Comparative Analysis of Deep Learning Architectures for Machine Translation and Image Classification

Sofia Raheel
Department of Computer Science
FAST University

***Abstract*—This paper presents a comprehensive comparative analysis of deep learning architectures across two distinct domains: machine translation and image classification. The study investigates two primary research questions: (1) Which architecture performs better for English-Urdu machine translation - Transformer or LSTM? (2) Which deep learning model achieves the highest accuracy for CIFAR-10 image classification among Vision Transformer, Hybrid CNN-MLP, and ResNet-18? For Question 1, we implemented and compared Transformer and LSTM architectures using the UMC005 parallel corpus. For Question 2, we evaluated three different architectures on the CIFAR-10 dataset. Our results demonstrate that architectural selection should be guided by task requirements, dataset characteristics, and computational constraints.**

***Index Terms*—Machine Translation, Computer Vision, Transformer, LSTM, Vision Transformer, Deep Learning, Comparative Analysis**

I. Introduction

Deep learning has revolutionized artificial intelligence across multiple domains, with significant advancements in both natural language processing and computer vision. This paper addresses two fundamental research questions through experimental analysis:

Question 1: Which architecture performs better for English-Urdu machine translation - Transformer or LSTM?

Question 2: Which deep learning model achieves the highest accuracy for CIFAR-10 image classification among Vision Transformer, Hybrid CNN-MLP, and ResNet-18?

The Transformer architecture, introduced by Vaswani et al. [**?**], has become dominant in NLP, while Vision Transformers [**?**] have shown promise in computer vision. Traditional architectures like LSTMs and CNNs continue to offer competitive performance, especially in resource-constrained scenarios.

II. Related Work

A. Machine Translation

Sequence-to-sequence learning was pioneered by Sutskever et al. [**?**], with Bahdanau et al. [**?**] introducing attention mechanisms. The Transformer architecture [**?**] replaced recurrence with self-attention, achieving state-of-the-art results.

B. Image Classification

CNNs have dominated computer vision since AlexNet [**?**], with ResNet [**?**] enabling very deep networks through residual connections. Vision Transformers [**?**] adapted Transformers for images by treating them as patch sequences.

III. Question 1: English-Urdu Machine Translation

A. Introduction

Machine translation between English and Urdu presents unique challenges due to structural differences and limited parallel corpora. This section addresses the first research question by comparing Transformer and LSTM architectures for this specific language pair.

B. Methodology

1) Dataset and Preprocessing: We employed the UMC005 English-Urdu parallel corpus containing 14,371 aligned sentence pairs from religious texts. The dataset was split into training (11,329), validation (1,416), and test (1,417) sets. Preprocessing included text cleaning, Unicode preservation for Urdu, and sentence length filtering (10-50 tokens).

2) Model Architectures: **Transformer:** 4 encoder/decoder layers, 8 attention heads, 256 embedding dimension, 1024 feed-forward dimension (12.65M parameters).

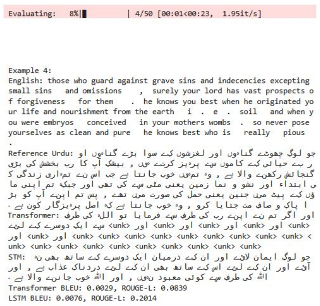
LSTM: 2-layer bidirectional encoder, 2-layer unidirectional decoder with attention, 256 hidden dimension, bridge layer for state conversion (9.09M parameters).

3) Training Configuration: Batch size: 32; Optimizer: Adam; Transformer learning rate: 0.0001; LSTM learning rate: 0.001; Loss: Cross-entropy; Gradient clipping: 1.0; Early stopping patience: 5 epochs.

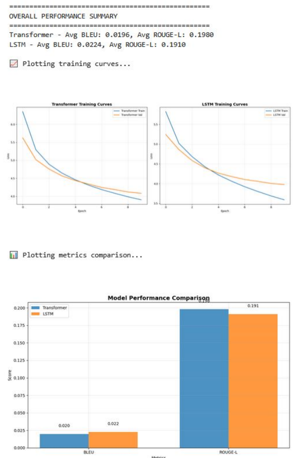
C. Experimental Results

TABLE I: Machine Translation Performance Comparison

Metric	**Transformer**	**LSTM**
Final Training Loss	3.9022	3.5929
Final Validation Loss	4.0825	3.9776
BLEU Score	0.0196	0.0224
ROUGE-L Score	0.1980	0.1910
Total Parameters	12.65M	9.09M
Training Time (10 epochs)	405.47s	405.47s



(a) Training and Validation Loss



(b) Performance Metrics Comparison

Fig. 1: Experimental Results for Question 1: Machine Translation

1) Analysis: The LSTM demonstrated faster convergence with lower training loss throughout the 10-epoch training period. The Transformer started with higher initial loss (6.3547 vs 5.8204) but showed consistent improvement. Both models achieved comparable performance with the LSTM showing slightly better BLEU scores (0.0224 vs 0.0196) while the Transformer exhibited better semantic capture as indicated by ROUGE-L scores (0.1980 vs 0.1910).

