Proposal Report (Rationale, Aims and Objectives)

1. Background, significance and basis

In recent years, machine learning (ML), especially deep learning (DL) technology has been increasingly applied in many fields. Meanwhile, material science experiences an essential transformation, i.e. from relying on empirical experience, being model-based theoretical science, and being computational science to data-driven science, as shown in Figure 1[1].

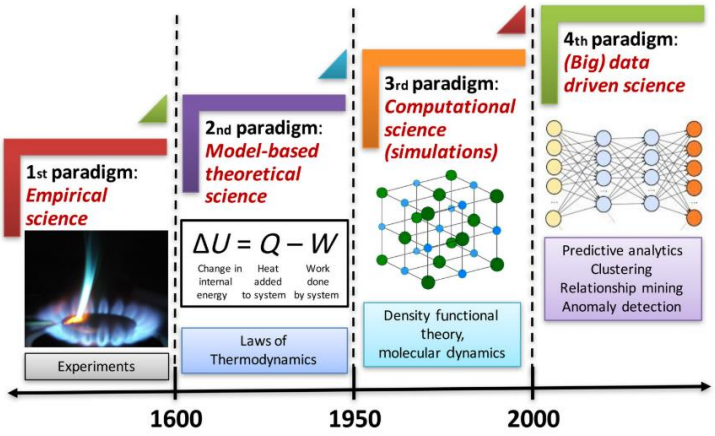

Figure 1. The paradigm transformation of material science

Machine learning has brought significant changes to our life. As a sub-branch of machine learning and a further development on perceptron, deep learning finds its specialization in data-based learning, which was initially brought in by McCulloch and Pitts in 1940s[2], when they were hoping to propose a system to mimic neurons of human brain. Thanks to the continuous efforts of AI engineers and data scientists, deep learning has been promising to learn from data in an accurate and computationally efficient manner. In computational material area, DFT based electronic calculation has brought fundamental changes to material simulation[3–6] and discovery, but soon they were found to be stuck by bottleneck at computational efficiency (typically in $O(n^3)$ scaling factor) with increasing system simulation size. Machine learning is thus the best approach to make up for the shortcomings of DFT but integrate its advantages. And due to these reasons, using ML/DL to explore material space has become one of hot researching fields.

2. Research status at home and abroad

Jha et al. developed a deep learning network called ElemNet, which could automatically

capture the physical and chemical interactions and similarities between different elements without manually featuring engineering requiring domain knowledge[7]. Artrith et al. used copper and zinc oxide as the reference system to prove the accuracy and effectiveness of the interatomic potential of the artificial neural network and described the copper oxide cluster compound Cu-Zn-O ternary composite system[8]. Huang et al. used network models such as DNN, CNN, and RNN to establish a deep learning model to describe the formation energy of inorganic compounds[9]. Xie and Grossman developed an interpretable CNN model, which could provide highly accurate prediction of DFT calculated properties for eight different properties of crystals with various structure types and compositions[10]. Ye et al. used a deep neural network and succeeded in predicting the DFT formation energies of $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskite with low mean absolute errors[11].

In summary, due to the development of machine learning technique, researchers can realize material property prediction based on elemental composition and structure. Comparing with DFT method and semi-empirical method, deep-learning-based material properties prediction provides a new way to do so in a both fast and accurate manner.

3. Research objectives, contents, methods and key technologies

This research aims to design and implement a Generative-Adversarial-Networks (GAN)-based machine learning framework (shown in Figure 2) to effectively predict material structures while taking NbO system as an example to testify the machine learning framework (shown in Figure 3).

Generative Adversarial Network is firstly brought in October 2014, by Goodfellow[12]. GAN model usually contains two main sub-networks: generator and discriminator. As their name suggested, the function of them is to generate data and discriminate the authenticity of generated data. They two constitute a typical dynamic two-player zero-benefit gaming process. Just like two boxers fighting with each other, both of which will improve themselves during the fighting. Based on the same principle, this dissertation hopes to train a GAN model to “learn” the rule of crystal structure and predict new structure.

