Few shot learning for cancer detection

Marco Russo, Rocco Zaccagnino

Department of Computer Science, University of Salerno

m.russo269@studenti.unisa.it, r.zaccagnino@unisa.it

Abstract

Abstract In recent years, thanks to the technological advancement of data processing capacity, it has been possible to almost completely sequence the human genome. This opened the doors to analyze in details mutational patterns in human cancers. Previews works [1] were able to identify correlations between mutations and type of cancer. We propose a Siamese Neural Network in order to find similarities between cancers with a few shot learning technique.

1. Introduction

1.1. Gene sequencing

The Human Genome Project was an international research effort that aimed to sequence and map the entire human genome. The project officially began in 1990 and was completed in 2003, with the publication of the complete human genome sequence.

The primary goal of the Human Genome Project was to identify and map all of the genes in the human genome, as well as to understand their function and how they interact with one another. By sequencing the entire human genome, researchers were able to identify approximately 20,000 to 25,000 protein-coding genes, as well as many other non-coding regions of DNA that play important roles in gene regulation and other cellular processes.

In addition to identifying and mapping genes, the Human Genome Project also had many other important implications and applications. For example, it has helped researchers to better understand the genetic basis of many diseases, including cancer, heart disease, and diabetes. It has also led to the development of new genetic testing and screening technologies, as well as new treatments for genetic diseases.

This data, combined with advances in machine learning and deep neural network (DNN) techniques, has enabled researchers to develop more accurate and effective methods for cancer detection and diagnosis.

For example, machine learning algorithms can be trained on large datasets of genetic and clinical data to identify patterns and biomarkers that are associated with specific types of cancer. These algorithms can then be used to develop predictive models that can accurately diagnose cancer and predict patient outcomes.

1.2. Siamese Network

A Siamese network is a type of neural network architecture that is used for tasks related to similarity learning. The architecture of a Siamese network typically consists of two or more identical subnetworks that share the same weights and are trained to perform a similarity function between two input data points. The output of each subnetwork is a feature vector that is used to calculate the similarity score between the two inputs.

Siamese networks have been used in a variety of applications, such as face recognition, image retrieval, and natural language processing, where the task involves comparing two inputs and determining how similar or dissimilar they are. One of the advantages of using a Siamese network for similarity learning is that it can be trained with relatively small amounts of data, as it does not require a large labeled dataset.

2. Related Works

2.1. PCAWG

The PCAWG (Pan-Cancer Analysis of Whole Genomes) [1] dataset is a large-scale genomic dataset that was generated by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) as part of a collaborative effort to comprehensively analyze the genomic landscape of cancer.

The PCAWG dataset includes whole-genome sequencing data from over 2,500 cancer patients, covering 38 different cancer types. The data includes information on somatic mutations, copy number alterations, structural variants, gene expression, DNA methylation, and other genomic features that are relevant for cancer research.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium created a deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns [**?**], this work was really focused on creating the dataset and used really simple models, in fact, it used a Random Forest Classifier and a DNN, which both gave good results.

2.2. Few Shot Learning

In few-shot learning [2], the goal is to learn from a small number of examples, and the model is trained on a support set that consists of a small set of labeled examples. The support set is used to assist in the training of the model, and data augmentation techniques are often applied to generate additional training examples. Few-shot learning is particularly useful in situations where the amount of labeled training data is limited, or where it is not feasible to obtain large amounts of labeled data.

In a siamese network, the support set is used to help the network learn to recognize similarities and differences between samples, and a specific type of data augmentation technique known as "pairwise coupling" is often applied. This involves combining each element of the support set with each element of the training set to create new pairs of data points. By generating new pairs of data in this way, the siamese network is able to learn to recognize similarities and differences between samples more effectively, leading to improved performance.

During training, the siamese network learns to encode each data point as a vector, and then calculates a distance metric between the vectors to determine the similarity between the two data points. By using pairwise coupling with the support set, the network is able to learn more robust and generalizable representations of the data, which can improve its performance on a variety of tasks. The use of a support set and pairwise coupling data augmentation is particularly useful for tasks such as image recognition or similarity matching, where it is important to be able to recognize similarities and differences between images or other types of data.

3. Result and Discussion

The dataset used for this project is the MSK-Impact [3]. The MSK-IMPACT dataset is a genomic profiling dataset generated by Memorial Sloan Kettering Cancer Center (MSKCC) in New York, USA. It contains molecular profiling data for over 30,000 cancer patients across multiple cancer types, including but not limited to breast, lung, prostate, and colorectal cancer.

The data is generated using next-generation sequencing (NGS) technologies, which allow for the comprehensive and high-throughput analysis of cancer genomes. Specifically, the dataset includes information on somatic mutations, copy number alterations, gene expression, and other molecular features that are relevant for cancer diagnosis, prognosis, and treatment.

3.1. Preprocessing

The dataset underwent a series of normalization procedures, beginning with the removal of columns that contained incomplete data. Next, non-numeric columns were encoded using a one-hot encoding scheme. The resulting data was then subjected to principal component analysis (PCA), after which it was again normalized.

3.2. Neural Network

The dataset was initially trained on a DNN for multiclass classification. Subsequently, the classification layer was removed from this model, the weights freezed, and the resulting network was used on the two branches of the siamese network to extract features from the data.

