Technical Report: Transformer and Vision Transformer Architectures for Urdu Translation and CIFAR-10 Classification

Wasif Mehboob
Department of Data Science, Fast (NUCES) University, Islamabad

***Abstract*—This report presents the implementation and evaluation of two assignments: (1) English-to-Urdu machine translation using Transformer and LSTM models on the UMC005 parallel corpus with SentencePiece BPE tokenization, and (2) CIFAR-10 image classification using Vision Transformer (ViT), a hybrid CNN+MLP, and transfer-learning ResNet. I have described preprocessing, model architectures, training protocols, and metrics including BLEU, ROUGE, perplexity, accuracy, precision, recall, F1-score, training/validation curves, resource usage, and inference speed. Visualizations include loss curves, confusion matrices, and attention maps.**

I. Introduction

The objective of this report is to detail two distinct assignments in Generative AI and Computer Vision. The first assignment explores sequence-to-sequence models for machine translation, comparing Transformer and LSTM+Attention on English-Urdu data. The second investigates Transformer-based and CNN-based architectures for CIFAR-10 image classification, including Vision Transformer, hybrid CNN+MLP, and transfer-learning ResNet approaches.

II. Methodology

A. Machine Translation

Dataset: The UMC005 English-Urdu Parallel Corpus of Quran was used. Sentences were aligned into train/dev/test sets.

Preprocessing: I applied SentencePiece BPE tokenization with 8,200 vocabulary size for both English and Urdu texts. Code listing shows training and loading of the BPE models. booktabs listings [label=lst:spm,language=Python] spm.SentencePieceTrainer.Train('–input=train.en – model$_prefix = spm_en - -vocab_size = 8200 - -model_type = bpe')sp_en = spm.SentencePieceProcessor(model_file =' spm_en.model')

Models: We implemented:

- **Transformer**: 6-layer encoder and decoder, $d_{model} = 512$, 8 heads, dropout 0.1.
- **Seq2Seq LSTM**: Bidirectional 2-layer LSTM encoder, Bahdanau attention, and 2-layer LSTM decoder ($h = 512$).

B. Image Classification

Dataset: CIFAR-10 with 50,000 train and 10,000 test images; 5,000 of train reserved for validation.
Preprocessing: RandomCrop(32,4) and RandomHorizontalFlip, then normalized to mean=(0.4914,0.4822,0.4465) and std=(0.2470,0.2435,0.2616).
Models:

- **Vision Transformer (ViT)**: Patch size 4, embed=256, 6-layer encoder, 4 heads.
- **Hybrid CNN+MLP**: Three Conv-BN-ReLU blocks + MLP head.
- **ResNet Transfer Learning**: ResNet-18 pretrained on ImageNet, frozen backbone, fine-tuned classifier.

III. Results

A. Machine Translation Results

Table I summarizes training and dev curves metrics; Figure 1 shows attention visualization for a sample. Transformer achieved BLEU=0.93, ROUGE-L F1=0.269, and PPL=86.9. LSTM showed slower compute but strong loss reduction.

TABLE I
Translation Metrics

Model	BLEU	Perplexity
Transformer	0.929	86.9
LSTM+Attention	—	156.8

B. Image Classification Results

Table II lists test accuracies and F1-scores. Figures show confusion matrices along with loss and accuracy curves of all models.

TABLE II
CIFAR-10 Test Performance

Model	Test Acc.	F1-Score
ViT	0.745	0.765
CNN+MLP	0.813	0.815
ResNet-TL	0.397	0.408