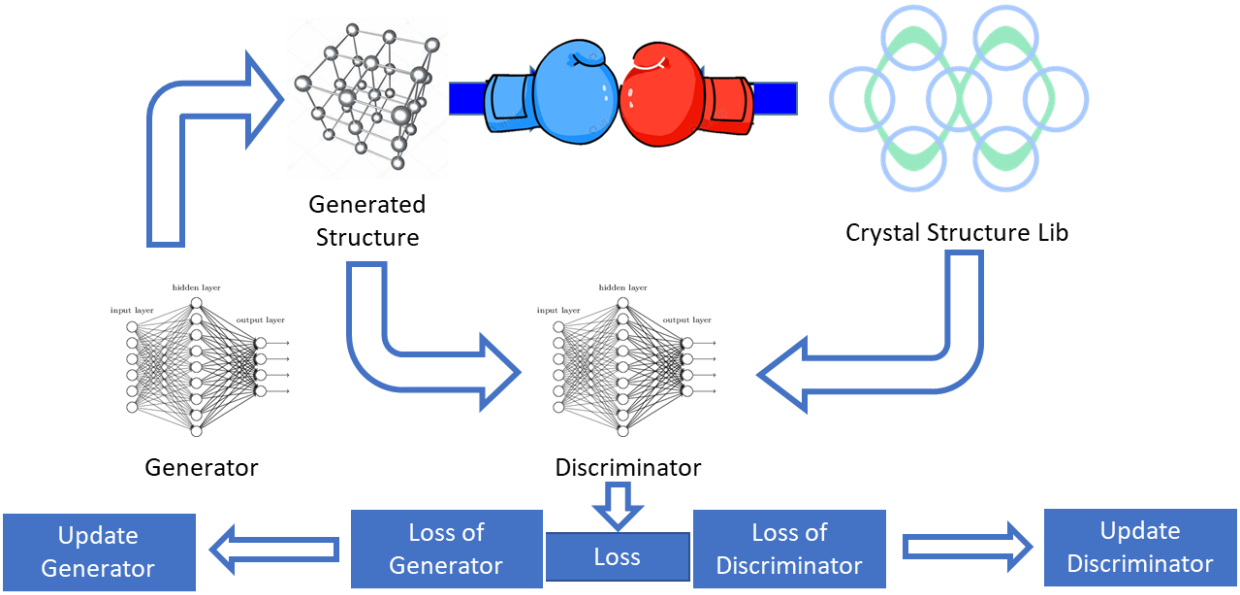

Figure 2. Illustration of Generative Adversarial Network (GAN) model

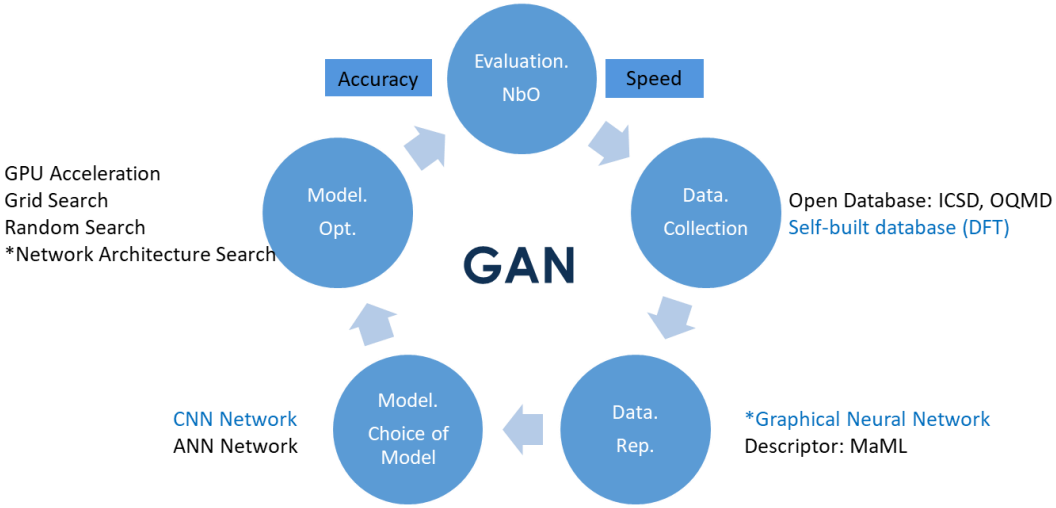

Figure 3. Flow chart of research content

3.1 Research on Data Collection

Training process of deep learning model requires an accurate database beforehand. Commonly, with the increase of training database size, performance of trained model improves. General curve of performance versus data amount is shown in Figure 4[13–15]. Thus, to keep the model performance, massive amount of training data is necessary.

