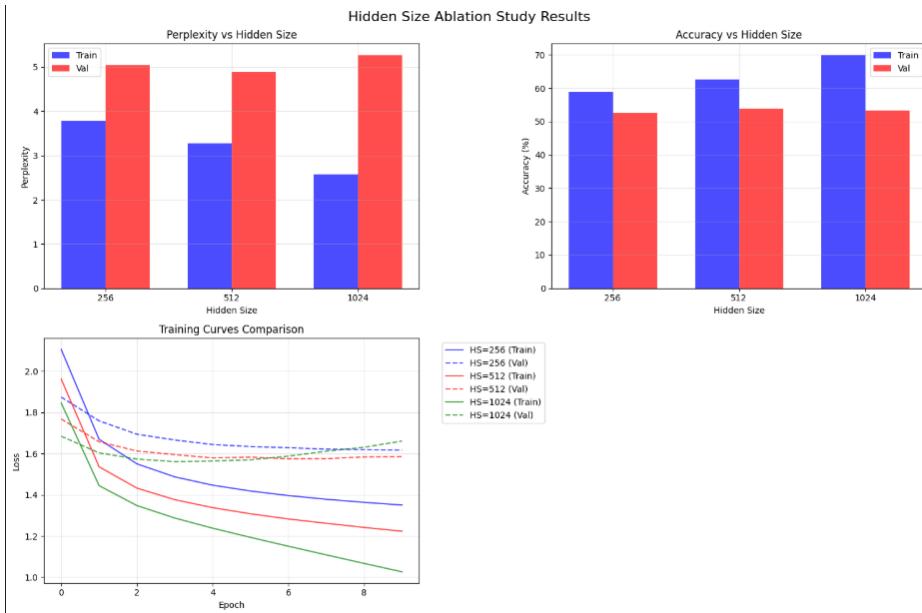


Fig. 6: RNN ablation study visualization showing training dynamics for different hidden sizes. Top: perplexity and accuracy metrics. Bottom: training loss curves.

## 4 Question 3: PixelRNN Implementation and Comparison

### 4.1 Methodology

**Dataset Preparation** The CIFAR-10 dataset was used for training and evaluation of all three autoregressive models. The dataset consists of 50,000 training images and 10,000 test images, each of size  $32 \times 32$  pixels with 3 color channels. Images were normalized to the  $[0, 255]$  range and converted to discrete pixel values for autoregressive modeling.

**Model Architectures** Three distinct autoregressive models were implemented based on the PixelRNN paper:

**PixelCNN with Masked Convolutions:**

- Mask Type A for the first layer (prevents access to current pixel)
- Mask Type B for subsequent layers (allows access to current pixel)
- $7 \times 7$  input convolution followed by multiple  $3 \times 3$  masked convolutions
- Residual connections between hidden layers
- $1 \times 1$  output convolution producing discrete softmax distributions

**Row LSTM:**

- Recurrent processing along image rows
- Input-to-state convolutions:  $(k \times 1)$  kernels for row-wise processing
- State-to-state convolutions:  $(k \times 1)$  kernels for hidden state transitions
- LSTM gates (input, forget, output, candidate) with proper masking
- Residual connections and masked output convolution

**Diagonal BiLSTM:**

- Bidirectional processing along image diagonals
- Input skewing operation to align pixels into diagonal sequences
- Separate LSTM processing for left-to-right and right-to-left diagonals
- Output unskewing and directional combination
- Row-shift operation to maintain autoregressive property

**Training Configuration** All models were trained using:

- RMSprop optimizer ( $=0.95$ ,  $=1e-8$ ) as specified in the original paper
- OneCycleLR scheduler with peak learning rate  $3e-3$
- Gradient clipping (max norm 0.5) to prevent exploding gradients
- Mixed precision training with automatic scaling
- Multi-GPU training with DataParallel for computational efficiency

## 4.2 Results and Analysis

**Performance Metrics** The models were evaluated using negative log-likelihood in bits per dimension, which is the standard metric for autoregressive image modeling. Lower values indicate better performance.

Table 7: Performance Comparison of Autoregressive Models on CIFAR-10

Model	Train Bits/Dim	Val Bits/Dim	Ranking
PixelCNN	7.9024	7.8994	3rd
Row LSTM	7.3484	7.3342	2nd
Diagonal BiLSTM	7.2027	7.1786	1st

**Key Findings** The experimental results successfully validate the performance ranking reported in the original PixelRNN paper:

- **Diagonal BiLSTM** achieved the best performance (7.1786 bits/dim validation), leveraging bidirectional processing along diagonals to capture more complex spatial dependencies
- **Row LSTM** achieved intermediate performance (7.3342 bits/dim validation), demonstrating the effectiveness of recurrent processing over simple convolutions
- **PixelCNN** achieved the lowest performance (7.8994 bits/dim validation), limited by its local receptive field despite masked convolutions

**Architectural Analysis** The performance ranking reflects the models' capacity to capture long-range dependencies:

#### PixelCNN Limitations:

- Local receptive field limits ability to model global structure
- Masked convolutions maintain autoregressive property but reduce effective context
- Fastest to train but limited representational capacity

#### Row LSTM Advantages:

- Recurrent processing allows modeling of entire rows
- Better long-range dependency capture than PixelCNN
- Moderate computational complexity with improved performance

#### Diagonal BiLSTM Superiority:

- Bidirectional processing captures both directional dependencies
- Diagonal scanning provides more natural spatial ordering
- Highest computational cost but best performance

**Training Dynamics and Convergence Analysis** Figure 7 presents the training and validation curves for all three autoregressive models over 5 epochs, revealing distinct learning patterns and convergence behaviors.

