International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952213

Elixir for Medicine

Thiyagarajan. B, Jamuna. I, Keerthana. G, Krithika. B
Department of Computer science and Engineering, Sri Manakula Vinayagar Engineering College, Pondicherry, Tamil Nadu, India

ABSTRACT

The Drug formulation is a proactive part of research in the existing Biomedical Industry is highly money consuming and the general pipeline involved in terms of time is huge that it takes around 10-12 years. People generally get typically skeptical as the time progresses. This is an essential need that needs to be addressed. The process of getting a drug from Initial Discovery to market can take years or decades. Experiments, clinical studies, and clinical trials are necessary.90% of all clinical trials in humans fail even after the molecule compounds have been successfully tested in animals.

Keywords : Artificial Intelligence, Machine Learning, Deep Learning, Generative Adversarial Networks.

I. INTRODUCTION

Artificial Intelligence (AI) has recently been developed into sizzling topic in the area of medical care industry. The biopharmaceutical industries are making efforts to approach AI to enhance drug discovery process, reduce research and development expenses, diminish failure rates in clinical trials and ultimately generate superior medicines. The accessibility of immense statistics in life sciences and a speedy development in machine learning algorithms led to an evolution of AI-based start-up companies focused on drug discovery over the recent years [1]. Numerous remarkable AI-biopharmaceutical alliance were declared in 2016-2017 that include Pfizer and IBM Watson, Sano Genzyme and Recursion Pharmaceuticals, AstraZeneca, Abbvie, Merck, Novartis, GSK and Exscientia

.

II. GENERATIVE ADVERSARIAL NETRWORKS IN DRUG DISCOVERY

A huge amount of literature in the recent years were turned onto the updation,improvisation,tuning GAN's to be fast and precise with respect to epochs. A problem in machine learning is to make an algorithm which will performs on the training data and also on new inputs. Developing more regularization techniques . one of the simplest and most common kinds of parameter: L2 penalty commonly known as weight decay. L2 regularization is also known as ridge regression or Tikhonov regularization.

The potential drugs contains more than 1060 molecules. Moreover, to test in a medical setting is time-consuming and expensive. Getting a drug to market can take up to 10 years and cost $2.6 billion. These computer-based methods are increasingly employed to accelerate drug discovery and it reduces the development costs. There is a growing interest in AI-based generative models. Their goal is to generate new lead compounds in silico, in a way that their medical and chemical properties are predicted in advance. Examples of this approach include Variational Auto Encoders, Adversarial Encoders, Reinforcement Learning in the stochastic contemporary combination initial stage: generated samples are evaluated only visually, or with metrics Evaluation would be regarding the internal chemical diversity. Generating a chemically diverse stream of molecules, because drug candidates can fail in many unexpected ways, later in the drug discovery pipeline. Based on visual inspection,reports that their model produce simplistic molecules. argues that their Objective-Reinforced Generative Adversarial Network (ORGAN) generates less repetitive and less

simplistic. It has own issues regarding the formulation of initial loss.

The major threat with training GANs are the gradual progress in the initial updation of Loss Function. This happens,Initial loss updation is slow since they generate random noises initially.Compared to GAN's in music generation and text generation,GAN's for image generation is of higher dimensional order. The attainment of local minima consumes a huge amount of time and epochs. In here, we propose a new form of regularization parameter to enthrill the process in ChemGAN challenge.

III. METRICS

The metric of internal chemical diversity Let a and b be two molecules, and ma and mb be their Morgan fingerprints (Rogers and Hahn 2010). Their number of common fingerprints is |ma ∩ mb| and their total number of fingerprints is |ma ∪ mb|. The Tanimoto-similarity Ts between a and b is defined by:

Ts(a, b) = |ma ∩ mb| /|ma ∪ mb| .

Their Tanimoto-distance is:

Td(a, b) = 1 – Ts(a, b)

We use rd kit implementation (Landrum 2017) of this distance.

