Convolutional Neural Networks: Estimating Relations in the Ising Model on Overfitting

Andrei Gavrilov
Bachelor of Applied Science
University of British Columbia
Vancouver, BC, Canada
andrei-gavrilov@outlook.com

Alex Jordache
Bachelor of Commerce
University of British Columbia
Vancouver, BC, Canada
alexonej@gmail.com

Maya Vasdani
Bachelor of Arts
University of British Columbia
Vancouver, BC, Canada
mvasdani@bell.net

Jack Deng
Bachelor of Applied Science
University of British Columbia
Vancouver, BC, Canada
jackdeng000@gmail.com

Abstract — A CNN (Convolutional Neural Network) is one of actively researched and broadly applied deep machine learning methods. A CNN is composed of a feed-forward neural network that takes in images as inputs, and outputs a probability value associated to a class that best describes the image. As well, it is constructed of multiple layers, which include convolutional, max-pooling and fully connected layers. However, the training set has a large influence on the accuracy of a network, and hence it is paramount to create a network architecture that prevents overfitting (when a trained model cannot differentiate newly input data from its test data) and underfitting (the inability of a model to find relationships among inputs). This paper addresses the above deficiencies by comparing the statistics of CNN image recognition algorithms to the Ising model. The Ising model consists of magnetic dipole moments that can be in one of two states: +1 or -1. Using a two-dimensional square-lattice array once a training set of such data is complete, we determine the impact that network parameters, specifically learning rate and regularization rate, have on the adaptability of convolutional neural networks for image classification. Our results not only contribute to a better theoretical understanding of a CNN, but also provide concrete guidance on preventing model overfitting and underfitting when a CNN is applied for image recognition.

Keywords—Machine Learning, Convolutional Neural Network, Image Processing, Ising Model, Data Overfitting, Learning Rate, Regularization Rate, Cognitive Systems

I. Introduction

In our modern world, machine learning has taken a rise as a form of artificial intelligence where computers can discover patterns and even learn from collected data to make intelligent autonomous decisions. A CNN is a class of deep learning neural networks that can be applied to various image recognition tasks, such as identifying a human in a surveillance video, recommending specific products for consumers, observing interesting weather phenomena, or identifying key chemical structure in various drug development [1] [2].

Data set training has a large influence on the accuracy of a network, and hence it is paramount to create a network architecture that prevents overfitting and underfitting [3]. By randomly trimming the data during training and providing more robust data sets, the network can become less reliant on similar pieces of data. Hence, this would improve its overall capability for higher accuracy image recognition.

This paper investigates the problem of overfitting and underfitting, comparing the statistics of CNN image recognition algorithms with the Ising model. We propose to use the analogy to the square-lattice array, in order to determine how the statistical mechanics of the Ising model phase transition are used in tuning parameters of the CNN. Once a training set of such data is complete, we explore the convolutional layers that propagate the neuron values, similar to the propagation of spins. The experimental results conducted on CIFAR image database demonstrate that a low regularization rate and learning rate yields overfit data, high regularization rate yields partially fit data, and high learning rate yields unfit data. Thus, this research provides insights that can assist in preventing model overfitting and underfitting when CNN is utilized in various image recognition applications. The conclusions section closes this chapter.

II. Literature Review

Recently, there has been a surge in interest in machine-learning methods for image recognition. They have been successfully applied to various intelligent systems in the fields of computer graphics, robotics, knowledge representation, virtual reality, situation awareness, decision-support systems, and many others [1] [2]. The introduction of CNN created excitement in image processing research community, with new opportunities to significantly increase image identification rate with a fraction of computational resources, thus making the recognition process more accurate and less expensive.