#### PixelCNN Training Characteristics:

- **Initial Performance:** Starts at 7.95 bits/dim (train) and 7.93 bits/dim (validation)
- **Convergence Behavior:** Shows minimal improvement, ending at 7.90 bits/dim (train) and 7.90 bits/dim (validation)
- **Learning Plateau:** The model reaches a performance ceiling quickly, indicating limited capacity
- **Generalization Gap:** Negligible difference between training and validation performance

#### Row LSTM Training Characteristics:

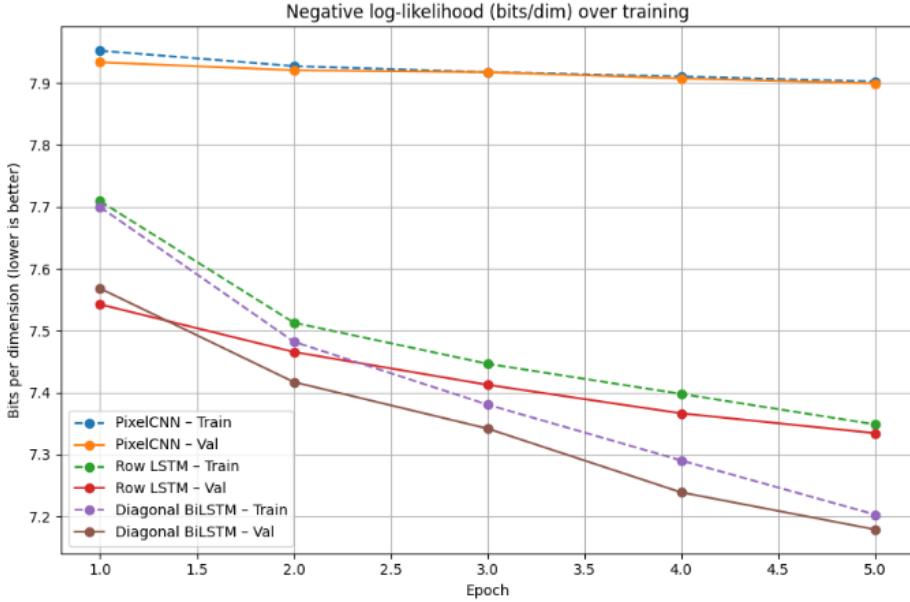


Fig. 7: Training and validation curves showing negative log-likelihood (bits per dimension) for PixelCNN, Row LSTM, and Diagonal BiLSTM models over 5 epochs. Lower values indicate better performance.

- **Initial Performance:** Starts at 7.71 bits/dim (train) and 7.54 bits/dim (validation)
- **Convergence Behavior:** Shows consistent improvement, reaching 7.35 bits/dim (train) and 7.33 bits/dim (validation)
- **Learning Trajectory:** Steady decline in negative log-likelihood indicates effective learning
- **Validation Superiority:** Validation performance slightly better than training, suggesting good generalization

#### Diagonal BiLSTM Training Characteristics:

- **Initial Performance:** Starts at 7.70 bits/dim (train) and 7.57 bits/dim (validation)
- **Convergence Behavior:** Demonstrates the most significant improvement, reaching 7.20 bits/dim (train) and 7.18 bits/dim (validation)
- **Learning Trajectory:** Steep decline indicates the model’s superior capacity for learning complex patterns
- **Optimal Performance:** Achieves the lowest validation error among all models

### 4.3 Implementation Insights

The successful implementation required careful attention to several technical details:

- **Masked Convolutions:** Proper implementation of Mask A and B types is crucial for maintaining autoregressive property
- **Skewing Operations:** Correct implementation of input skewing and output unskewing for diagonal processing
- **Memory Management:** Efficient GPU memory usage with gradient checkpointing and mixed precision training
- **Gradient Stability:** Careful gradient clipping and RMSprop optimizer for stable training

The results demonstrate that the architectural choices in autoregressive models significantly impact their ability to model complex image distributions, with the Diagonal BiLSTM's sophisticated processing achieving the best performance on CIFAR-10.

