Model Architecture

We implemented the **TURBO (Target-aUgmented shaRed fusion-Based sarcasm explanatiOn)** model, integrating a **Vision Transformer (ViT)** for image encoding and **BART** for text-based generation. The architecture is composed of the following components:

Visual Encoder

A ViT-base model is employed to extract high-level visual features from input images.

Text Encoder

The encoder of the BART model processes the textual inputs, including:

- Social media captions
- Object detection outputs
- Image descriptions
- Explicit sarcasm targets

Cross-Modal Fusion

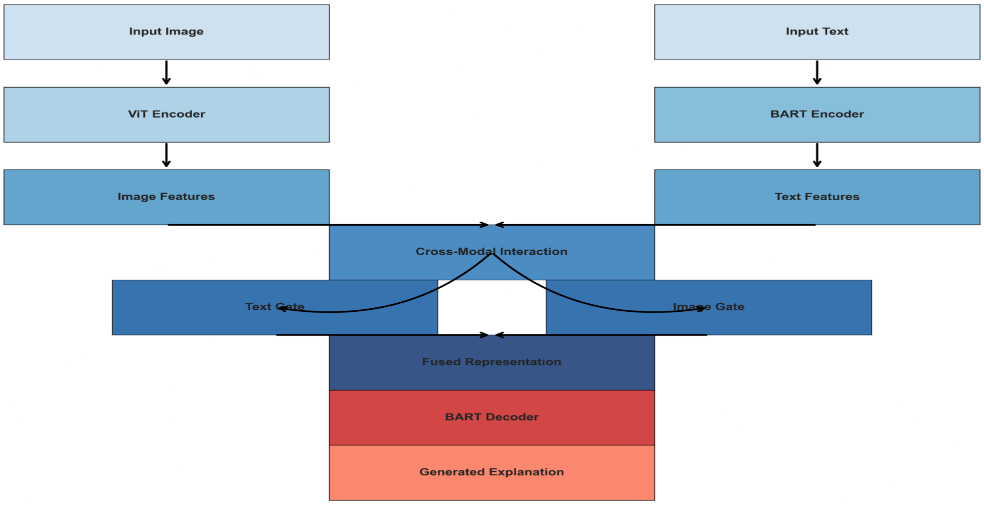
A shared fusion mechanism integrates visual and textual modalities using:

- Global representation extraction
- Cross-modal interaction layers
- Gated fusion with trainable parameters

Decoder

The BART decoder generates the final explanation, attending to the fused multimodal representation.

TURBO Model Architecture



This modular architecture preserves the flexibility and learnability of the fusion layers, facilitating nuanced sarcasm interpretation.

Data Processing

The preprocessing pipeline accommodates the model's multimodal nature:

- **Visual Input**: Images are processed via the ViT feature extractor.
- **Textual Input**: Captions, sarcasm targets, object labels, and image descriptions are tokenized and concatenated with special separator tokens.
- **Decoder Targets**: Output sequences are formatted with start-of-sequence and end-of-sequence tokens for autoregressive generation.

