Using Artificial Intelligence to Prescribe Medicine

Xander Bush[1] and Louis Livingston[2]

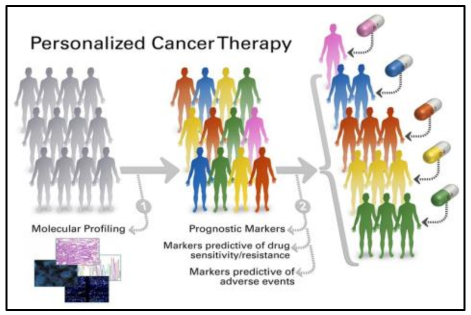
[1]abush27@students.kennesaw.edu, [2]lliving8@students.kennesaw.edu

CS 4267 Project Advisor: Aledhari, Mohammed

Introduction

Personalized Medicine is becoming a more popular solution to modern day medical problems by combining Clinical Diagnostics and Machine Learning techniques. By adopting powerful preprocessing techniques such as word2vec and the TensorFlow platform, it is possible to create Artificial Neural Network inputs from notes made by real doctors.

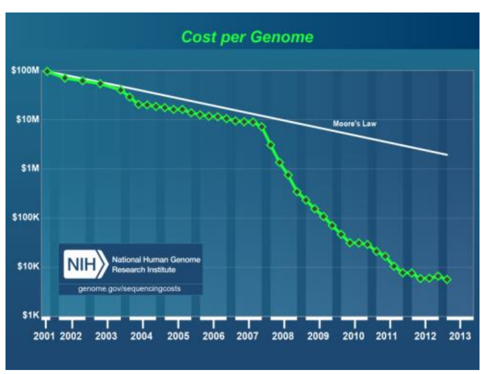
Machine Learning Techniques to imitate Doctor's observations



https://pct.mdanderson.org/

As cancer treatment can vary heavily from case to case, the usage of personalized treatment plans are used. This can be simplified through the usage of Machine Learning techniques like the Artificial Neural Network as demonstrated on the left.

According to the National Human Genome Research Institute, the cost of sequencing a human genome has gone down much more than originally expected and has allowed for high throughput programs to better examine the highly complex structure more efficiently.

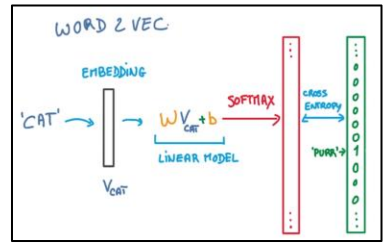


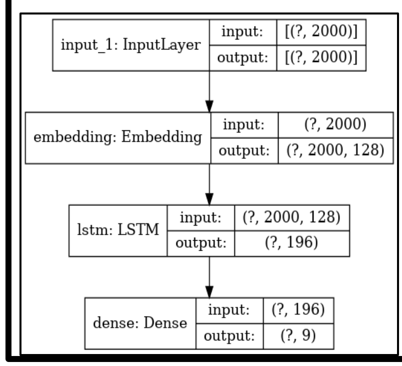
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Genomics: interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.

Techniques Used

To the right is an example of how the Word2Vec preprocessing technique transforms and obtains quantitative values out of text-based documents. This technique allows for the usage of our clinical notes database as input for the ANN.





To the left is a visualization of the project's Artificial Neural Network process. Through the combination of Embedding, LSTM, and Dense style layers, the encoded clinical notes are processed and subsequently classified through this ANN architecture.

Summary /Gathered Data

The usage of Clinical Diagnoses as input data for our hidden layers allows the ANN to process large amounts of text extremely quickly.

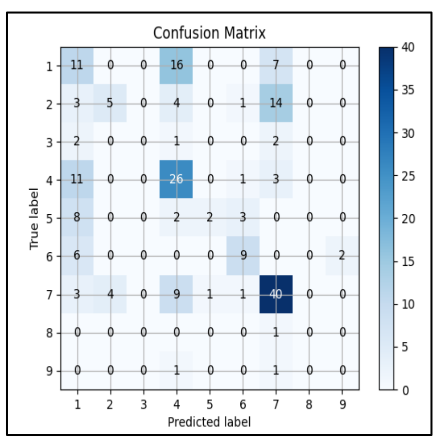
Layer (type)	Output Shape	Number of Params
Embedding	(None, 2000, 128)	256000
LSTM	(None, 196)	254800
Dense	(None, 9)	1773
	Total params:	512,573
	Trainable params:	512,573
	Non-trainable params:	0
	Log Loss:	1.444628581
	Accuracy:	0.445

	ID	Gene	Variation	Text
0	0	ACSL4	R570S	2. This mutation resulted in a myeloproliferat...
1	1	NAGLU	P521L	Abstract The Large Tumor Suppressor 1 (LATS1)...
2	2	PAH	L333F	Vascular endothelial growth factor receptor (V...
3	3	ING1	A148D	Inflammatory myofibroblastic tumor (IMT) is a ...
4	4	TMEM216	G77A	Abstract Retinoblastoma is a pediatric retina...

Through the usage of the baseline project code, we were able to gather large amounts of analytic data on the efficiency of the project as well as the loss functions and accuracy score.

	ID	Gene	Variation	Class	Text
count	3321	3321	3321	3321	3316
unique	0	264	2996	0	1920
top	0			0	
freq	0	264	93	0	53
mean	1660	0	0	4.365854	0
std	958.834449	0	0	2.309781	0
min	0	0	0	1	0
25%	830	0	0	2	0
50%	1660	0	0	4	0
75%	2490	0	0	7	0
max	3320	0	0	9	0

Results



This confusion matrix shows the distribution of our first Neural Network Model through the usage of gradients.

Model	Baseline Project	New Model	% Change
Loss Function	1.444	7.222	+400.14%
Accuracy	44.5%	45%	+1.12%

Future Plans

Our future plans for this project include the following improvements:

1. Improvement to throughput
2. Integration of GPU processing
3. Improvements to accuracy
4. Integration of multiple data sets
5. Allow for usage of various types of input aside from specifically Clinical Research Notes

Source Code and Resources

Dataset: https://www.kaggle.com/c/msk-redefining-cancer-treatment/data

Source Code: https://github.com/louisliv/machinelearningproject