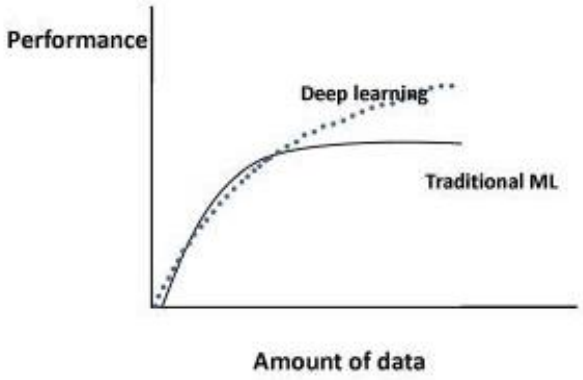

Figure 4. Learning curve of deep learning and traditional machine learning

This research aims to gather data from two main sources: 1) open-access massive database, and 2) self-built database. Open-access massive database will include: Open Quantum Mechanism Database (OQMD) and Inorganic Crystal Structure Database (ICSD), all of which are commonly used in the field of computational materials and chemistry[5,16]. Self-built database will include DFT results of NbO system obtained under several different calculation scenarios.

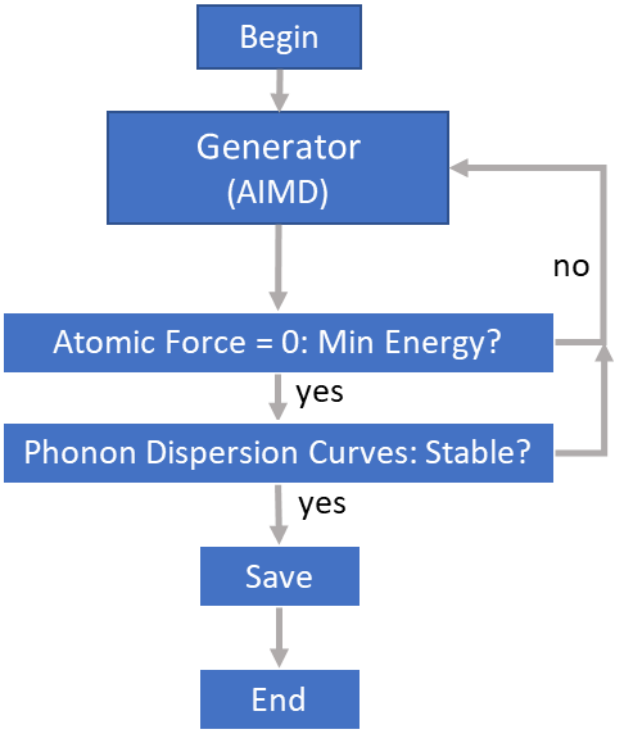

Figure 5. The workflow chart of the generation of self-built database

For the self-built database to be established, VASP will be employed for DFT calculation and potential structural generation[17]. We shall generate structure by molecular dynamic calculation and calculate the atomic force and phonon spectrum (shown in Figure 5). If there is no imaginary frequency and atomic force is zero, we consider this structure have reached a **local**

minimum of energy and thus, save it into database.

3.2 Research on Data Representation Method

DL community believe that data featurization will determine the ultimate performance of DL model. Commonly, feature engineering needs to be completed by scholars specializing in the research field, because it requires researchers to have a deep knowledge of the comprehensive meaning explained by the database. For DL, this condition is loosened, but for optimal performance, part of this research will be attributed to this part.

Traditional descriptor method and Graph Neural Network (GNN) method (shown in Figure 6) is potential candidates. Based on them, this dissertation shall investigate the possible featurization method while taking NbO crystal lattice as an example.

