# MLSS. Final project proposal

**Option Five: Paper Implementation**

I intend to re-implement the Transformer architecture as per the paper [1] using core PyTorch functionality (not using their torch.nn.Transformer), i.e. their Encoder and Decoder parts with regularization and hyperparameters as in the paper. The dataset for the training and testing is the WMT 2014 English to German corpus [2]. Specifically, for the training set, it is Europarl Parallel Corpus with English, German and German-English [3] datasets. And for the test set, it is Q4/2000 portion of the data. Hardware is that available at SageMaker or Google Colab Pro (to train the base model for 100K steps, 12 hours)

The deliverable is a github repository with ipynb, py files, a README file describing how to run the code, a report about what results the model achieves on the Europarl Parallel Corpus.

Ways to extend the work:

1. Apply adversarial attack as per [Linyang Li, 2020]
2. Try the hallucinations in NMT as per [5]. Then detect them, apply a fix.
3. Other possible monitoring directions as per [D.Hendrycks et al. 2021]

References:


- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762
- [2] https://www.statmt.org/wmt14/translation-task.html
- [3] https://www.statmt.org/europarl/
- [Linyang Li, 2020] BERT-ATTACK: Adversarial Attack Against BERT Using BERT. https://arxiv.org/abs/2004.09984
- [5] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. "Hallucinations in neural machine translation"
- [D.Hendrycks et al. 2021] Unsolved Problems in ML Safety. https://arxiv.org/abs/2109.13916