Final Project Proposal

(1) Project title:

Visualizing And Mitigating Biases in Encoder-Decoder Language Models

(2) Team members:

(Yoda Group)

Lorraine Chen (xc2425@nyu.edu)
Rakshit Lodha (rl4563@nyu.edu)
Swapan Jain (sj2594@nyu.edu)

(3) **Key idea** and datasets:

The key idea of our project is to investigate the sensitivity of the attention mechanisms towards the input data bias. We will adopt pretrained language models from Huggingface libraries which have the encoder-decoder structure to train on neural machine translation tasks on both unfair datasets and fair datasets. Through visualizing attention heads on both datasets, we hope to further find out the bias-sensitive tokens. To mitigate the influence of input data bias, we propose a method by fine tuning the model on a fair dataset. In the result, we will present our comparison analysis and preliminary results on the effectiveness of our proposed retraining method. If time allowed, we would like to explore more mitigating methods based on visualizing observation.

Datasets:
https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification
https://github.com/pliang279/LM_bias

Visualization Toolkit Github:
https://github.com/jessevig/bertviz

(4) Deliverables:

1. Proposed mitigation strategy towards input textual data bias respect to some sensitive social topics.
2. Github repo including our code implementation and a pdf file of report
3. Report including
   a. preliminary results on bias mitigation experiment and its comparison to baseline
   b. visualization of attention heads in encoder/decoder
   c. result analysis
   d. If any, mathematical analysis explanation

(5) Timeline:
Nov 30: We will be done with the model training and attention head visualization, finding both unfair and fair datasets and detecting bias.

Dec 16: Done with the finetuning and mathematical reasoning analysis.

(6) Novelty concern: To the best of our knowledge, this proposal is the first to study bias mitigating methods on encoder-decoder language models via attention head visualization and fair dataset retraining in application of neural machine translation(NMT).

References

Liang, P. P., Wu, C., Morency, L. P., & Salakhutdinov, R. (2021, July). Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning* (pp. 6565-6576). PMLR.

Li, B., Peng, H., Sainju, R., Yang, J., Yang, L., Liang, Y., ... & Ding, C. (2021). Detecting Gender Bias in Transformer-based Models: A Case Study on BERT. *arXiv preprint arXiv:2110.15733*.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.