Adrian Kukla

CS7643 HW3

Question 4

Question 4.7

Key contributions

- Introduces BART which is a denoising sequence-to-sequence pre-training approach that unifies the strengths of bidirectional encoders (like BERT) and autoregressive decoders (like GPT) by learning to reconstruct corrupted text.
- Proposes a framework for pre-training using various noising strategies that can be adapted for a wide range of NLP tasks.
- Provides ablation studies that analyse the impact of different noise functions on downstream task performance, offering insights into the design of pre-training goals.

Strengths

- Multiple comparisons to appropriate benchmarks and provided the magnitude of the improvement. Quantitative metrics help provide an objective basis for comparing different models.
- A thorough experimental approach was used by the authors of the paper due to multiple different end objectives tested like natural language generation, translation and reasoning/comprehension.
- Graphical representation of models and transformations is helpful for understanding, especially for lay people.

Weaknesses

- The extensive computational demands pre-training and fine-tuning required for BART can be computationally expensive, potentially limiting accessibility for some research teams.
- The effectiveness of the chosen noising strategies can vary across tasks and BART's performance varies materially by different objective, so there is room for improvement.
- Need to tune hyperparameters to particular downstream tasks requires significant tuning to balance performance and model capacity.

Question 4.8

Personal takeaways

I learned that there is potential to innovate by implementing masking transformations to introduce noising schemes for model pre-training, that the authors haven't considered in the paper. There is further investigation warranted to evaluate task-specific noise schemes and optimization parameters and computational resources which could lead to better NLP frameworks.

Moreover, I realized that it is possible to combine various Transformer models to create ensemble models, where one can stack models and use the output of one model as an input into another. I knew this was possible with boosting/bagging models like XGBoost, but didn't realize the possibility to explore in this area in the sphere of Transformer models.

Lastly, I learned that it's important to test models on a range of different objectives and to highlight the importance of data and transformations in experimental design.