AI powered Clinical Pathologist

Abdullah Moazzam, Faisal Maqbool, Syed Shahbaz Hussain

Information Technology University Lahore, Pakistan.

SPIDER Lab
Lab Dir. Ali Ahmed

Abstract

We have developed a deep learning based system that can mimic the work of a clinical pathologist by distinguishing between different types of genetic mutations, based on the associated clinical evidence. Our approach is tested on Memorial Sloan Kettering Cancer Centers' (MSKCC) expert annotated dataset, which contains a list of 3500 genetic mutations along with the associated scientific literature that was used to assign them 9 different labels.

Problem Statement

- Cancer tumors can have thousands of mutations ranging from driver mutations (those that contribute towards cancerous growth) to passenger mutations (those that do not)
- Advances in DNA sequencing technology have made DNA sequencing of cancerous tumors readily available
- The challenge, now, is to distinguish between different types of genetic mutations
- Currently, this requires a lot of manual labor, as clinical pathologists have to manually classify each and every gene by going through scientific literature
- This project seeks to automate this task by using machine learning models to classify genetic mutations based on clinical evidence

Applications/Project Promise

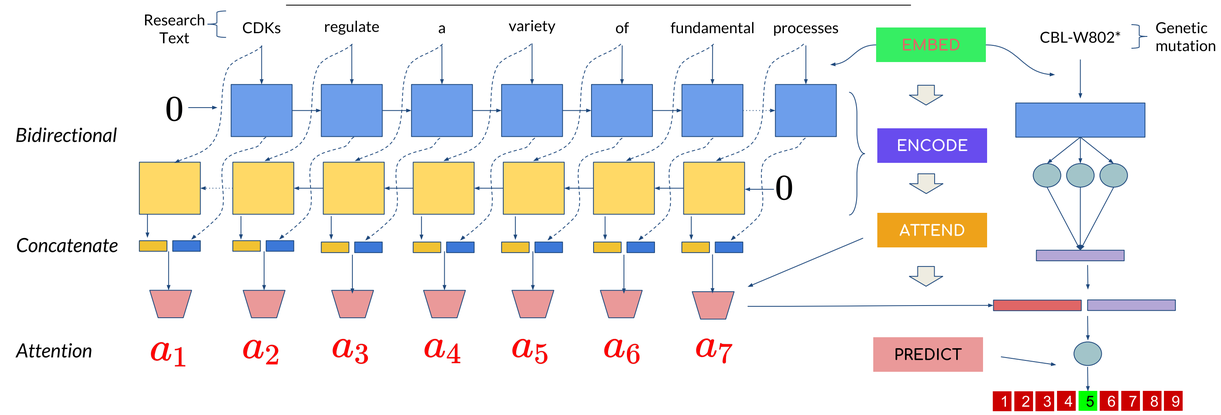
The project falls into the domain of the rapidly evolving field of BioNLP and still needs to mature before being of any major use. However, potential applications include:

- Acting as an assistant to clinical pathologists as they scan through literature in order to distinguish between different genetic mutations
- With the availability of more annotated data, the model can become more robust and may have the potential to automate the task of classifying all genetic mutations in a cancerous tumor.

Associated clinical evidence is available for every genetic mutation that needs to be classified

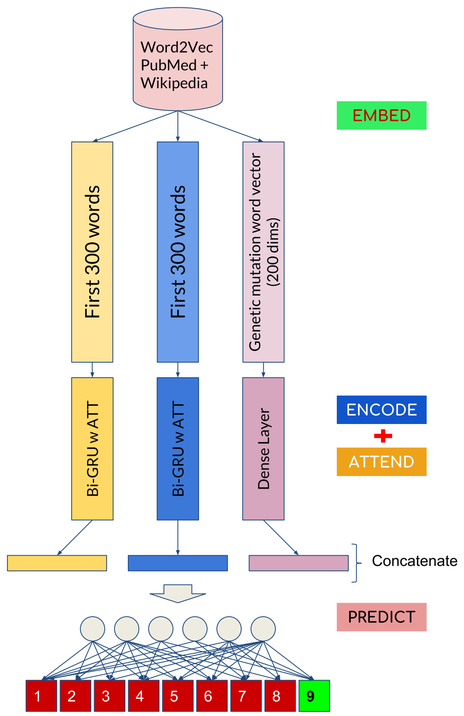
Key Assumption!