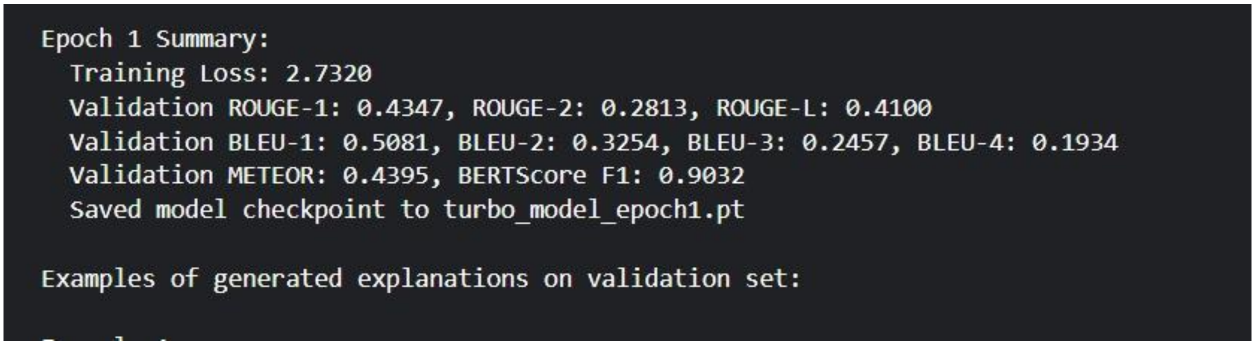
Hyperparameters Used

1. **Model Architecture Hyperparameters**
    - Visual Encoder : ViT-base (google/vit-base-patch16-224-in21k)
    - Text Encoder : BART-base (facebook/bart-base)
    - Hidden Dimension : 768 (derived from BART's configuration)
    - Cross-Modal Fusion :
        - - Gating layers with 2*d → d dimensions (where d=768)
        - - Sigmoid activation for gating mechanism
2. **Training Hyperparameters**
    - Optimizer : AdamW
    - Learning Rate : 1e-4
    - Weight Decay : 0.01 (for L2 regularization)
    - Batch Size : 8
    - Number of Epochs : 3
    - Loss Function : Cross-entropy with padding tokens ignored
3. **Input/Output Processing Hyperparameters**
    - Max Input Length : 64 tokens
    - Max Output Length : 32 tokens
    - Tokenizer : BART tokenizer with special tokens (BOS, EOS, PAD)
    - Image Processor: ViT feature extractor (224×224 pixels)
4. **Inference Hyperparameters**
    - Decoding Strategy : Greedy decoding (argmax)
    - Generation Stopping Criteria : EOS token or maximum length reached
    - Temperature : 1.0 (no temperature scaling applied)
5. **Model Regularization**
    - Parameter Freezing : ViT parameters frozen, only BART and fusion components trained
    - Gradient Accumulation : 1 (no gradient accumulation)

Model checkpoints were saved at the end of each epoch. Cross-entropy loss was used for optimization, with padding tokens excluded from loss computation.

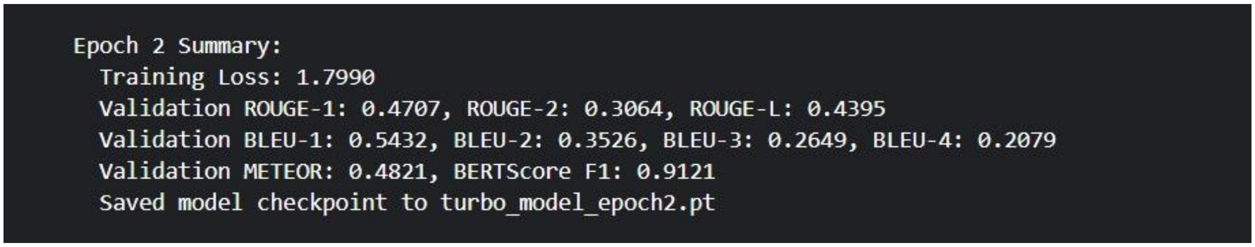
Results and Analysis

The model demonstrated progressive performance improvements across epochs. Below are the results on the validation set. The model exhibited steady performance improvements across training epochs. The table below summarizes training loss and evaluation metrics on the validation set:

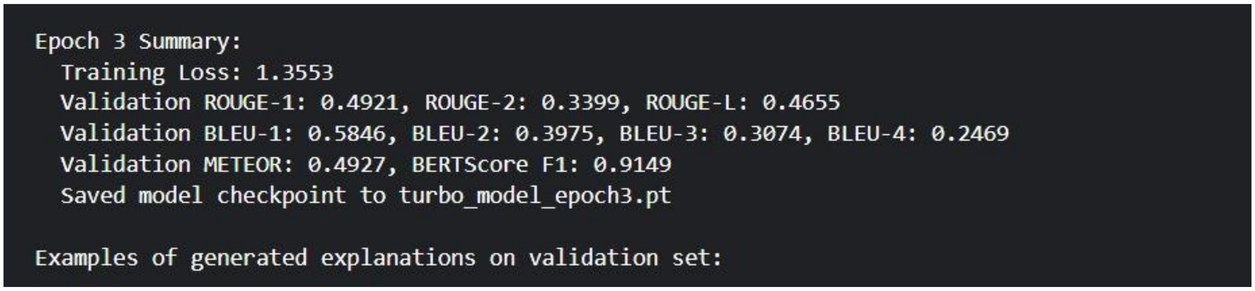


```
Epoch 1 Summary:
  Training Loss: 2.7320
  Validation ROUGE-1: 0.4347, ROUGE-2: 0.2813, ROUGE-L: 0.4100
  Validation BLEU-1: 0.5081, BLEU-2: 0.3254, BLEU-3: 0.2457, BLEU-4: 0.1934
  Validation METEOR: 0.4395, BERTScore F1: 0.9032
  Saved model checkpoint to turbo_model_epoch1.pt

Examples of generated explanations on validation set:
```