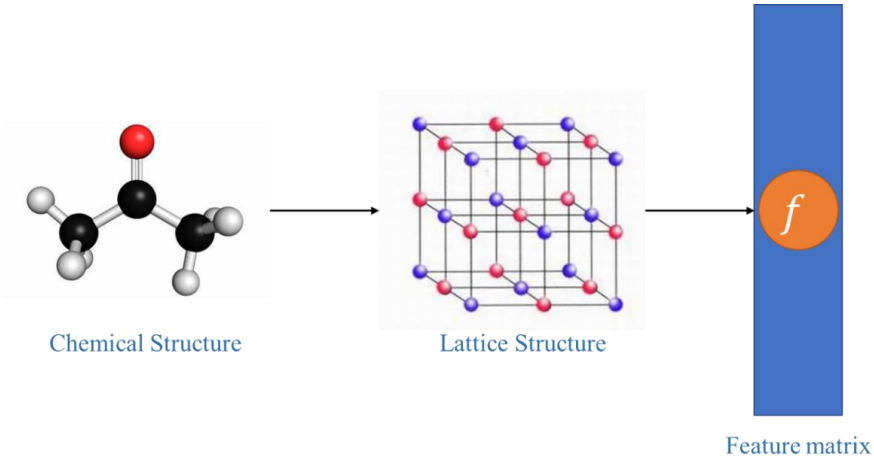

Figure 6. Illustration of featurization network

3.3 Research on Choice of Model

The prediction accuracy depends on the proper choice of model. The general model used in this research is GAN, the model for generator and discriminator is to be explored. Other potential candidate neural-network schemes include Convolutional Neural Network (CNN) and Artificial Neural Network (ANN)[18].

This dissertation shall investigate the suitable GAN model that could bring the best performance from the aspect of stability, accuracy and speed.

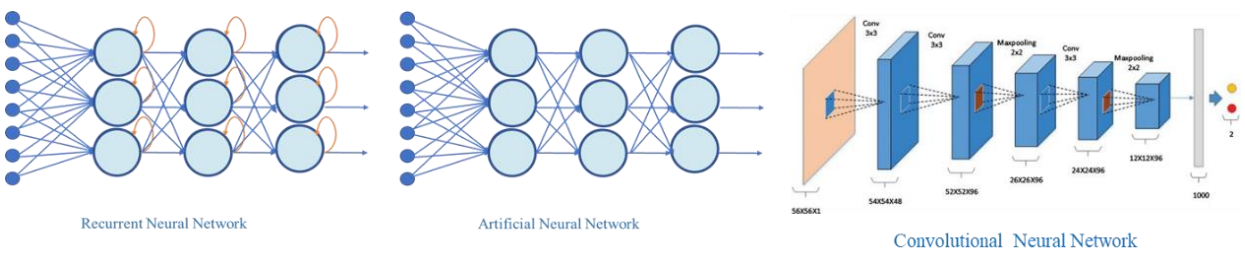

Figure 7. Comparison of CNN, RNN and ANN

3.4 Research on Algorithms for Model Optimization

Hyperparameters of DL model is the key to its performance. The error between the predicted result and the real result is rooted in the balance between the model's bias and variance, apart from the irreducible system error. This dissertation will focus on model bias.

The bias of the model itself is caused by incorrect assumptions of the model, algorithm or by the absence of important interaction relationships. This dissertation shall adjust topology structure and hyperparameters of GAN for optimization.

To be more specific, application on unseen test data and cross-validation will further be conducted to verify the optimization performance. Optimization method such as Grid Search (GS) algorithm, Random Search (RS) algorithm (shown in Figure 8[19,20]) and Network Architecture Search (NAS) algorithm are potential candidates.