Network Architecture



Methodology

- We have used a modified version of the NLP framework used in seq2seq model [2], popularized by Mathew Honiball as the 'Embed, Encode, Attend and Predict' framework for Deep NLP models [3]
- A visual depiction of our method is given below



Data Encoding 1: One bidirectional GRU trained on the first 300 words from research text + 30 most similar words to genetic mutation (concatenated)

Data Encoding 2: Two bidirectional GRUs trained on the first 300 and last 300 words of research text + feature vector corresponding to genetic mutation acquired from Word2Vec model (concatenated before softmax)

Class Imbalance

- The dataset is highly imbalanced
- To cope with class imbalance we have used a loss function called focal loss [1].

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

- Focal loss reduces the relative loss for well-classified examples, putting more emphasis on hard, misclassified examples.

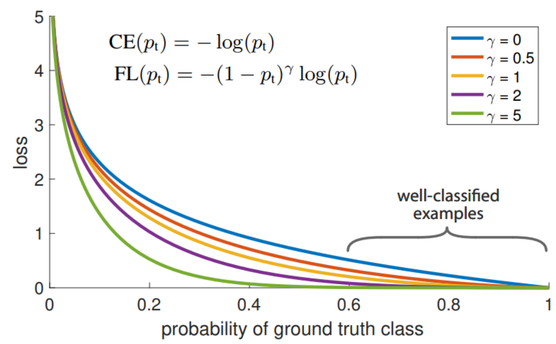


Figure 1. Increasing the value of γ reduces the relative loss for well-classified examples while placing greater focus on hard examples.

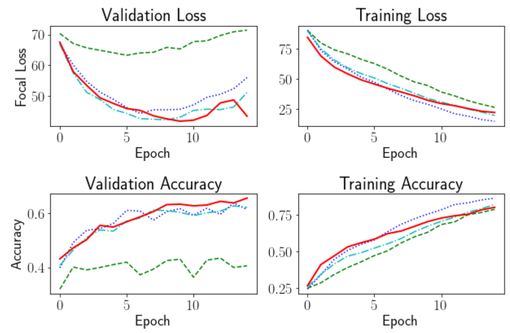


Figure 2. Line in red corresponds to our proposed model, Bi-GRU with attention using Data Encoding 2. All the rest follow Data encoding 1, with green being a simple GRU, blue being Bi-GRU and cyan being Bi-GRU with attention.

Kaggle Competition Result

Sr. No	Model Description	Private Leaderboard Score	Public Leaderboard Score
1	Top in private leaderboard (ilmirashaim)	2.03027	1.47447
2	Simple GRU (DE1)	2.58223	1.79227
3	Bi-GRU (DE1)	**2.46368**	1.32249
4	Bi-GRU (DE1)	2.52356	1.30881
5	Bi-GRU with Attention (DE1)	2.51774	1.29545
6	Bi-GRU with Attention (DE2)	2.56705	**1.10625**

Table 1. Kaggle Submissions. The best performing model belongs in the top 30 in the competition's private leaderboard

References

[1] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. arXiv preprint arXiv:1708.02002.

[2] Luong, M. T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114.

[3] Honnibal, Mathew. "Embed, Encode, Attend, Predict: The New Deep Learning Formula for State-of-the-Art NLP Models · Blog · Explosion AI." Explosion AI, explosion.ai/blog/deep-learning-formula-nlp.

[4] Moen, S.P.F.G.H., & Ananiadou, T. S. S. (2013). Distributional semantics resources for biomedical text processing. In Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan (pp. 39-43).