IV. INTERNAL DIVERSITY

We define the internal diversity I of a set of molecules A of size |A| to be the average of the Tanimoto-distance Td of molecules of A with respect to each other. Formally, we have: I(A) = 1 |A| 2 X (x,y)∈A×A Td(x, y) (1) For a sufficiently large set A, any sufficiently large subset A′ ⊂ A, sampled with uniform probability, has the same internal diversity as A. This property follows from the law of large numbers. We can thus define the internal diversity of a generative model, by computing the internal diversity of a sufficiently large generated sample. This allows to formalize our challenge: Challenge (restatement): Let N be the molecules observed in nature. Is there a non-trivial generative model G and a non-trivial chemical property P such that: I(G ∩ P) ≥ I(N ∩ P) .Internal chemical diversity is always smaller than 1 (because the Tanimoto-distance is smaller than 1), and it is usually much smaller. That's why we prefer this definition to the Tanimoto-variance of a set of molecules A, which is:

V (A) = 1 |A| 2 X (x,y)∈A×A Td(x, y) 2

V. EXTERNAL DIVERSITY

A related notion is external diversity. Let A1 and A2 two sets of molecules. The relative diversity E of A1, A2 is defined by:

E(A1, A2) = 1 |A1| × |A2| X

(x,y)∈A1×A2 Td(x, y)

The external diversity of a generative model is defined as the relative diversity between the training set and a sufficiently large generated sample. External diversity essentially corresponds to the notion of diversity. A measure of the Tanimoto similarity between generated and natural molecules is also considered . The main insight of our paper is to compare internal diversities of generated and natural molecules respectively, instead of considering the relative diversity between generated and natural molecules (and also, we measure this internal diversity with respect to the subset of molecules satisfying the property of interest). We think measuring internal diversity is a good way to quantitatively capture the visually observed fact that generated molecules can be repetitive and simplistic.

VI. DEEP GENERATIVE MODELS

Deep generative models attempt to capture the probability distributions over the given data samples. Deep generative models have been applied to data generation, image classification and multimodal learning . Restricted Boltzmann Machines (RBMs) are the basis of many other hierarchical models.RBMs have been used to model the distributions of images and documents . Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) are extended from the RBMs. The top two layers of DBNs are identical to RBMs but several directed layers are added under the top two

undirected layers. The most successful application of DBNs is for image classification, where a DBN is used to extract feature representations.

The Machines are totally undirected models. One important property of DBMs compared with DBNs is that the hidden variables in a layer are conditionally independent given the other layers. This makes the posterior distribution of hidden variables more elegant than the one of DBNs. However, all of RBMs, DBNs and DBMs have the diculties of an intractable partition function or an intractable posterior distribution. Therefore approximation methods have to be used to learn the models. Another important deep generative model is Variational Autoencoders (VAE) . The VAE is a directed model and can be trained with gradient-based optimization methods. But VAEs also make an approximation for the objective function, which introduces small errors to the model. Generative Adversarial Network (GAN) is another type of generative deep model. One advantage of GANs is that it does not require any approxi-mation method and can be trained through the differentiable networks. GANs

VII. MULTI CLASS GAN

Generative adversarial networks (**GANs**) are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks contesting with each other in a zero-sum game framework. They were introduced by Ian Goodfellow *et al.* in 2014. This technique can generate photographs that look at least superficially authentic to human observers, having many realistic characteristics (though in tests people can tell real from generated in many cases).

For learning a GAN on MNIST, it is able to generate high-quality images. However, the Chinese characters are more complex than digits. For Chinese characters, it is difficult to achieve a good convergence by GANs and the generated images is difficult to be recognized For multiple classes datasets such as Chinese characters, the experimental results indicate that the conditional GANs can achieve a better convergence. The reason is that for GANs without conditions, the generator may generate images combined from two or more characters and the discriminator treats these images as real images. But for conditional GANs, the training images are conditioned on their labels, which equals to train multiple models for multiple classes. Another benet of conditional GANs learning is that the images are generated with labels and the labeled images can be more useful for other applications such as data augmentation. We leave the GANs learning on Chineses characters for data augmentation for the future work.One-hot encoding for labels is adopted for learning conditional GANs on MNIST . However, there are much more classes for characters.For examples, the Handwritten Characters dataset contains 3740 classes. If one-hot encoding vectors are directly concatenated to the network layers, the size of the network will become huge such that both the GPU memory cost and the computational time cost are infeasible. We propose to use a linear mapping layer before concatenation, which maps the big one-hot vectors into small vectors. The linear mapping layer is used as an encoder. The experimental results show that the linear mapping layer works well. To make the learning process stable, the L2 loss function is used for the discriminator.The network architecture for handwritten characters is shown ReLU activations are used for the generator and Leaky ReLu activations are used for the discriminator. The layers to be concatenated are determined by empirical results.