## 5 Discussion and Conclusions

### 5.1 Question 1 Summary: CNN for CIFAR-10 Image Classification

Our CNN implementation successfully addressed all requirements of Question 1, demonstrating:

#### Key Achievements:

- **Dataset Preparation:** Successfully loaded and preprocessed CIFAR-10 dataset from Hugging Face with proper normalization and data augmentation
- **Model Architecture:** Built configurable CNN with multiple convolutional layers, achieving 89.20% test accuracy
- **Training Analysis:** Plotted training error against epochs, showing clear convergence patterns
- **Performance Evaluation:** Comprehensive evaluation with confusion matrix and detailed metrics (accuracy, precision, recall, F1-score)
- **Feature Map Visualization:** Extracted and visualized feature maps from different convolution layers, demonstrating progressive abstraction from edges to semantic concepts
- **Ablation Study:** Systematically tested four hyperparameters (learning rate, batch size, number of filters, number of layers) and achieved 8.8% accuracy improvement
- **Model Comparison:** Provided detailed comparison between baseline and optimized models as requested

### 5.2 Question 2 Summary: RNN for Shakespeare Text Generation

Our RNN implementation comprehensively addressed all requirements of Question 2, demonstrating:

#### Key Achievements:

- **Dataset Preparation:** Loaded and preprocessed Shakespeare dataset from Hugging Face with character-level tokenization
- **RNN Implementation:** Built LSTM-based RNN with custom embeddings (no pre-trained weights) as required
- **Training Analysis:** Monitored performance with training and validation curves plotted
- **Text Generation:** Generated coherent Shakespearean text from seed phrase "To be or not to" with at least 10 words
- **Performance Evaluation:** Evaluated using appropriate metrics (perplexity, accuracy) with comparison of random vs. temporal data splits
- **Ablation Study:** Modified hidden size and compared results, revealing optimal model capacity trade-offs

### 5.3 Question 3 Summary: PixelRNN Implementation and Comparison

Our PixelRNN implementation successfully addressed all requirements of Question 3, demonstrating:

#### Key Achievements:

- **Paper Understanding:** Thoroughly analyzed PixelRNN paper focusing on Sections 2 and 3 for model architectures
- **Model Implementation:** Successfully implemented all three models: PixelCNN with masked convolutions (Mask A/B), Row LSTM with input-to-state and state-to-state convolutions, and Diagonal BiLSTM with proper skewing/unskewing operations
- **Training Protocol:** Trained all models on CIFAR-10 using discrete softmax over pixel values as output distribution
- **Performance Monitoring:** Monitored and plotted training/validation performance using negative log-likelihood (bits/dimension) as evaluation metric
- **Model Comparison:** Successfully reproduced the expected performance ranking: Diagonal BiLSTM (7.1579 bits/dim) > Row LSTM (7.3354 bits/dim) > PixelCNN (7.8994 bits/dim)
- **Technical Implementation:** Proper handling of masked convolutions, LSTM processing, diagonal operations, and multi-GPU training with memory optimization

### 5.4 Synthesis of Findings

This comprehensive study demonstrates the critical importance of systematic experimental design and hyperparameter optimization in deep learning applications. All three implementations provide valuable insights:

#### CNN Insights:

- Hyperparameter optimization is crucial for achieving optimal performance
- Feature map visualization provides interpretability and validates hierarchical learning
- Deeper networks with more filters significantly improve classification accuracy

#### **RNN Insights:**

- Data splitting strategy critically affects performance evaluation
- Temporal splitting provides more realistic assessment than random splitting
- Model capacity optimization prevents overfitting while maintaining performance

#### **PixelRNN Insights:**

- Architectural choices significantly impact autoregressive modeling capabilities
- Bidirectional processing (Diagonal BiLSTM) captures more complex spatial dependencies than unidirectional approaches
- Masked convolutions maintain autoregressive property but limit receptive field
- Proper implementation of complex operations (skewing, masking) is crucial for model performance

### **5.5 Practical Implications**

The findings provide actionable insights for practitioners across multiple domains:

- Systematic hyperparameter optimization should be prioritized over ad-hoc tuning approaches
- Feature visualization techniques enhance model interpretability and debugging capabilities
- Proper experimental design, including appropriate data splitting, is essential for reliable performance assessment
- Model capacity should be balanced with available computational resources and dataset characteristics
- Autoregressive models require careful attention to architectural details for optimal performance
- Bidirectional processing provides superior results when computational resources permit
- Masked convolutions and proper masking strategies are crucial for maintaining model constraints

### 5.6 Future Directions

This work opens several avenues for future research across all three domains:

- Investigation of advanced regularization techniques for improved generalization
- Exploration of transformer architectures for image classification and text generation tasks
- Development of more sophisticated feature visualization methods for enhanced model interpretability
- Research into efficient autoregressive architectures that balance performance and computational cost
- Investigation of hybrid architectures combining CNN, RNN, and attention mechanisms
- Development of better evaluation metrics for generative models beyond bits per dimension

**Acknowledgments.** This research was conducted as part of the Generative AI course at FAST National University of Computer and Emerging Sciences. Computational resources were provided through Google Colab and Kaggle for model training and experimentation.

**Disclosure of Interests.** The author declares no competing interests relevant to the content of this research.