MediAssist

A Generative AI Model for Diagnostic Report Generation

Group Members :

BOYA ANUDEEP (230304)
A S PRAJIT (230001)
ASWAAJIEET J M (230236)
SHREEHARI P (230985)

Abstract :

Automated generation of radiology reports from chest X-ray images can greatly reduce radiologists' workload and improve patient care. We propose *MediAssist*, a multimodal deep learning model that generates diagnostic reports from chest X-rays. MediAssist is inspired by TieNet, a 2018 model that jointly learned image and text features for thorax disease classification and reporting, but improves on it by replacing the original RNN-based decoder with a Transformer-based decoder.

Our approach combines a convolutional image encoder with a multi-head attention transformer for language modeling. MediAssist's design leverages prior work in image captioning (e.g. "Show, Attend and Tell" with visual attention) and multi-modal medical imaging models, and demonstrates the benefit of Transformer-based language decoders for structured medical report generation.

Introduction :

Chest X-rays (CXRs) are one of the most common radiological exams, and accurate interpretation is critical for patient diagnosis. However, the global shortage of radiologists and high imaging workload lead to delays and errors. Automating the report-writing process for CXRs can augment radiologist efficiency. Recent advances in deep learning, especially in vision-language models, make it possible to generate free-text descriptions of images. Classical image captioning models like "Show and Tell" and "Show, Attend and Tell" demonstrated that attention-based encoder-decoder networks can describe image content. In medical imaging, TieNet (Wang *et al.*, 2018) introduced a CNN-RNN framework with multi-level attention that simultaneously classified diseases and generated preliminary reports. These works show that combining convolutional encoders with attention-driven decoders can align visual features to words.

However, traditional RNN/LSTM decoders have limitations in capturing long-range dependencies in text. Recent work has shown that Transformer architectures, with self-attention mechanisms, improve language modeling for image captioning. Inspired by these advances, MediAssist replaces TieNet's RNN-based language model with a Transformer-based decoder. This allows the model to better capture complex language structures and long-range context in radiology reports.

TieNet's original multi-level attention and joint image-text embedding idea serves as a foundation. We credit TieNet for pioneering automated CXR report generation and using it as a base architecture. Our key modification is to use a Transformer decoder with multi-head self-attention, while still using a ResNet CNN to encode images.

Dataset :

We use the **MIMIC-CXR** dataset, a publicly available collection of chest X-rays paired with free-text radiology reports physionet.org. MIMIC-CXR contains over 377,000 frontal/posterior CXR images from more than 227,000 studies. From MIMIC-CXR we sampled 10,000 image–report pairs for our experiments: 7,000 for training, 1,000 for validation, and 2,000 for testing (a representative subset of the full database). Each sample consists of a single CXR image and its associated radiology report.

Preprocessing: We apply standard image preprocessing (resize 224×224, tensor conversion, normalize (as ResNet-50 expects)) and convert each report to a sequence of BERT token IDs and truncated/padded to a fixed length of 128 tokens. Special tokens [CLS] and [SEP] are added to each report. We reserve one index as the <PAD> token for padding shorter reports. During training, only valid (non-padded) tokens are considered in the loss.

Model Architecture :

MediAssist follows a **CNN–Transformer** encoder-decoder design. The overall pipeline for a training sample is as follows:

1. **Image Encoding:** The input CXR image is fed into a ResNet-50 convolutional neural network pretrained on ImageNet. We remove ResNet's final classification layer to output a feature map of shape (C, H, W) (typically C=2048, H=W=7 for a 224×224 input). This feature map encodes spatial visual features of the radiograph.

2. **Spatial Attention (Feature Refinement):** We apply a learned spatial attention mechanism to the feature map. Concretely, a 1×1 convolution produces an attention map over the H×W spatial grid. The feature map is multiplied by this attention map, highlighting relevant regions (e.g., areas of abnormal opacity) while suppressing irrelevant background. This yields an "attended" feature map of the same dimension. We then flatten this map into a sequence of M=H×W feature vectors of dimension C.

3. **Text Embedding:** In parallel, the target report text (during training) is tokenized into IDs and embedded into a sequence of D-dimensional vectors. We use a learned embedding layer (dimension D=512) followed by positional encodings to represent each token in context. The maximum sequence length is 128 tokens (including special tokens).

4. **Transformer Decoder:** The core of MediAssist is a multi-layer Transformer decoder. At each decoding step, the model takes the embedded token sequence up to the current position and attends to both itself (masked self-attention) and the image features (cross-attention). We use a Transformer decoder with 2 layers, 4 attention heads, hidden dimension 512, and dropout 0.1 (hyperparameters). In each layer, the

decoder first applies masked multi-head self-attention over the target tokens, then multi-head cross-attention over the flattened image features, and finally a feed-forward network. The output of the decoder layers is a sequence of contextualized vectors (one per token position).

