iML Project Proposal

Alexey Wratschinski, David Schell, Lukas Marusenko
github.com/Lukas311202/Oldboy

1 Motivation

Complex models often behave like black boxes, making their decision processes difficult to interpret. To improve our understanding of these models, we aim to identify and penalize the use of *shortcuts*, spurious patterns that allow a model to make predictions without genuine reasoning. We use the term shortcut to describe situations in which a model, for example, encounters a specific token in a review and immediately assigns a particular rating based solely on that token. By discouraging reliance on such shortcuts, we seek to encourage more robust and meaningful model behavior.

2 Related Topics

Integrated Gradients is a feature attribution method that helps interpret black-box models by measuring how much each input feature contributes to a prediction. By comparing the model's gradients along a path from a chosen baseline to the actual input, it allows us to identify whether the model relies on shortcut features or meaningful signals.

3 Idea

Our project aims to improve the robustness of a BERT model on the IMDB movie review dataset by discouraging reliance on spurious shortcuts. We will inject irrelevant tokens into positive reviews to simulate shortcuts and then fine-tune the model on this augmented dataset. Using local explanation methods, such as Integrated Gradients (possibly via the Captum library), we will identify when the model relies on these shortcuts. A penalty term will be applied to the explanation if it highlights irrelevant features, and the total loss will be defined as something like

$$\text{Loss}_{total} = \text{Loss}_{classification} + \lambda \cdot \text{Loss}_{explanation}.$$

Finally, the model will be refit based on this adjusted loss to encourage more meaningful feature use.

4 Experiments

Dataset & Metrics We will use the IMDB movie review dataset and fine-tune a pre-trained BERT model on it. Metrics will include classification accuracy as well as explanation fidelity, measuring whether the model's attributions correspond to meaningful tokens rather than injected shortcuts.

Experimental Scope We plan to run several experiments including: baseline BERT training, training with injected shortcuts, and training with the explanation-adjusted penalty. We will use multiple random seeds and explore different values of λ for the explanation loss possibly with methods like Grid Search.

Estimated Computational Load Fine-tuning BERT on the IMDB dataset with explanation-based penalties is expected to take some minutes to few hours per experiment on a single GPU. Memory requirements will be moderate, mostly for storing gradients during Integrated Gradients computation.

We will run all experiments on a GPU-enabled environment and monitor resource usage to adjust batch sizes as needed.

5 Timeline

Given the 2-week timeframe, we plan the project as follows:

- **Research (2 days):** Quickly review relevant literature on BERT fine-tuning, shortcut learning, and local explanation methods such as Integrated Gradients. Familiarize with the Captum library for interpretability.
- **Implementation (5 days):** Set up the BERT model and IMDB dataset. Implement shortcut injection and integrate the explanation-based penalty into the training loop. Ensure the total loss $\text{Loss}_{total} = \text{Loss}_{classification} + \lambda \cdot \text{Loss}_{explanation}$ works correctly.
- **Experiments (4 days):** Run baseline and adjusted training experiments with a few selected seeds and hyperparameter settings. Collect classification and explanation metrics.
- **Analysis (2 days):** Analyze results to assess whether the model relies less on shortcuts and produces more meaningful explanations. Compare metrics across experiments.
- **Reporting (1 day):** Prepare the final report summarizing methodology, experiments, results, and conclusions.