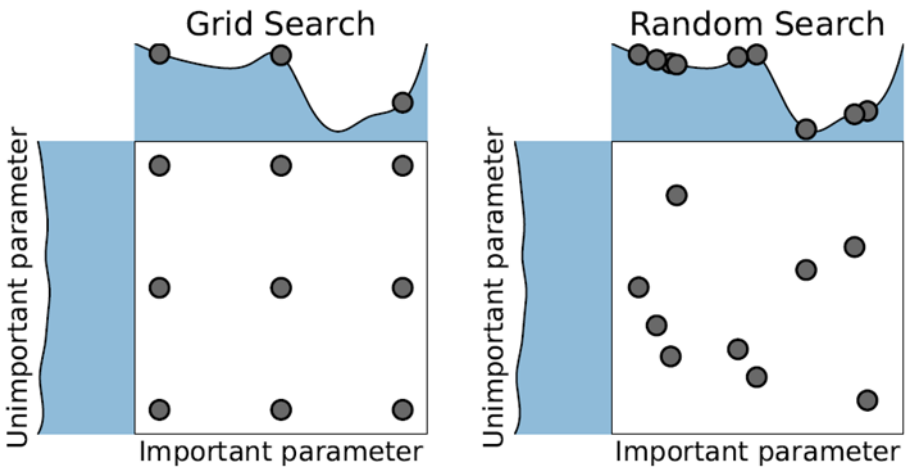

Figure 8. Illustration of grid search and random search

3.5 Testification on NbO System

To quantitatively evaluate the performance of proposed model, this dissertation intends to take the self-built NbO dataset as example to test the performance of various proposed methods. Meanwhile, dataset should be large and precise enough to obtain the optimal performance of deep learning model.

This dissertation is planned to use AIMD to generate serials of data and find out meta-stable structure[17]. As for the verification of generated data, the phonon spectrum and atomic force criteria will be employed (shown in Figure 5). By comparing the prediction cost and speed of DFT and proposed method, the relevant performance indexes for the proposed model will be

obtained, which is expected to prove the importance of GAN model for accelerating new material discovery. To calculate data of NbO system, crystal structure database of NbO system will be studied via open database and literatures.

4. Possible difficulties and problems, possible solutions, and the objectives of the research

4.1 Establishment of training dataset

The performance of deep learning model depends on the training data size. Therefore, the self-built database needs to include accurate and sufficient data. How to obtain a large number of accurate data is the key issue for this research.

The open database is going to include the DFT data of NbO system on different structural configurations both at and far from the ground states. The energy data of NbO system from stable to metastable state is derived by molecular dynamic simulation. To assure the accuracy of the model, cross validation will be used on the database. Further schemes to validate the database will be used, such as by verifying the continuity of input variables such as temperature and doping concentration of energy or other structural properties, by verifying the calculated properties with standard database, or by verifying obtained data with existing literature.

Besides, for massive data processing, appropriate data structure, hardware and software framework will be carefully considered.

4.2 Featurization of crystal lattice

Featurization is the key to ultimate data performance. Thus, it is necessary to establish a crystal lattice characteristic method satisfying the following requirements:

1. reference independent
2. degeneracy
3. smoothness
4. transformability.

Besides, it must represent the energy and atomic force properties of system.

4.3 Suitable GAN model and analysis on the reason

Firstly, the reason why detection task employs CNN model is that the hieratical structure of CNN fits the hieratical structure of image element. Similarly, RNN with sequential feature are employed for NLP tasks. The appropriate network architecture is the most efficient way to solve

certain problem.

Secondly, with the network structure deepens, precision will increase. However, when it reaches a certain threshold, the precision decrease instead of continuously increase, while it is a paradox that the huge amount of data requires a deeper and wider network structure.

This dissertation shall properly establish hyperparameters and network architecture. To better fit binary crystal structure data, this dissertation shall investigate the appropriate network structure and hyperparameters to assure the robustness, speed and accuracy when making material predictions. Secondly, for network which have tens of deep layers, normalize the data at the input layer and the middle layer will greatly help to ensure the convergence in the process of back propagation. When it comes to deeper neural structure, shortcut connection (shown in Figure 9 [21]) is a potential solution in case of avoiding the overfitting of deep network. Detailed calculation with this improved scheme is going to be carried out in the dissertation.

