Detailed Report: Multimodal Sarcasm Explanation Generation Using a TURBO-like Model

1. Preprocessing Steps

- **Data Sources:**
  - The dataset is provided in TSV files containing columns:
    - `pid` – Unique identifier for each sample.
    - `text` – The original text or caption accompanying the image.
    - `explanation` – The gold (human-annotated) sarcasm explanation.
    - `target_of_sarcasm` – The specific target or entity that is being ridiculed by the sarcasm.
- **Additional Modalities:**
  - **BLIP Descriptions:** A pickle file (`D_train.pkl` / `D_val.pkl`) contains automatically generated image descriptions.
  - **YOLO Object Labels:** Another pickle file (`O_train.pkl` / `O_val.pkl`) contains lists of objects detected in the images.
- **Text Concatenation:**
  - For each sample, the final input text is formed by concatenating:
    - The original text (caption).
    - The BLIP description.
    - The YOLO object labels (converted into a space-separated string).
    - A special separator token (`</s>`) and then the target of sarcasm.
  - This concatenated text is tokenized using the BART tokenizer with a maximum token length of 128 tokens (with truncation and padding applied).
- **Image Processing:**
  - Images are loaded from the specified directory using the sample's `pid` (matching either a `.jpg` or `.png` file).
  - Each image is resized to 224×224 pixels and normalized using ImageNet's mean and standard deviation.

2. Model Architecture and Hyperparameters

- **Architecture Overview:**
  - **Text Generation Backbone:**
    - Uses **BART (base)** as the sequence-to-sequence model for generating sarcasm explanations.
  - **Image Feature Extraction:**
    - Uses **ViT (Vision Transformer)** with a base patch size of 16 (224×224) to extract high-level visual features.
  - **Shared Fusion Module:**
    - A custom module that applies self-attention separately to text embeddings and image embeddings.
    - A gating mechanism combines the two modalities by computing learnable weights and fusing the attended features.
  - **Fusion Output:**
    - The fused representation is fed to the BART decoder for generating the final explanation.

- **Hyperparameters:**
  - **Batch Size:** 4
  - **Learning Rate:** 1e-4
  - **Number of Epochs:** 10
  - **Maximum Token Length (for input):** 128 tokens
  - **Maximum Explanation Length:** 64 tokens (i.e., half of the max input length)
  - **Multihead Attention:** 4 heads are used in both text and image attention modules
  - **Checkpointing:** Only the best-performing model (based on validation loss) is saved as `best_checkpoint.pt`.

3. Training Loss Report

During training, the following losses were observed (example values):

- **Epoch 1:**
  - Train Loss: 3.9897
  - Validation Loss: 3.4910
  - *Best model saved at epoch 1 (lowest validation loss so far)*
- **Epoch 2:**
  - Train Loss: 3.0947
  - Validation Loss: 3.5719
- **Epoch 3:**
  - Train Loss: 2.6224
  - Validation Loss: 3.7753
- **Epoch 4:**
  - Train Loss: 2.1162
  - Validation Loss: 3.9799
- **Epoch 5:**
  - Train Loss: 1.6882
  - Validation Loss: 4.2517
- **Epoch 6:**
  - Train Loss: 1.3525
  - Validation Loss: 4.5408
- **Epoch 7:**
  - Train Loss: 1.1259
  - Validation Loss: 4.7143
- **Epoch 8:**
  - Train Loss: 0.9746
  - Validation Loss: 4.8841
- **Epoch 9:**
  - Train Loss: 0.8863
  - Validation Loss: 5.0165
- **Epoch 10:**
  - Train Loss: 0.7935
  - Validation Loss: 5.1695

Observation:
The validation loss is lowest at epoch 1, which is why that checkpoint was saved as the best model.

4. Evaluation Metrics on the Validation Set (After Each Epoch)

Example metrics from Epoch 1 and later epochs (values may vary):

- **Epoch 1 Evaluation:**
    - ROUGE-1: 0.1918 (≈ 19.18%)
    - ROUGE-2: 0.0732 (≈ 7.32%)
    - ROUGE-L: 0.1549 (≈ 15.49%)
    - BLEU-1: 0.2144 (≈ 21.44%)
    - BLEU-2: 0.1158 (≈ 11.58%)
    - BLEU-3: 0.0745 (≈ 7.45%)
    - BLEU-4: 0.0515 (≈ 5.15%)
    - METEOR: 0.1345 (≈ 13.45%)
    - BERT Precision: 0.4669 (≈ 46.69%)
    - BERT Recall: 0.4986 (≈ 49.86%)
    - BERT F1: 0.4805 (≈ 48.05%)
- **Other Epochs:**
    - Evaluation metrics for subsequent epochs are computed similarly and printed out after each epoch.

5. Generated Sarcasm Explanations (Sample Predictions on Validation Set for the Last Epoch)

At the end of training, the following sample predictions were generated from the validation set using the best model checkpoint:

- **Sample 1:**
    - **Reference:** "the author is pissed at <user> for not getting network in malad."
    - **Prediction:** "the author is pissed at <user> for not sending an advanced notice for construction."
- **Sample 2:**
    - **Reference:** "nothing worst than waiting for an hour on the tarmac for a gate to come open in snowy, windy chicago."
    - **Prediction:** "the author is sad to have to mop up muddy doggy prints."
- **Sample 3:**
    - **Reference:** "nobody likes getting one hour of their life sucked away."
    - **Prediction:** "it's a very hot weather today."
- **Sample 4:**
    - **Reference:** "having a salivary gland biopsy on monday morning is not a good way to start the new week."
    - **Prediction:** "the author doesn't think midterm is the best assessment ever created, there are other ways to determine knowledge application."
- **Sample 5:**

  - **Reference:** "the author is worried that the weekend is going to be freezing with a high of -1 and windchill probably -30."
  - **Prediction:** "it's not a beautiful day driving back home from the grocery store in the rain."

Note: The generated explanations are not yet close to the reference ones, indicating that further tuning or model improvements might be needed.

6. Conclusion and Observations

- **Model & Training:**
  - We implemented a TURBO-like model combining BART (for text generation) and ViT (for image encoding) with a Shared Fusion module.
  - Hyperparameters were set as described above (batch size = 4, LR = 1e-4, 10 epochs).
  - The training loss decreased over epochs; however, the best validation loss was observed at epoch 1, suggesting possible overfitting in later epochs.
- **Evaluation:**
  - Evaluation metrics (ROUGE, BLEU, METEOR, BERTScore) were computed after each epoch.
  - The best model (based on the lowest validation loss) was saved as `best_checkpoint.pt`.
- **Qualitative Analysis:**
  - Sample predictions show that the model captures some aspects of the sarcasm but does not fully match the nuanced explanations provided in the gold references.
- **Future Work:**
  - Further tuning of hyperparameters and model architecture (e.g., advanced gating, regularization techniques, extended training) is needed.
  - Consider incorporating external knowledge or improving data augmentation for better performance.