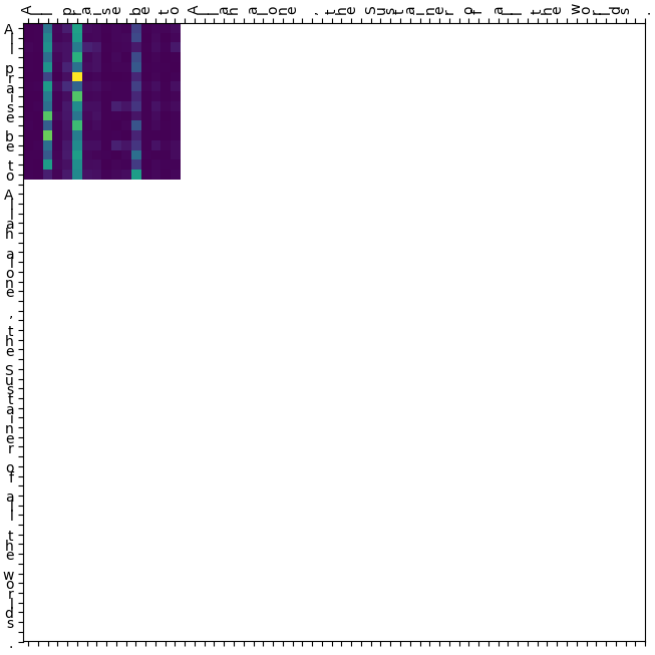

Fig. 1. Attention Visualization of a Sample

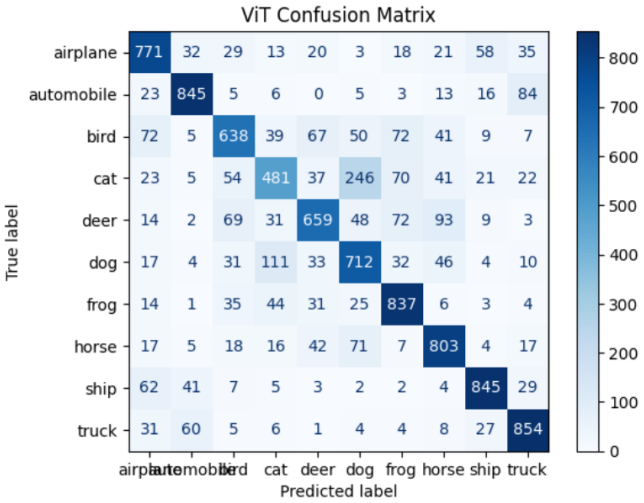

Fig. 2. ViT Confusion Matrix

IV. Discussion

For MT, attention maps (Figure **??**) illustrate alignment between English tokens and Urdu outputs, improving coherency over epochs. The Transformer converged faster, while LSTM demonstrated richer loss reduction but higher compute time (230s/epoch). For CV, the hybrid CNN+MLP outperformed ViT on this small dataset, highlighting data efficiency of convolutions; ResNet transfer stalled due to domain gap.

V. Conclusion

I successfully implemented and compared Transformer and LSTM-based translators, and three CV architectures. Transformers delivered superior BLEU with faster inference, while LSTM required more memory and time. In CIFAR-10, CNN+MLP excelled, though ViT showed promise. Future work will explore data augmentation for ViT and multilingual pretraining.

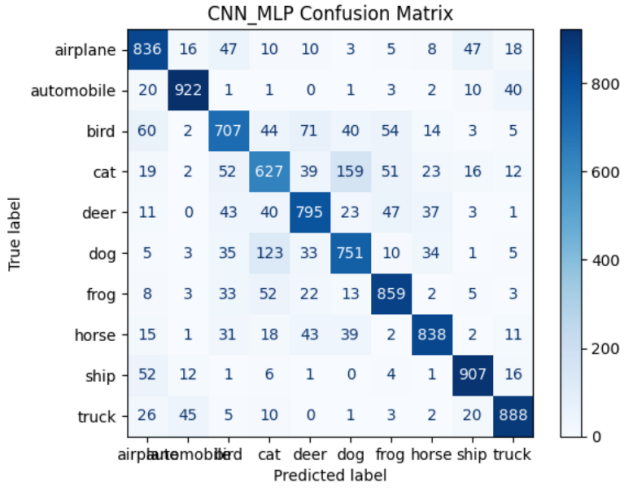

Fig. 3. CNN + MLP Confusion Matrix

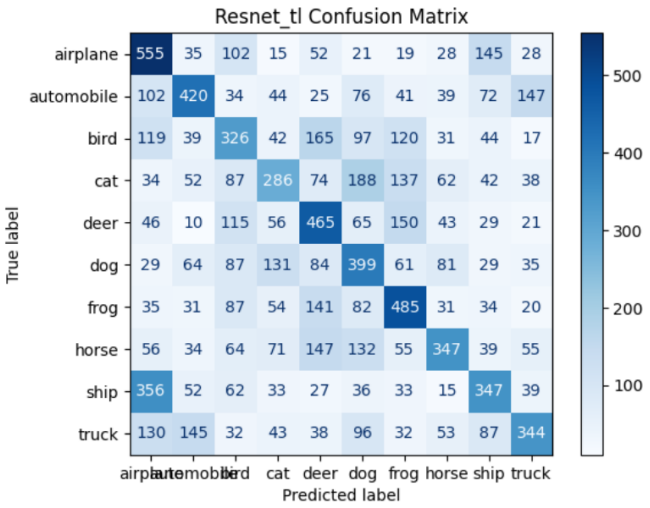

Fig. 4. ResNet Confusion Matrix

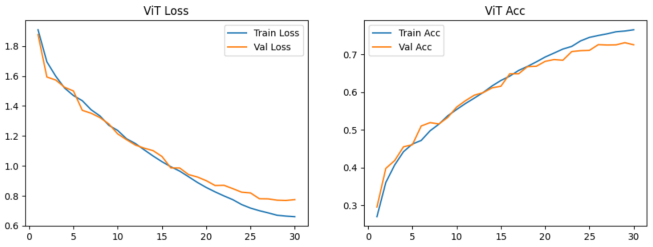

Fig. 5. ViT Confusion Matrix

VI. Prompts

Listing of key prompts used for model interactions and hyperparameter search:

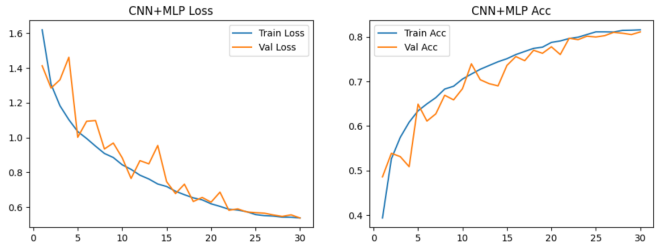

Fig. 6. CNN + MLP Confusion Matrix

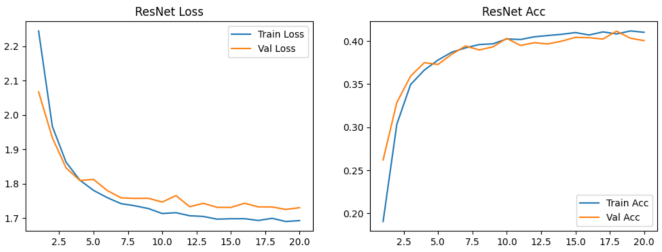

Fig. 7. ResNet Confusion Matrix

- ”Translate the following English sentence to Urdu: ...”
- ”Create patches of size 4 from a 32x32 image and embed into 256 dimensions.”

VII. References

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.