Finally, the resulting features were passed through a distance function (Euclidean distance), processed through hidden layers, and finally a binary classifier determined whether the two input elements belonged to the same class or not (0 or 1).

3.3. Results

Table 1. DNN Model evaluation

Loss	Accuracy	Precision	Recall	AUC
0.5457	0.8975	0.9196	0.8861	0.9744

Cancer	Loss	Accuracy	Precision	Recall	AUC	F1 Score
Non-Small Cell Lung Cancer	0.222917	0.962963	0.962500	0.950617	0.993014	0.956522
Breast Cancer	0.403626	0.931035	0.929825	0.913793	0.972661	0.921739
Bladder Cancer	0.907647	0.833333	1.000000	0.750000	0.945833	0.857143
Head and Neck Cancer	3.557362	0.500000	0.555556	0.500000	0.747333	0.526316
Bone Cancer	0.874127	0.785714	0.846154	0.785714	0.957313	0.814815
Soft Tissue Sarcoma	2.417866	0.629630	0.680000	0.629630	0.866530	0.653846
Prostate Cancer	0.011922	1.000000	1.000000	1.000000	1.000000	1.000000
Ovarian Cancer	2.306796	0.375000	0.500000	0.375000	0.852083	0.428571
Glioma	0.017366	1.000000	1.000000	1.000000	1.000000	1.000000
Esophagogastric Cancer	1.004104	0.833333	0.833333	0.833333	0.954398	0.833333
Hepatobiliary Cancer	0.395946	0.944444	0.944444	0.944444	0.970988	0.944444
Colorectal Cancer	0.110219	0.958333	0.956522	0.916667	0.999653	0.936170
Cancer of Unknown Primary	1.058346	0.777778	0.875000	0.777778	0.930041	0.823529
Endometrial Cancer	1.526419	0.600000	0.750000	0.600000	0.948000	0.666667
Melanoma	0.315476	0.888889	0.941176	0.888889	0.997737	0.914286
Pancreatic Cancer	0.445821	0.857143	0.850000	0.809524	0.994709	0.829268

Table 2. Classification performance metrics for various cancer types on the DNN

Table 3. Siamese Model evaluation

Loss	Accuracy	Precision	Recall	AUC
0.1226	0.9661	0.8490	0.5569	0.8958

Cancer	Loss	Accuracy	Precision	Recall	AUC	F1 Score
Non-Small Cell Lung Cancer	0.074384	0.982253	0.960317	0.746914	0.960181	0.840278
Breast Cancer	0.084038	0.973060	0.883721	0.655172	0.975840	0.752475
Bladder Cancer	0.129697	0.953125	0.687500	0.458333	0.925463	0.550000
Head and Neck Cancer	0.221959	0.943750	0.666667	0.200000	0.680583	0.307692
Bone Cancer	0.091722	0.973214	0.785714	0.785714	0.960162	0.785714
Soft Tissue Sarcoma	0.185826	0.945602	0.594595	0.407407	0.791427	0.483516
Prostate Cancer	0.101098	0.969122	0.967033	0.523810	0.960138	0.679537
Ovarian Cancer	0.227238	0.917969	0.368421	0.437500	0.786068	0.400000
Glioma	0.237458	0.943257	1.000000	0.092105	0.707768	0.168675
Esophagogastric Cancer	0.104480	0.973958	0.733333	0.916667	0.971123	0.814815
Hepatobiliary Cancer	0.064282	0.987847	0.939394	0.861111	0.985597	0.898551
Colorectal Cancer	0.082385	0.971354	0.842105	0.666667	0.977836	0.744186
Cancer of Unknown Primary	0.107919	0.961806	0.733333	0.611111	0.969753	0.666667
Endometrial Cancer	0.204249	0.918750	0.333333	0.300000	0.874667	0.315789
Melanoma	0.074745	0.989583	1.000000	0.833333	0.966641	0.909091
Pancreatic Cancer	0.274922	0.936012	0.400000	0.047619	0.532445	0.085106

Table 4. Performance metrics for various cancer types on the Siamese Network

4. Conclusion

In conclusion, our study has demonstrated that the siamese neural network architecture can be successfully applied for the task. Our model achieved good performance on the given dataset, with promising results for future application in real-world scenarios. However, it is worth noting that the limited size of the dataset used in our study may have contributed to the model's performance, and we believe that the performance of the model can be further improved with the inclusion of additional datasets. Therefore, we suggest that future studies should focus on collecting larger and more diverse datasets to evaluate the performance of the model in a wider range of contexts. Overall, our findings suggest that the siamese neural network architecture has great potential for improving the accuracy and efficiency, and we look forward to seeing its continued development and application in the field.

References

[1] Wei Jiao, PCAWG Consortium, et. al: A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, (2020) 11:728. [Online]. Available: https://www.nature.com/articles/s41467-019-13825-8.

[2] Andrea Fedele, Riccardo Guidotti,Dino Pedreschi: Explaining Siamese Networks in Few-Shot Learning for Audio Data, *University of Pisa, Pisa, Italy.*

[3] Zehir A,et. al.: "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients". Nat Med. 2017 Jun;23(6):703-713. doi: 10.1038/nm.4333. Epub 2017 May 8. Erratum in: Nat Med. 2017 Aug 4;23 (8):1004. PMID: 28481359; PMCID: PMC5461196.