A CNN can learn the set of features important for image recognition that typically were hand-picked in traditional sequential algorithms [4]. There have been numerous examples of CNNs performing very well for a variety of applied problems [5]. For example, they have been successfully used in engineering domains for fuzzy-logic control and decision

Proc. 2018 IEEE 17th Int'l Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC'18)
Y. Wang, S. Kwong, J. Feldman, N. Howard, P. Sheu & B. Widrow (Eds.)
978-1-5386-3360-1/18/$31.00 ©2018 IEEE

systems [6]. The very recent overview of their applications and open questions can be perhaps found in 2017 research article by McCann et.al. [7]. In their state-of-the-art paper, they describe a gamut of uses of CNNs to solve both direct and inverse problems in imaging. They point out that once it became possible to train deep CNNs on large databases of images, this approach demonstrated exceptional performance for image classification and segmentation tasks. Recently, researchers started apply CNNs to inverse problems such as denoising, deconvolution, super-resolution, and medical image reconstruction, which quickly outperformed traditional image processing, statistics, bag-of-words, and matrix-based methods. An interesting observation is that the authors of this work state that some of the most critical design decisions questions are where the training data comes from, and what the architecture for optimal CNN performance is [7]. Our research directly answers the question of what the most efficient mechanism to prevent data overfitting and underfitting during CNN training is through insights learnt in the Ising model.

To summarize, this paper is one of the first studies linking the structure of the Ising model to the CNN entropy loss concept, with the goal of finding out optimal paraments to prevent data overfitting and underfitting during training phase. The methodology developed will be presented in the next section.

III. METHODOLOGY

A. Convolution Neural Network

A CNN is a type of deep machine learning method, normally represented as a feed-forward artificial neural network that can learn the set of features important for image recognition that typically were hand-picked in traditional sequential algorithms. A CNN consists of neurons which have various weights and biases. The network assumes that the inputs are images from the CIFAR database and expresses a single output score from the input raw image pixels on one end to class scores on the other end [4]. A typical convolutional neural network consists of a sequence of layers. Every layer transforms an input 3D volume and transforms it to an output 3D volume through a differentiable function. There are mainly three types of layers in a convolutional neural network: convolutional layers, pooling layers, and fully-connected layers [4]. One way in which the recognition accuracy can be improved is by increasing the number of layers. Our network architecture used three distinct layers following the following process: Image selection, convolution, pooling, convolution, pooling, connect images, outcome predictions. However, this also significantly increases the training time of such data sets. In this paper, we consider the effects of learning rate and regularization in the CNN model on its recognition performance. Over time, as the CNN learns the features of the test data set, we find that the recognition rate increases. However, if the model is trained for too long, the performance may decrease as the model becomes overfit and becomes oversaturated from the training dataset.

After the network architecture is applied to each image, our system performs machine learning on the training data sets; hence training our system to recognize various images from the CIFAR database. The system is trained by passing labels with the images to determine which image score outputs correspond which images shown in Fig. 1 [8]. After training is complete, the parameters of the dataset are altered to observe the performance of the trained network on testing data sets, and finally to observe overfitting, underfitting, and analyze statistics of the overall network.

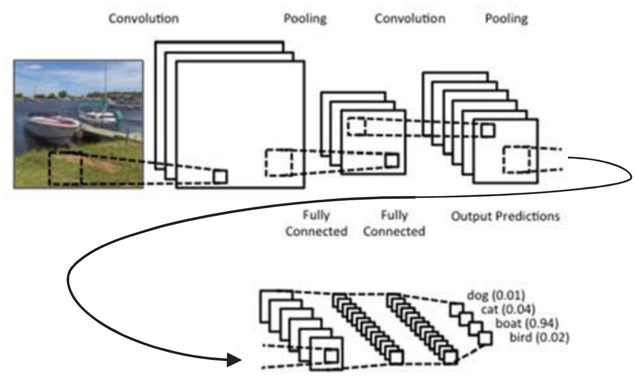

Fig. 1. An example of various layers in a CNN architecture

B. The Loss Function

An observation on which we base the premises of this paper is that an extended network architecture is analogous to physical mechanical concepts of macrostates, microstates and probabilities. In information theory to define an entropy, it is imperative to define an ensemble of states with respective probabilities. This concept is related, if not equivalent to the microstate, where each network is thus its own microstate. A pixel in the filter is just like a particle, and the filter itself at a given moment is just a microstate.

