Reading 1

(1) Title, author and where it was published

- Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
- Authors: Sergey Ioffe, Christian Szegedy
- Published: ArXiv

(2) Summary of the content

This paper introduces a machine learning technique called batch normalization which could alleviate the internal covariate shift. In the next section, the paper then discusses some details about its implementation, like adding parameters to restore the representation power of each layer's input, and the specific batch normalization algorithm which is basically scaling and shifting the normalization based on the mean and variance of the minibatchs. Later, this paper goes over some of the advantages of batch normalization, which include generalizability over all activation functions, applicability on fully-connected and convolutional layers, and allowance for high learning rates. In addition, the paper provides an experiment examining the ability of Batch Normalization in combating with internal covariate shift by comparing the test accuracy and evolution of input distribution using the MNIST dataset, and concluded that Batch normalization indeed helps achieving higher accuracy, making more stable input distribution and reducing the internal covariate shift. Last but not least, the authors also compare the Batch normalization variants on the Inception network and proved that Batch normalization could actually accelerate the training process of deep learning networks.

(3) Strengths of the paper

First of all, this paper illustrates the concepts and mechanism of batch normalization in a comprehensible way. The paper first explains the definition of internal covariate shift and explains why it is problematic to the training process. Internal covariate shift happens when the choice of parameters in the preceding layers changes the distribution of the inputs to the current layer, and one needs to carefully tune the parameters in each layer to solve this problem, causing the learning process to be time-consuming and ineffective. Therefore, with normalization on input data prior to learning by fixing the means and variances of the inputs, the authors claim that the internal covariate shift issue could be effectively reduced.

Moreover, this paper is convincing in that it provides enough data to support the claim that batch normalization is effective in enhancing the training process for deep learning networks. It not only demonstrates its ability on simple networks with three fully-connected hidden layers and nonlinear activation functions but also validates that it could also be effectively applied on convolutional networks to perform more complicated tasks like image classification. The authors clearly present that batch-normalization-modified Inception network achieves accuracy of 72.2% using much fewer steps, reducing from 31×10^6 to 2.1×10^6. And we could see that higher learning rate does leads to faster training while preserving the accuracy of the network, again validating that batch normalization is an useful technique for deep learning algorithms.

(4) Major critiques

Batch normalization is helpful in reducing internal covariate shift by making the loss surface smoother, but apparently it does not solve this issue completely. Therefore, future research is needed in investigating how to address internal covariate shifts more thoroughly.

Reading 2

(1) Title, author and where it was published

- Title: Dropout: A Simple Way to Prevent Neural Networks from Overfitting
- Authors: Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sustskever, Ruslan Salakhutdinov
- Published: University of Toronto

(2) Summary of the content

This paper introduces a powerful and predominant machine learning optimization technique called dropout that aims to solve the overfitting issues in deep neural networks. It is common for a training dataset to have noises, and it is computationally very expensive to combine large neural networks. Therefore, it is necessary to randomly drop out some units during training time. This paper discusses the motivations and previous works regarding dropout, and then it provides the definition and algorithms of dropout. Later, it presents several experiments comparing the performance of deep learning tasks with and without dropout, and explains in detail about the properties behind. Last but not least, the paper introduces the dropout Restricted Boltzmann Machines (RBM) and compares it with the standard RBM, claiming that dropout improved the performance in various aspects.

(3) Strengths of the paper

Combining graphs and mathematical formulas, this paper explains the mechanism of dropout in a comprehensible way. The formulas define dropout in each layer to be a normal node despite a Bernuolli random variable with an input probability p that determines whether this node would be included in the training time. In addition, it also mentions a procedure for training dropout neural nets, which includes backpropagation and unsupervised pre-training. Dropout could be applied to finetune nets for pre-training the neural networks accompanied with a small learning rate to achieve a better performance.
Furthermore, the experimental results provide the readers with strong results that support its argument. Dropout has been tested on a wide range of datasets that include handwritten digits, speech recognition, image classification, newswire articles, and genes information, and it has been shown that dropout leads to better performance on all of these scenarios. For example, for the MNIST dataset, Dropout NN with Logistic activation has error (1.35%) smaller than that of Standard neural net (1.60%), and Deep Boltzmann Machines gives even better performance.
After proving that dropout is indeed effective, the paper then goes on to discuss specific features that dropout has. First of all, dropout help prevents co-adaptation that might make the hidden units unreliable. In addition, dropout rate p is also an important hyperparameter that needs to be tuned carefully. The authors examine the effect of dropout rate by comparing the classification error when the number of hidden units is fixed with when the number of hidden units varies with probability p. It turns out that $p = 0.6$ appears to be the best value for this situation, but the authors also claim that 0.5 is a generally optimal result.

(4) Major critiques

Dropout is an effective technique for preventing overfitting, but it would increase the training time since there is an extra layer of computation. Further studies should be focused on how to incorporate dropout with other regularization and optimization techniques so that they will lead to better performance together.

Reading 3

(1) Title, author and where it was published

- Title: Effective and Inconspicuous Over-the-air Adversarial Examples With Adaptive Filtering
- Authors: Patrick O'Reilly, Pranjal Awasthi, Aravindan Vijayaraghavan, Bryan Pardo
- Published: Northwestern University, Google Research

(2) Summary of the content

This paper introduces a new technique called adaptive filtering that aims to improve the robustness of deep neural networks in the audio domain by providing many interpretable and differential parametric transformations on audios as adversarial examples in a simulated over-the-air setting. This approach is motivated by the fact that the current deep neural networks are vulnerable to artificially generated noises and perturbations that might cause the network to make incorrect predictions, and thus the researchers must test the network's ability against such perturbations in audio recognition learning tasks by using effective attacks. Then, the authors discuss in detail about the basic setup, algorithm, objective function, adversarial and auxiliary loss functions for adaptive filtering and the simulation setting. Over-the-air attack is defined as the kinds of audio that are flawed and are captured by microphone before entering the victim model, and the authors propose that they could make differentiable distortions through adaptive filtering which could dynamically shape the frequency content of audio. Later, the paper went over several experiments regarding the effectiveness of this attack method over the state-of-the-art speaker verification model, and concluded that the proposed attack is less conspicuous than the baseline model.

(3) Strengths of the paper

The authors explain the mechanisms of adaptive filtering in a very detailed fashion. It specifies how to process the audio input through a parametrized finite impulse response (FIR) filter. The objective function is defined as the set of parameters that could minimize the combination of adversarial loss and auxiliary loss. The adversarial loss measures the success of the attack in achieving the outcome using the neural network model with adaptively filtered audio input as data, and the auxiliary loss is defined as how much it resembles the benign audio x.

Moreover, this paper also employs convincing experimental results that demonstrate the effectiveness of adaptive filtering in attacking deep neural networks. This experiment consists of generated and baseline attacks applied to a speaker verification model, and tested the effectiveness of these attacks through a perceptual study that let the participants rate the conspicuousness of each attack. The results show that the baseline attack is almost two times more conspicuous than the proposed attack, suggesting that filtering can significantly affect attack success and conspicuousness.

(4) Major critiques

This paper mentions several integration with other techniques or methods yet fails to explain such choices, making the reading less comprehensible over all. For example, it mentions that the expectation-over-transformation approach could help produce more robust adversarial examples in the over-the-air setting, but it remains a little ambiguous as to how it could improve its performance.

Furthermore, it looks like the experiments are carefully planned and there are lots of constraints about the settings, while real-world audio learning tasks might face more noises and uncertainties, so the applicability of this approach needs to be further tested on a wider range of environments.