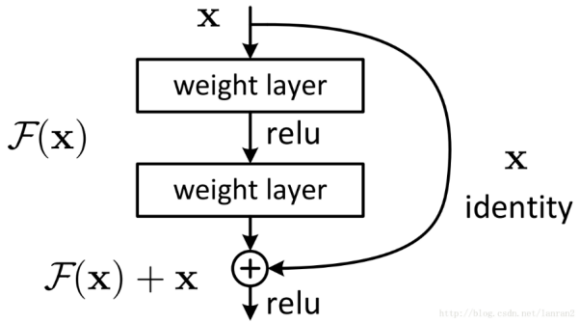

Figure 9. Illustration of shortcut connection

5. Research schedule

Time	Task
1/1/2021- 15/1/2021	Finish the feasibility research of proposed method and find out the potential problems; finish proposal report and send to both supervisors for feedback
16/1/2021- 31/1/2021	Collect required data for network training; design and develop the framework of GAN model for crystal structure prediction
1/2/2021- 14/2/2021	Collect required data for network training; test the performance of GAN model; analyze the calculation results
15/2/2021 – 28/2/2021	Optimize selected GAN model; analysis collected data

1/3/2021- 15/3/2021	Prepare for mid-term report; backup time for potential difficulties
16/3/2021 – 31/3/2021	Finish the writing up of mid-term report and send to both supervisors for feedback; work on plots and demonstrating of the calculation results
1/4/2021- 15/4/2021	Preparing for mid-term inspection; preparing for final dissertation
16/4/2021 – 30/4/2021	Modify and improve the dissertation report; send dissertation paper to both supervisors and improve with feedbacks
1/5/2021- 15/5/2021	Check the final version of dissertation
16/5/2021- 31/5/2021	Prepare for the presentation of the dissertation

6. References

1. Agrawal, A. & Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Commun.* **9**, 779–792 (2019).
2. Yan, L., Alfredo, C., Mark, G. & Zeming, L. Deep Learning. *NYU CENTER FOR DATA SCIENCE* https://atcold.github.io/pytorch-Deep-Learning/ (2020).
3. Ryczko, K., Strubbe, D. A. & Tamblyn, I. Deep learning and density-functional theory. *Phys. Rev. A* **100**, (2019).
4. Ye, W., Chen, C., Wang, Z., Chu, I. H. & Ong, S. P. Deep neural networks for accurate predictions of garnet stability. *Nat. Commun.* 1–18 (2018).
5. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
6. Karamanis, P. The Importance of the DFT method on the computation of the second hyperpolarizability of semiconductor clusters of increasing size: A critical analysis on prolate aluminum phosphide clusters. *International Journal of Quantum Chemistry* vol. 112 (John Wiley & Sons, Ltd, 2012).
7. Jha, D. *et al.* ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **8**, 1–14 (2018).
8. Artrith, N. & Kolpak, A. M. Understanding the Composition and Activity of Electrocatalytic Nanoalloys in Aqueous Solvents: A Combination of DFT and Accurate

Neural Network Potentials. *Nano Lett* **14**, 2670–2676 (2014).

9. Huang, L. & Ling, C. Practicing deep learning in materials science: An evaluation for predicting the formation energies. *J. Appl. Phys.* **128**, (2020).
10. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, (2018).
11. Ye, W., Chen, C., Wang, Z., Chu, I. H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **9**, 1–6 (2018).
12. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
13. Latest Trends on Computer Vision Market – InData Labs Blog. https://indatalabs.com/blog/trends-computer-vision-software-market?cli_action=1555888112.716.
14. Zhu, X., Vondrick, C., Fowlkes, C. C. & Ramanan, D. Do We Need More Training Data? *Int. J. Comput. Vis.* **119**, 76–92 (2016).
15. Hestness, J. *et al.* Deep learning scaling is predictable, empirically. *arXiv* 1–19 (2017).
16. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, (2015).
17. Georg, K., Doris, V. & Martijn, M. Vienna Ab initio Simulation Package. (2017).
18. Lu, M. CNN vs.RNN vs.ANN - Analysis of Three Deep Leaning Networks. *Zhihu.com* https://zhuanlan.zhihu.com/p/107993566 (2020).
19. Hoogle. Introduction to Black Box Optimization. *Zhihu.com* https://zhuanlan.zhihu.com/p/66312442 (2019).
20. Da, B. Deep Learning No.10: Optimization of Hyperparameters. *Zhihu.com* https://zhuanlan.zhihu.com/p/69025104 (2018).
21. Lanren, Y. ResNet Analysis. *CSDN blog* https://blog.csdn.net/lanran2/article/details/79057994 (2018).