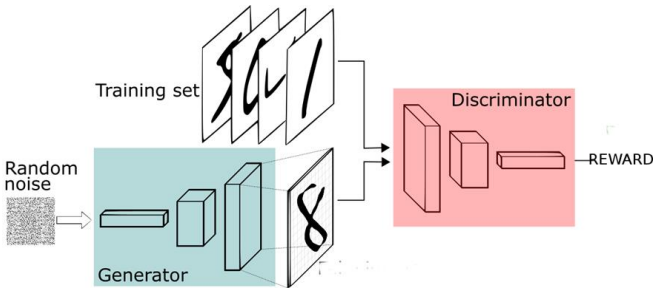

Fig 1.

VIII. EXPERIMENTATION

During the initial stage of identifying the lead molecules, we cannot be sure of anything.Live experiments in the lab are still very slow and

expensive, so we would like to find lead molecules as accurately as we can.Even if the goal is to treat cancer there is no hope to check the entire endless variation of small molecules in the lab.72 million is just the size of a specific database, the total number of small molecules is estimated to be between 10^{60} and 10^{200}.synthesizing and testing a single new molecule in the lab may cost thousands or tens of thousands of dollars.The early guessing stage is really, really important.We can use machine learning models to try and choose the molecules that are most likely to have desired properties.

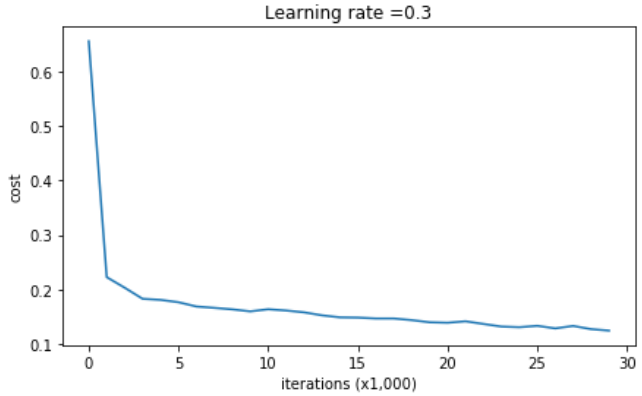

Fig 2.

IX. CONCLUSION

With our proposed solution, we could shrink the potential time taken by featuring the possible strength of the GAN's and incubating the power of autoencoders, we could possibly build something which will be able to transcribe the role of the research scientist who studies the complete bio chemical potential of the Drug and evaluating the drug, If It'll be possible to hold the facilities to cure a disease.This could be built with an amorphous Probabilistic model approach, Which possibly reduces the time taken by the human agent.

X. REFERENCES

[1]. Kustrin AS, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal 22: 717-727

[2]. Ian Goodfellow,Jean Pouget-Abadie,Mehdi Mirza,et al.

[3]. Ian J Goodfellow.On distinguishability criteria for estimating generative models.arXiv preprint arXiv:1412.6515,2014.

[4]. Daniel Jiwoong Im,Chris Dongjoo Kim,Hui Jiang and Roland Memisevic.Generating images with recurrent adversarial networks.arXiv preprint arXiv:1602.05110,2016.

[5]. Donggeun Yoo,Namil Kim,Sung gyun Park,Anthony S Paek,and In So Kweon.Pixel-level domain transfer.arXiv preprint arXiv:1603.07442,2016

[6]. Ian Goodfellow,Yoshua Bengio, and Aaron Courvilee.Deep Learning 2016.MIT Press.

[7]. Christian Szegedy,Wojciech Zaremba,Ilya Sutskever,et al.Intriguing properties of neural networks.arXiv preprint arXiv.1312.6199,2013.

Cite this article as :

Thiyagarajan. B, Jamuna. I, Keerthana. G, Krithika. B, "Elixir for Medicine", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 911-914, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952213
Journal URL : http://ijsrcseit.com/CSEIT1952213