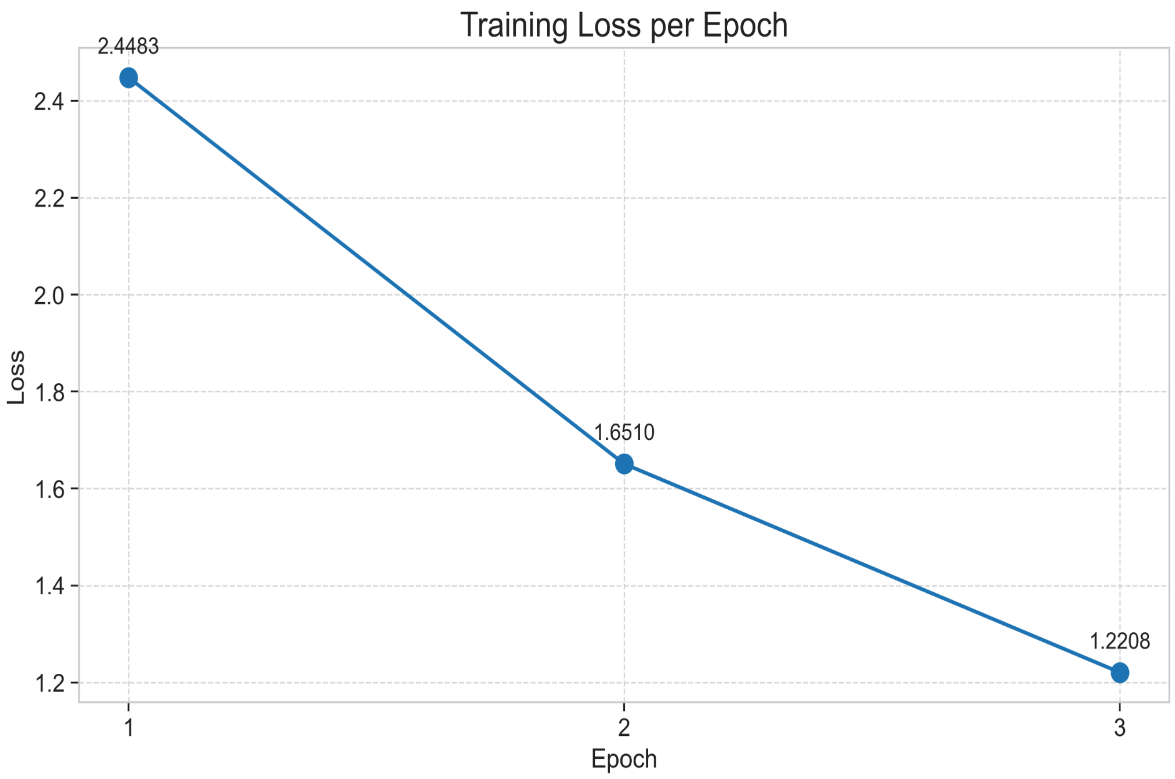


```
Epoch 2 Summary:
  Training Loss: 1.7990
  Validation ROUGE-1: 0.4707, ROUGE-2: 0.3064, ROUGE-L: 0.4395
  Validation BLEU-1: 0.5432, BLEU-2: 0.3526, BLEU-3: 0.2649, BLEU-4: 0.2079
  Validation METEOR: 0.4821, BERTScore F1: 0.9121
  Saved model checkpoint to turbo_model_epoch2.pt
```



```
Epoch 3 Summary:
  Training Loss: 1.3553
  Validation ROUGE-1: 0.4921, ROUGE-2: 0.3399, ROUGE-L: 0.4655
  Validation BLEU-1: 0.5846, BLEU-2: 0.3975, BLEU-3: 0.3074, BLEU-4: 0.2469
  Validation METEOR: 0.4927, BERTScore F1: 0.9149
  Saved model checkpoint to turbo_model_epoch3.pt

Examples of generated explanations on validation set:
```



The **consistent decline in training loss and improvement in evaluation metrics across epochs** highlight the model's effective learning and generalization capabilities for the sarcasm explanation task. The metrics indicate effective learning and strong generalization in sarcasm explanation generation.

Qualitative Analysis

The model generates fluent, semantically accurate explanations that reflect the underlying sarcastic intent. Examples include:

Example 1

- *Caption*: “@ thank u for this awesome network in malad (see pic) . #patheticcs”
- *Target*: “’s network in malad”
- *Generated Explanation*:
  “The author is frustrated with for the poor network in Malad, despite the sarcastic use of ‘awesome’.”

Example 2

- *Caption*: “Nothing like waiting for an hour on the tarmac for a gate to come open in snowy, windy Chicago!”
- *Target*: “gate not opening”

- *Generated Explanation*:
  "The author expresses frustration about the delay in gate opening, masked by a sarcastic remark."

These outputs demonstrate the model's ability to decode sarcasm and provide context-aware explanations.

Conclusion

Our implementation of the TURBO model successfully addresses the MuSE task, leveraging cross-modal fusion to produce coherent and meaningful sarcasm explanations. The model achieves competitive performance across standard metrics and demonstrates consistent improvements throughout training.