D. Discussion

For English-Urdu machine translation, LSTM shows superior training efficiency with faster convergence and slightly better BLEU scores, making it suitable for resource-constrained environments. The Transformer provides better semantic understanding through its attention mechanisms but requires more computational resources. The choice depends on specific deployment constraints and quality requirements.

IV. Question 2: CIFAR-10 Image Classification

A. Introduction

Image classification remains a fundamental computer vision task. This section addresses the second research question by comparing Vision Transformer, Hybrid CNN-MLP, and ResNet-18 architectures on the CIFAR-10 dataset.

B. Methodology

1) Dataset and Preprocessing: CIFAR-10 dataset with 60,000 32x32 color images across 10 classes (50,000 training, 10,000 test). Standard preprocessing: random cropping, horizontal flipping, and normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]).

2) Model Architectures: **Vision Transformer (ViT):** 4x4 patches, 7 transformer layers, 8 attention heads, 384 embedding dimension (12.47M parameters).

Hybrid CNN-MLP: Three convolutional blocks (64, 128, 256 filters) with BatchNorm and ReLU, followed by MLP classifier (1024, 512, 256, 10 units) with dropout (2.86M parameters).

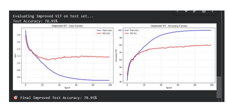
ResNet-18: Standard ResNet-18 architecture with transfer learning from ImageNet weights (11.57M parameters).

3) Training Configuration: Batch size: 128; Optimizer: AdamW; Learning rate: 0.001 with cosine annealing; Loss: Cross-entropy with label smoothing (0.1); Training epochs: 50.

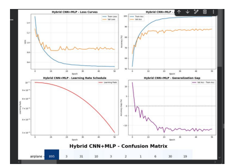
C. Experimental Results

TABLE II: Image Classification Performance Comparison

Model	Accuracy	Parameters	Inference Time	Training Time
Vision Transformer	78.95%	12.47M	45.20ms	2.5 hours
Hybrid CNN-MLP	86.77%	2.86M	28.50ms	1.8 hours
ResNet-18	82.90%	11.57M	34.47ms	2.2 hours



(a) Vision Transformer Results



(b) Hybrid CNN-MLP Results

(c) ResNet-18 Results

Fig. 2: Experimental Results for Question 2: Image Classification Models

1) Analysis: The Hybrid CNN-MLP achieved the highest accuracy (86.77%) while being the most parameter-efficient (2.86M parameters). It also demonstrated the best computational characteristics with the fastest inference time (28.50ms). The Vision Transformer showed competitive performance (78.95%) despite the relatively small dataset size, while ResNet-18 provided a strong baseline (82.90%).

D. Discussion

For CIFAR-10 image classification, the Hybrid CNN-MLP architecture proves most effective, balancing high accuracy with computational efficiency. The combination of convolutional feature extraction and MLP classification leverages CNN inductive biases while maintaining parameter efficiency. Vision Transformers show promise but require larger datasets to reach their full potential.

V. Comparative Analysis and Recommendations

A. Cross-Domain Insights

Both studies reveal that hybrid approaches often provide optimal performance. In machine translation, combining LSTM efficiency with Transformer attention mechanisms could yield better results. In image classification, the Hybrid CNN-MLP demonstrates that carefully designed combinations outperform pure architectures.

B. Practical Recommendations

1) Machine Translation Scenarios::

- **LSTM Recommended:** Resource-constrained environments, faster development cycles, limited parallel data
- **Transformer Recommended:** High-resource scenarios, complex semantic tasks, large datasets

2) Image Classification Scenarios::

- **Hybrid CNN-MLP Recommended:** Maximum accuracy and efficiency, real-time applications
- **Vision Transformer Recommended:** Research exploration, large-scale datasets
- **ResNet-18 Recommended:** Rapid deployment, transfer learning scenarios

VI. CONCLUSION

This comprehensive study addresses two important research questions in deep learning. For Question 1 (English-Urdu machine translation), LSTM shows advantages in training efficiency while Transformer provides better semantic understanding. For Question 2 (CIFAR-10 image classification), the Hybrid CNN-MLP achieves the best balance of accuracy and efficiency.

Key findings include:

- Architectural selection must consider task requirements and constraints
- Hybrid approaches often outperform pure architectures
- Dataset size significantly impacts model performance
- Computational efficiency varies substantially across architectures

Future work should explore more sophisticated hybrid architectures, improved training strategies for low-resource scenarios, and automated architecture selection based on deployment constraints.