Background and Objectives

A typical machine learning (ML) model development flow is shown in the figure above.

- Numerical tables are widely employed to communicate or report the classification performance of machine learning (ML) models with respect to a set of evaluation metrics. For non-experts, domain knowledge is required to fully understand and interpret the information presented by numerical tables.

This paper proposes a new natural language generation (NLG) task where numerical models are trained to generate textual explanations, analytically describing the classification performance of ML models based on the metrics' scores reported in the tables.

- Presenting the generated text along with the numerical tables (as shown below) will allow for a better understanding of the classification performance of ML models.

Model Performance Narration Model

- Our experiment and analysis are based on the T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) language models.
- The input to the NLG models is a numerical table containing the evaluation metrics' scores of a classifier, the list of class labels and the distribution of the underlying data across the two classes: C1 and C2.
- The performance report tables are converted to flat-strings by a linearizer based on a pre-defined template.
- To improve the quality of the narrative generated, the neural generator as shown below is augmented with semantic representations of the metrics information generated by a neural module: Metrics Processing Unit (MPU).

Experimental Setup: Dataset, Model Settings, Training, and Inference

- The dataset for this task comprises 825 analytical textual explanations written by experts summarizing the performance of ML models trained on 59 classification tasks across different application domains. 100 data-narrative pairs are sampled as the test set.
- Permutating the order in which the evaluation metrics and their corresponding scores are passed to the models increased the training set size from 725 to 4529.
- The fine-tuning of the T5 and BART models is performed using the Adam optimizer [Kingma and Ba, 2014] with an initial learning rate of 3e-4.

- For simplicity, the T5-variant+MPU and BART-variant+MPU respectively, denote the T5 and BART model variants augmented with the auxiliary information from the MPU.

Experimental Results

- The performance of the models are evaluated based on the metrics: BLEU, METEOR, BLEURT, and PARENT.
- The different variants of the pre-trained models achieved varying scores.
- Augmenting the linearised representation with the semantic representations of the metrics information in the table with MPU is shown to further improve the generation performance of the underlying models. Example compare the performance of T5-small and T5-small+MPU models.

Examples of Generated Model Performance Narrations

- The performance of the classifier on the table conditions were evaluated based on the metrics, METEOR, BLEURT, and PARENT.
- The different variants of the pre-trained language models to generate accurate analytical textual explanations for numerical tables contain metric information in the table with MPU is shown to further improve the generation performance of the underlying models. Example compare the performance of T5-small and T5-small+MPU models.

Sample Narrative output from the T5-small and base model variants with and without the MPU

- As shown, the neural generators can produce high-quality classification performance explanations capturing the information presented in the input structured data. There were some cases where the generator fails (highlighted in red) to accurately verbalise the content of the performance metric table.
- However, the metrics-values-ratings contextual information from the MPU allow the pre-trained language models to generate accurate analytical textual explanations based on the related table.