Notes on papers relevant to our project

Stephen Malina

October 14, 2019

Abstract

There be dragons here, except the dragons are my raw thoughts. Yes, my thoughts have in fact been known to eat people.

Contents

1 Project Criteria & Ideas 3
1.1 Criteria 3
1.2 Ideas 3

2 Transformers: DNA Disguised 3
2.1 Relevant Work 3
2.2 Elevator Pitch 3
2.3 Risks 3
2.4 Potential Tasks 4

3 Chromatin & Epigenomics 4
3.1 ChromHMM 4

1 Project Criteria & Ideas

1.1 Criteria

- We're interested in doing something that illuminates a mechanism behind some sort of biological phenomenon.
- We're both interested in recurrent-style NNs (includes Transformers).
- Stephen's interested in causality.
- Stephen's interested in network learning.
- Stephen's interested in model interpretability.

1.2 Ideas

- Participate in the single cell breast cancer prediction challenge.
- Building blocks of interpretability inspired new interpretability technique that doesn't just look at specific examples.
- (Stephen's favorite): Pre-train a Transformer to predict DNA sequences and then aggressively fine-tune it on a task like TF binding prediction.

2 Transformers: DNA Disguised

2.1 Relevant Work

- Genomic ULMFit: uses an NLP pre-training technique called ULMFit to do SOTA on a bunch of different genomic prediction tasks. Only code, no paper though.
- Rives et al. (2019) scales Transformer training to 250 million proteins and test their model's ability to classify proteins, predict their alignment features, and more.
- Quang and Xie (2016) uses a hybrid convolutional/recurrent architecture to predc

2.2 Elevator Pitch

- The ground truth for DNA is the sequence information.
- Transformers seem to work fairly well for learning info about proteins.
- Biology people aren't going to buy into these sorts of methods until someone shows that they can actually be used to go deep into a biological problem.

2.3 Risks

Will Transformers work well on sequences pulled from a 4-character alphabet?

2.4 Potential Tasks

- Identifying chromatin state, i.e. compare to ChromHMM.
- Predicting transcription factor binding.
- Replicating DeepCpg.

3 Chromatin & Epigenomics

3.1 ChromHMM

ChromHMM is (surprise!) a Hidden Markov Model that annotates the genome with presence/absence tags for a large number of different chromatin annotations, i.e. different types of histone marks.

References

Quang, D. and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, 44(11):e107.

Rahimi, A. and Recht, B. (2009). Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In *Neural Information Processing Systems*, pages 1–8.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Lawrence Zitnick, C., Ma, J., and Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.