From the information theory, we know that uncertainty can be expressed through the concept of entropy: more underteach corresponds to a higher entropy of the system. Thus, we base our methodology for data overfitting on the concept of Loss, with the goal of CNN is to minimize the loss through using Stochastic Gradient Descent, where TrainLoss(w) is the is the direction that minimizes the loss the most [3]:

$$\nabla_{\mathbf{w}} \mathsf{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\mathsf{train}}|} \sum_{(x,y)\in\mathcal{D}_{\mathsf{train}}} 2(\underbrace{\mathbf{w}\cdot\phi(x) - y}_{\mathsf{prediction-target}})\phi(x)$$

$$l_{reg}(\vec{w}) = \lambda \sum_{i=1}^{N} w_i^2$$

C. The Ising Model

Based on the concepts of entropy and of loss in the information theory, we now propose to compare the statistics of CNN image recognition algorithms to the Ising model [9]. The Ising model consists of magnetic dipole moments that can be in one of two states: +1 or -1. Using a two-dimensional square-lattice array, we investigate how the statistical mechanics of the Ising model phase transition are used in tuning parameters of the CNN to best fit the data.

Consider a set of lattice sites Λ, each with a set of adjacent sites forming a d-dimensional lattice. For each lattice site $k \in \Lambda$ there is a discrete variable σ_k such that $\sigma_k \in \{+1, -1\}$, representing the site's spin. A spin configuration, $\sigma = (\sigma_k)_{k \in \Lambda}$ is an assignment of spin value to each lattice site [10].

For any two adjacent sites $i, j \in \Lambda$ there is an interaction J_{ij}. Also a site $j \in \Lambda$ has an external magnetic field h_j interacting with it. The energy of a configuration σ is given by the Hamiltonian function:

$$H(\sigma) = -\sum_{\langle i\ j \rangle} J_{ij}\sigma_i\sigma_j - \mu\sum_j h_j\sigma_j$$

The first sum is over pairs of adjacent spins (every pair is counted once). The notation $\langle ij \rangle$ indicates that sites i and j are nearest neighbors and that the configuration probability has a Boltzmann distribution [10].

The configuration probabilities $P_\beta(\sigma)$ represent the probability that the system (in equilibrium) is in a state with configuration σ. Ising models can be simplified by assuming that all of the nearest neighbors have the same interaction strength. The Hamiltonian is thus the following equation [10]:

$$H(\sigma) = -J\sum_{\langle i\ j \rangle} \sigma_i\sigma_j.$$

The illustrations of the Ising model and Gradient Descent function are given in Fig. 2 and Fig. 3.

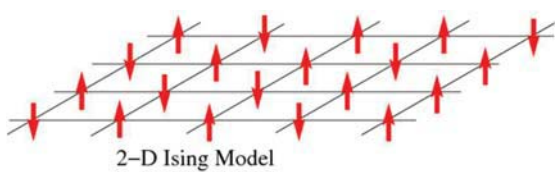

Fig. 2. Illustration of 2-D Ising Model

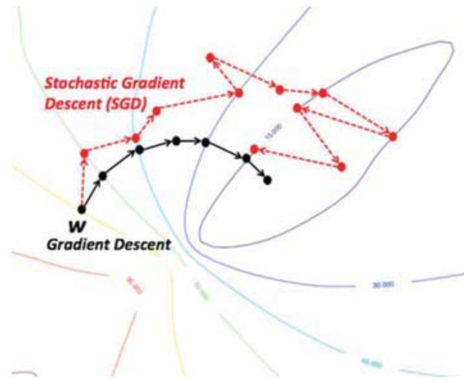

Fig. 3. Stochastic Gradient Descent (SGD)

We note that in the two-dimension problem, the Ising model provides exact solutions; however, in a higher-dimensional case, the Ising model cannot always produce global exact solutions. Thus, the use of Ising model for parameter fitting in CNN is proposed for two-dimensional cases.

IV. Experimental results

The goal of the experiment was to establish which network parameters, specifically learning rate and regularization rate, have the largest impact on adaptability of convolutional neural networks for image classification. In our investigation of the trained model, CNN layers were considered to be a function of the interaction strength of the Ising model. The CIFAR Image Dataset was used for our experiments [4] as shown in Fig. 4.

Fig. 4. CIFAR sample image dataset

We then performed machine learning using the CNN on training data sets, observed the performance of the trained network on testing data sets, recorded the observed overfitting, and analyzed statistics of the network.

The first experiment was run on a small training set ($n_1 = 500$) with a large number of epochs ($n_2 = 60$) as shown in Fig. 5. An epoch value represents a number of times that the training architecture runs through each image. This experiment had a low learning rate (0.01) and a very low regularization rate (0.0001) as the system parameters. Regularization rate is the process of introducing additional information into the system to prevent overfitting, adds inertia to the system, and resists the urge of the network to produce outcomes. In this experiment at the 200 iteration mark, the loss function becomes very close to zero. As the loss function stops decreasing, this means that the training set is returning 100% accurate images. Hence, the training set was overfit, and testing accuracy was low. Overall, training accuracy was 100%, while test accuracy reached only 30.6%. It can be concluded that low test accuracy was due to the CNN being over trained and data overfit.

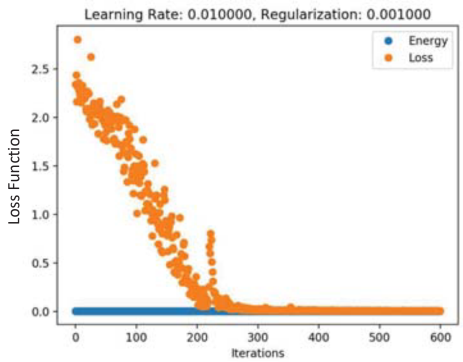

Fig. 5. Low Regularization, Low Learning Rate

In the second experiment shown in Fig. 6, the data size and epoch parameters were kept constant, including the same learning rate of 0.01. The regularization rate was changed from 0.001 to 1.000. In this test, the loss function begins at a high rate, and as the number of iterations increases, the loss function decreases. At 200 iterations, the loss function stabilizes at a lower level, indicating that the network is slowly decreasing the loss function. The training accuracy was 41.4%, much lower than in the first experiment, however test accuracy achieved 25.4%. We can conclude that the network has not been overtrained, as the training test accuracy did not reach 100% saturation. Thus, the high regularization rate yields a low training and test accuracy.

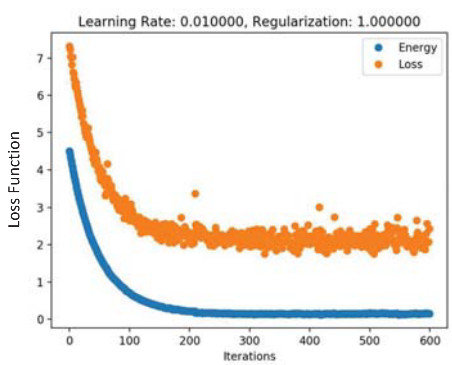

Fig. 6. High Regularization, Low Learning Rate

In the final experiment shown in Fig. 7, the data size and epoch parameters were kept constant. The learning rate was increased to 0.100, and the regularization rate was decreased to 0.100. In this test, the loss function begins at a high rate, and as the number of iterations increases, the loss function decreases. At 200 iterations, the loss function stabilizes at a lower level, as in the previous experiment. As can be seen below, there is little variation in the test results as a combination of the low regularization and learning rates. The training accuracy was ~10% and the testing accuracy was also ~10%. We can conclude that the network has not been overtrained, but yields low results as a low regularization rate and learning rate yields a low training and test accuracy.

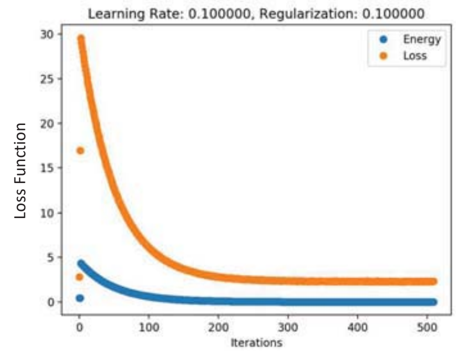

Fig. 7. Low Regularization Rate

We can summarize that low a regularization rate and learning rate yields overfit data (100% training accuracy), high regularization rate yields partially fit data where training accuracy is approximately the same as the testing accuracy, and that a high learning rate yields unfit data with low training and testing accuracies.

V. CONCLUSIONS

This paper presented a study on how the statistical mechanics of the Ising model phase transition are used in tuning parameters of the CNN to best fit the data. Once a training set of such data is complete, we explore the convolutional layers that propagate the neuron values, similar to the propagation of spins. The experimental results conducted on CIFAR image database demonstrate that low regularization rate and learning rate yields overfit data, high regularization rate yields partially fit data, and high learning rate yields unfit data. Overall, we find that properly tuning the learning and regularization rate parameters will prevent the model from overfitting or underfitting the data. These parameters can also be used on their own or in combination with other mechanisms to find the best fit: resampling methods and the use of validation datasets in image recognition applications. Future work will focus on experimenting with larger data set, various CNN architectures, and different system parameters.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Aaron Palmer of the UBC Department of Mathematics for his time, support, and advice obtained during this project.

REFERENCES

[1] Y Wang, N. Howard, J. Karcprzyk, P. Frieder, P. Sheu, R. Fiorini, Marina L. Gavrilova, S. Patel, J. Peng, B. Widrow Cognitive Informatics: Towards Cognitive Machine Learning and Autonomous Knowledge Manipulation, International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) vol 12, no 1, pp1-18, 2017.

[2] M. Gavrilova, Y. Wang, F. Ahmed, P.P. Paul KINECT Sensor Gesture and Activity Recognition for Consumer Cognitive Systems, IEEE Consumer Electronics Magazine, Special Issue on Consumer Electronics, vol 4, pp 88-96, 2017.

[3] L. Faussett, Fundamentals of Neural Networks: Architectures, Algorithms and Applications, 1st edition, Pearson, 2004.

[4] A. Krizhevsky, I. Sutskever, and G.E. Hinton ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS), 2012.

[5] Wang, Y., & Tian, Y. A Formal Knowledge Retrieval System for Cognitive Computers and Cognitive Robotics. International Journal of Software Science and Computational Intelligence (IJSSCI), 5(2), 37-57, 2013.

[6] C-T Lin and C.S. G. Lee Neural Network Based Fuzzy-Logic Control and Decision Systems, IEEE Transactions on Computer, vol 40, no 12, 1991.

[7] M.T. McCann, K. H. Jin, M. Unser A Review of Convolutional Neural Networks for Inverse Problems in Imaging, Image and Video Processing, Computer Vision and pattern Recognition, eess.IV, arXiv: 1710.04011, 2017.

[8] CNN Network Architecture Image Reference:
https://www.clarifai.com/static/images/cnn.png.

[9] E. Ising, Z. Physik 31, 253, 1925.

[10] Roy J. Glauber Time-Dependent Statistics of the Ising Model Journal of Mathematical Physics 4, 294, 1963.