5. **Output Projection:** The decoder's output vectors are projected through a linear layer to the size of the vocabulary, producing logits for each token. During training, these logits are compared to the ground-truth next token via cross-entropy loss (with teacher forcing). At inference, we generate tokens autoregressively using a beam search (beam width 5) until an end-of-sequence token is produced or max length is reached.

By replacing TieNet's LSTM decoder with a Transformer, our model can more effectively capture long-range dependencies in report text. The Transformer's multi-head attention allows the decoder to focus flexibly on different image regions and previously generated words. This architecture draws on the success of attention-based captioning and extends TieNet's CNN+attention idea with modern Transformer decoders.

Training :

We train MediAssist end-to-end on the training split of MIMIC-CXR.

- **Loss Function:** We use cross-entropy loss between the decoder's output logits and the ground-truth report tokens (teacher forcing). Padding tokens do not contribute to the loss.

- **Optimization:** Parameters are optimized with the Adam optimizer (learning rate 1e-4). We employ gradient clipping (value 1.0) to stabilize training.

- **Batching:** Batch size is set to 8 images per step. We shuffle the training data each epoch.

- **Training Schedule:** We train for 100 epochs. To preserve learned image features, the ResNet encoder is frozen (no weight updates) for the first few epochs and then gradually fine-tuned (as in TieNet). This helps the model adapt to radiology images without catastrophic forgetting.

- **Regularization:** Dropout (0.1) is used in the Transformer layers to prevent overfitting. We also use data augmentation on images (random flips) during training.

- **Evaluation:** After each epoch, we measure validation loss on the 1,000-image validation set. The best model (lowest validation loss) is saved. For final testing, we generate reports on the 2,000-image test set using beam search (beam width = 5).

Results :

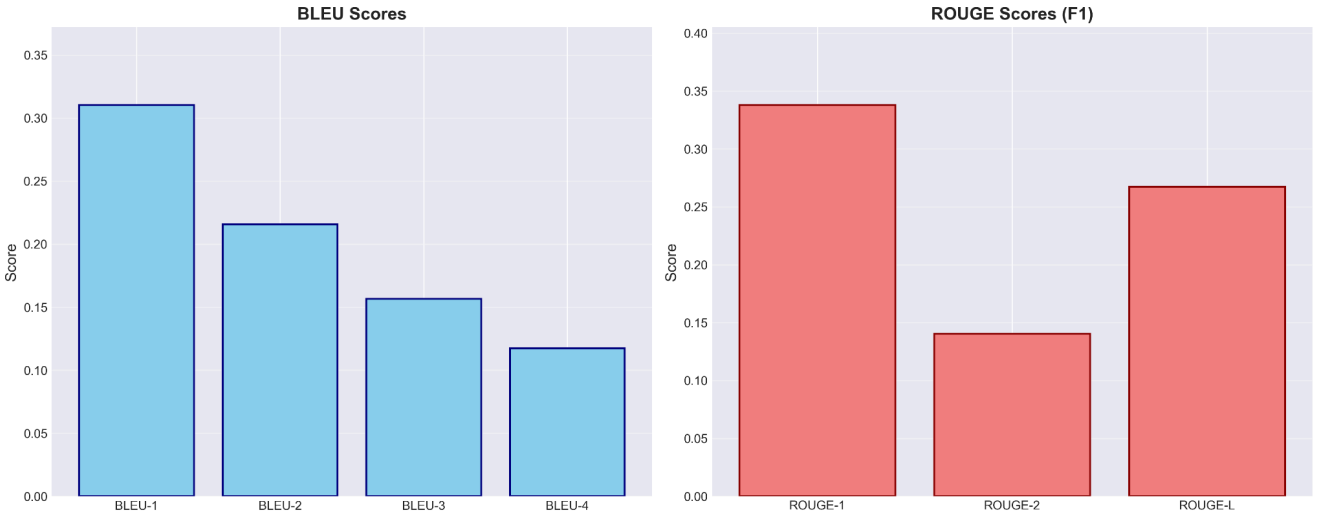
We evaluate MediAssist using standard language generation metrics: BLEU-1 through BLEU-4 and ROUGE-1, ROUGE-2, ROUGE-L. These metrics measure n-gram overlap between the generated report and the reference report.

Our model achieved the following scores on the test set :

- BLEU-1: 0.31015
- BLEU-2: 0.21564
- BLEU-3: 0.15667
- BLEU-4: 0.11740

- ROUGE-1: 0.33773
- ROUGE-2: 0.14042
- ROUGE-L: 0.26728

Compared to a purely CNN-LSTM baseline , using a Transformer decoder tends to yield better BLEU and ROUGE scores .



Training , Validation loss vs Epoch: