Performance Enhancement

of

CKKS Encrypted Neural Networks During Training

Samuel Atkins – 1002951754

Thesis Supervisor: Prof. Glenn Gulak

January 27, 2020

Table of Contents

[Introduction 1](#_Toc31061036)

[Literature Review 2](#_Toc31061037)

[Encryption 2](#_Toc31061038)

[Public-Key Cryptography 2](#_Toc31061039)

[Homomorphic Encryption 2](#_Toc31061040)

[Fully Homomorphic Encryption 2](#_Toc31061041)

[Artificial Intelligence 3](#_Toc31061042)

[Machine Learning Categories 3](#_Toc31061043)

[Neural Network Training and Forward Propagation 4](#_Toc31061044)

[Backpropagation and Hyperparameter Tuning 5](#_Toc31061045)

[Convolutional Neural Networks 6](#_Toc31061046)

[Modern CNN Architectures 7](#_Toc31061047)

[Objectives and Metrics 9](#_Toc31061048)

[Progress to Date 11](#_Toc31061049)

[Future Work 13](#_Toc31061050)

[References 15](#_Toc31061051)

[Appendix A: Public-Key Encryption 19](#_Toc31061052)

[Appendix B: Forward Propagation 21](#_Toc31061053)

[Appendix C: Activation Functions 25](#_Toc31061054)

[Sigmoid 25](#_Toc31061055)

[Tanh 25](#_Toc31061056)

[Arctan 26](#_Toc31061057)

[ReLU 26](#_Toc31061058)

[Leaky ReLU 27](#_Toc31061059)

[Appendix D: Loss Functions for Binary and Multi-Class Classification 29](#_Toc31061060)

[Appendix E: Loss Curves and Neural Network Training 31](#_Toc31061061)

[Appendix F: Linear Image Filtering and Convolutions 32](#_Toc31061062)

[Appendix G: Types of Pooling Layers 34](#_Toc31061063)

[References 36](#_Toc31061064)

Table of Figures

[Figure 1: Structure of a multi-layer neural network [12] 4](#_Toc31061107)

[Figure 2: Illustration of the structure of a CNN [14] 6](#_Toc31061108)

[Figure 3: LeNet5 architecture [15] 7](#_Toc31061109)

[Figure 4: AlexNet architecture [16] 8](#_Toc31061110)

[Figure 5: An illustration of the GoogLeNet inception module [18] 9](#_Toc31061111)

[Figure 6: An illustration of the general structure of a candidate model 11](#_Toc31061112)

[Figure 7: A graph [29] of the x-squared activation function 11](#_Toc31061113)

[Figure 8: Gantt chart detailing a procedural timeline leading up to the final evaluations 13](#_Toc31061114)

[Figure 9: Alice sends Bob a public key (red), keeping her private key (green) to herself 19](#_Toc31061115)

[Figure 10: Bob uses Alice’s public key to encrypt a message 19](#_Toc31061116)

[Figure 11: Bob sends his encrypted message to Alice 19](#_Toc31061117)

[Figure 12: Alice uses her private key to decrypt Bob’s encrypted message 20](#_Toc31061118)

[Figure 13: Alice can now view Bob’s message 20](#_Toc31061119)

[Figure 14: An example of a fully-connected neural network prior to forward propagation 21](#_Toc31061120)

[Figure 15: Illustration of the forward propagation process prior to the application of the activation function 22](#_Toc31061121)

[Figure 16: The forward propagation process after applying the activation function to the nodes in the hidden layer 23](#_Toc31061122)

[Figure 17: An illustration of the output of the neural network after forward propagation has occurred 24](#_Toc31061123)

[Figure 18: A graph [29] of the Sigmoid activation function 25](#_Toc31061124)

[Figure 19: A graph [29] of the tanh activation function 25](#_Toc31061125)

[Figure 20: A graph [29] of the arctan activation function 26](#_Toc31061126)

[Figure 21: A graph [29] of the ReLU activation function 27](#_Toc31061127)

[Figure 22: A graph [29] of the Leaky ReLU function 27](#_Toc31061128)

[Figure 23: A graph [29] of the negative natural logarithm, y = -ln(x) 29](#_Toc31061129)

[Figure 24: A typical graph of the validation and training loss of a neural network as a function of the update iterations 31](#_Toc31061130)

[Figure 25: Grayscale image of Jenga blocks 32](#_Toc31061131)

[Figure 26: Illustration of the 3x3 right Sobel filter 32](#_Toc31061132)

[Figure 27: Result of the application of the right Sobel filter 33](#_Toc31061133)

[Figure 28: 4x4 matrix of integers representing the input to a pooling layer 34](#_Toc31061134)

[Figure 29: Result of applying average pooling to the input matrix in Figure 25 34](#_Toc31061135)

[Figure 30: Result of applying max-pooling to the input matrix of Figure 25 35](#_Toc31061136)

[Figure 31: Result of applying global average pooling to the input matrix 35](#_Toc31061137)

# Introduction

In 2009, Gentry utilized lattice-based cryptography to construct the first fully homomorphic encryption (FHE) scheme [1]. Fully homomorphic encryption is named as such because it allows one to perform an arbitrary sequence of additions and multiplications on cipher texts. Given the security requirements of private data processing, FHE was a tremendous breakthrough for machine learning. A neural network modified to operate on homomorphically encrypted data is called a CryptoNet. CryptoNets are attractive because they allow a data owner to present their data, already encrypted, to a cloud service. This cloud service can then apply their CryptoNet to the encrypted data to provide encrypted inferences. These encrypted inferences can be sent back to the data owner to be decrypted. During this entire process, the cloud service remains unaware of the content of the data because the data owner maintains the private key.

In 2016, Microsoft Research developed a high-throughput CryptoNet capable of achieving 99% accuracy and 51,000 inferences per hour on a single PC [2]. Despite Microsoft’s strong per hour throughput on the MNIST dataset, applications that have larger and more complex datasets may yield fewer inferences per unit time. There is significant room for improvement in terms of latency and throughput. These improvements will be valuable because they will increase the applicability of CryptoNets in a manner proportional to the magnitude of the efficiency improvements.

# Literature Review

## Encryption

### Public-Key Cryptography

Public-key cryptography uses key-pairs to encrypt and decrypt data. Plaintext is unencrypted, readable text. Ciphertext, conversely, is encrypted text. To encrypt plaintext into ciphertext using public-key encryption, one first must generate a public key and a private key. The public key is used to encrypt plaintext into ciphertext, and the private key is used to decrypt the ciphertext back into plaintext. Public-key cryptography allows parties to communicate private information through public mediums. For a visualization of this process, see Appendix A.

### Homomorphic Encryption

Homomorphic encryption is a form of encryption that supports arithmetic computation on ciphertexts. Using a homomorphic encryption scheme allows one to perform mathematical operations on ciphertexts. One can, using homomorphic encryption, create two ciphertexts, multiply these two ciphertexts together, and then decrypt the encrypted result to obtain the product of the two original plaintexts.

There are three different types of homomorphic encryption: partially homomorphic encryption (PHE), somewhat homomorphic encryption (SHE), and fully homomorphic encryption (FHE). PHE allows a smaller set of mathematical functions to be performed on ciphertexts an arbitrary number of times. SHE supports a limited number of operations that can only be performed a finite number of times. The most desirable of the three, FHE, allows one to perform both addition and multiplication on ciphertexts a semi-infinite number of times.

### Fully Homomorphic Encryption

In 2009, Gentry introduced a limited SHE scheme [1]. The SHE scheme that he introduced was limited because after each multiplication operation between two ciphertexts, the noise budget corresponding to the result would grow. Ultimately, the noise budget would grow too large and render the decrypted result inaccurate. Gentry elucidates, however, that a modified version of the proposed SHE scheme is *bootstrappable*. An encryption scheme is bootstrappable if the scheme can homomorphically evaluate its own decryption circuit plus one additional NAND gate. Gentry then illustrates that any SHE scheme that is bootstrappable can be recursively self-embedded to make it fully homomorphic. This discovery outlined the basis for the first FHE scheme.

The security of Gentry’s FHE scheme was based on ideal lattices. The implementation of Gentry’s lattice-based scheme showed that secure bootstrapping took up to 30 minutes per operation [3]. Following this implementation was an integer-based approach [4] that was simpler and more efficient than the previous lattice-based approach. Despite the efficiency improvement, the uses of homomorphic encryption were severely limited by computational efficiency bottlenecks due to the large number of mathematical operations involved. From 2011 to 2016, many variations of FHE schemes followed Gentry’s breakthrough. These encryption schemes include the BGV [5], LTV [6], BFV [7] [8], BLLN [9], and CKKS [10] schemes.

The scheme that we will focus on for the purpose of this thesis is the CKKS scheme. This scheme is significantly faster than the other schemes because it utilizes a rescaling procedure. During rescaling, the ciphertext is truncated into a smaller modulus. As a result, the decrypted result is rounded and less accurate. This thesis will focus on speed (i.e., problem instance throughput) and will take advantage of approximate computing to enhance computational speed.

## Artificial Intelligence

### Machine Learning Categories

There are two general categories of machine learning: supervised and unsupervised learning. Supervised learning requires labelled training data to improve the performance of a model through backpropagation. The supervised learning problem of particular interest is multi-class classification, specifically in the context of image classification. Convolutional neural networks are comprised of a convolutional segment and a fully-connected segment. These types of neural networks are the most effective networks for multi-class image classification.

### Neural Network Training and Forward Propagation

An artificial neural network, also known as a neural network, or multi-layer perceptron, is a circuit of neurons connected in a network. Neural networks are used to classify data and perform data driven tasks. Unlike traditional classification engines, neural networks learn from their input data to produce stronger results over time. The connections between the neurons in a neural network are known as the weights of the neural network. These weights are malleable and define the state of the network. As the neural network learns from its input data, the weights of the network change. The learning process in which the network learns from its input data is known as the training phase.

Prior to the training phase, the dataset must be split into three partitions: the training dataset, the validation dataset, and the testing, or holdout, dataset. Certain data splitting strategies lead to stronger generalization. The Kennard-Stone algorithm and SPXY systematic sampling method are two examples of algorithms that attempt to maximize the generalizability of the network through effective data splitting [11]. During training, we feed the data entries from the training dataset into the neural network. The neural network tries to infer the labels attached to the training data. It then uses the feedback from its correct and incorrect inferences to gradually update its inner parameters.

The structure of a neural network is one of the factors that dictates the post-training performance of a neural network. Figure 1 illustrates the structure of a simple neural network:

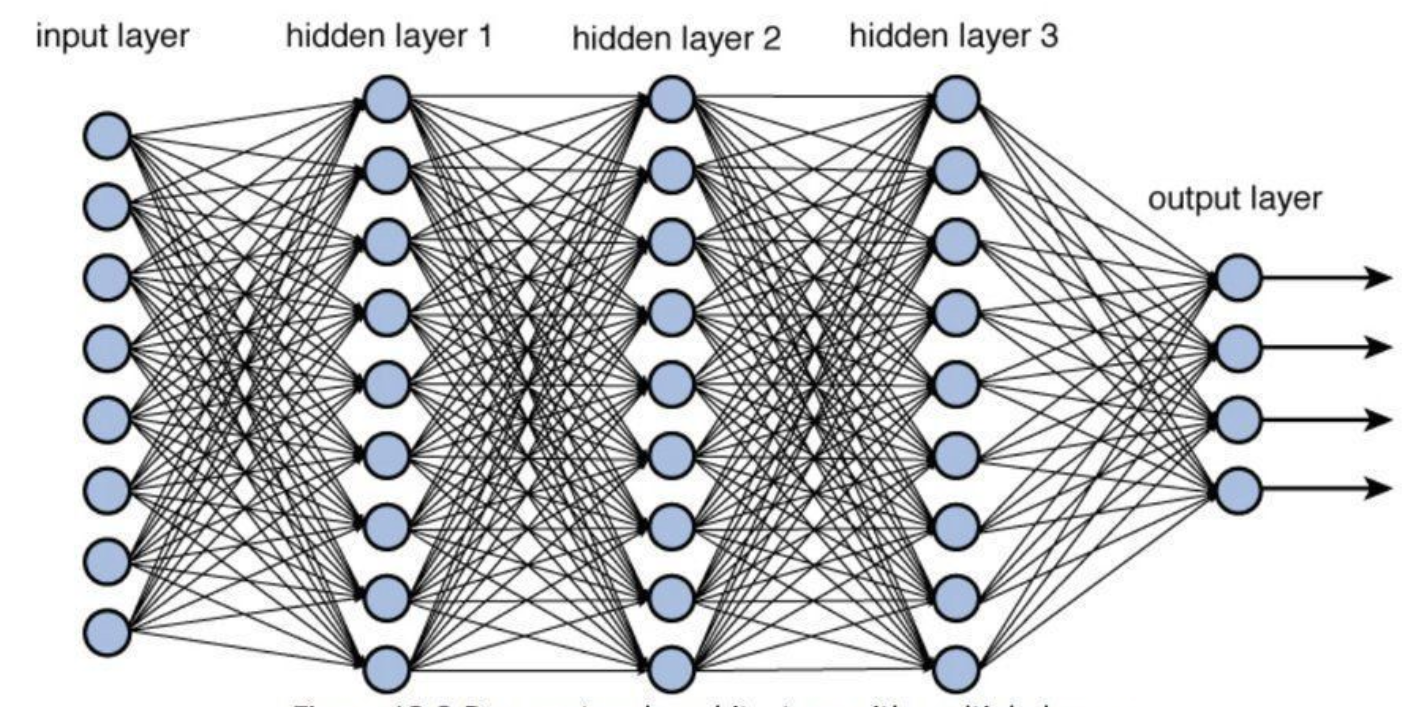


Figure : Structure of a multi-layer neural network [12]

The dimension of a layer in a neural network is defined by the number of nodes in that layer. The neural network in Figure 1 has an input layer of dimension 8, three hidden layers of dimension 9, and an output layer of dimension 4. The network above is called a *feed-forward, fully-connected* neural network. It is fully-connected because every node in each layer is connected to every node in the following layer. It is called a feed-forward network because during a process called *forward propagation*, the information propagates forward from the input layer to the next layer, and then to the subsequent layers that follow. An in-depth example that delineates the forward propagation process is detailed in Appendix B.

A crucial element that influences the predictive power of a neural network is the selected *activation function*. The activation function is applied to the output of every layer in a neural network. Activation functions introduce non-linearity into the weighted sums that define the connections between each layer. This non-linearity allows one to approximate “an arbitrary continuous function, defined on [0,1]” using “a multilayer feed-forward neural network with one hidden layer (that contains only finite number of neurons) using neurons with arbitrary activation functions in the hidden layer and a linear neuron in the output layer” [13] due to the Universal Approximation Theorem [13]. Examples of common activation functions are illustrated in Appendix C.

### Backpropagation and Hyperparameter Tuning

Backpropagation is the process that complements forward propagation. During backpropagation, the neural network uses the results from each attempted inference to update its parameters. It uses a *loss function* to measure how well each prediction fares with respect to the ground truth label. The nature of the loss function is dictated by the task at hand. Each task requires a metric that successfully defines the quality of each prediction. The objective of the neural network is to minimize the defined loss function. Examples of typical binary classification and multi-class classification loss functions are detailed in Appendix D.

When using a gradient-descent-based backpropagation algorithm, the parameters are modified in a way that decreases the loss of the network by the greatest amount each iteration. The speed in which the parameters are updated is proportional to a pre-defined term called the *learning rate.* The learning rate greatly affects the nature of the training process. Small learning rates result in networks that train slowly, but carefully. If the learning rate is too small, however, the network may get stuck in a local minimum and stop learning. Conversely, if the learning rate is too large, the network may become unstable.

Another parameter that is of significant importance during training is the *batch size*. The batch size is the number of data entries that are used during each parameter update. A batch size of 128, for example, means that 128 data points will be used to calculate the gradient for the current update iteration. Large batch sizes result in controlled and accurate updates because the network uses more information to calculate each parameter update. The trade-off is that large batch sizes decrease the number of update iterations that the network experiences.

The batch size and the learning rate are two common examples of *hyperparameters*. Hyperparameters are parameters that are not adjusted automatically during training. These parameters are manually adjusted to increase performance during a process called hyperparameter tuning. Neural network training and hyperparameter tuning are executed intermittently to minimize the pre-defined loss metric on the validation dataset. This process is illustrated in Appendix E.

### Convolutional Neural Networks

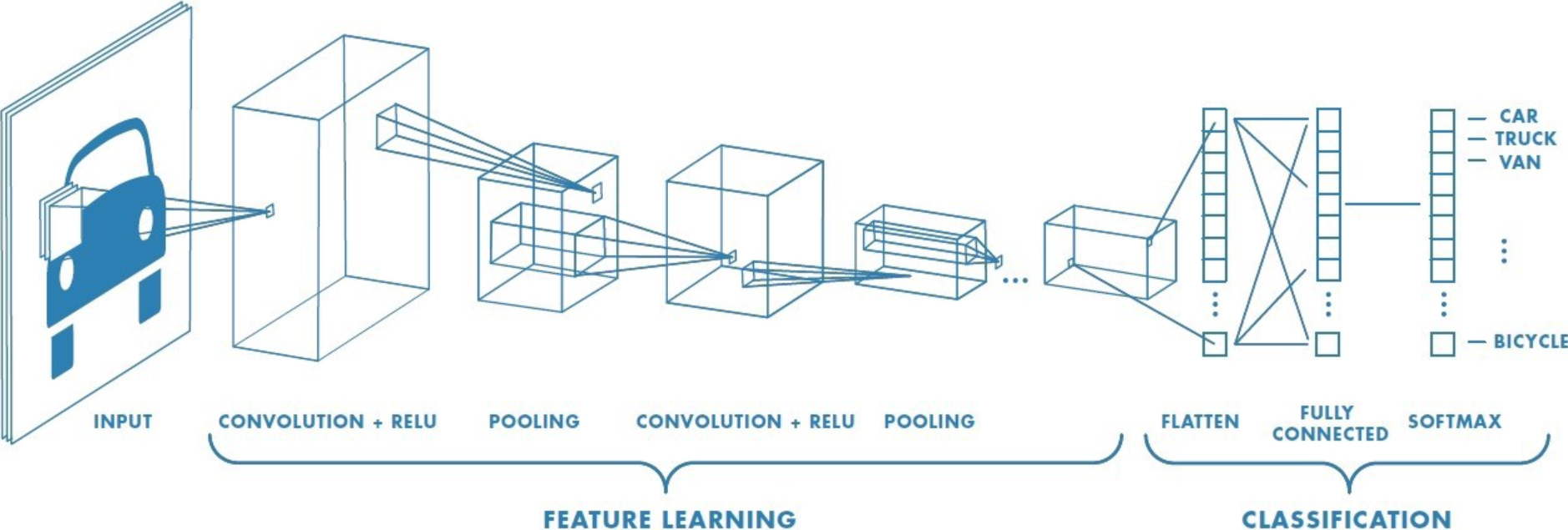
A convolutional neural network (CNN) is a type of neural network that is applied to image classification problems. The following Figure illustrates the structure of a CNN: 

Figure : Illustration of the structure of a CNN [14]

The CNN in the above Figure is an example of a common convolutional structure. A CNN is comprised of a feature learning segment and a classification segment. During feature learning, learnable image filters called kernels are convolved over each convolutional layer. Using these learnable image filters, CNNs are able to discern which characteristics present in the input image data are relevant to the defined image classification problem. An explanation on learnable image filters and convolutions is included in Appendix F.

A pooling layer immediately follows each convolutional layer. Pooling layers reduce the size of the feature representations formed by the convolutional layers. These layers reduce the computational complexity of the network by extracting the most prominent features from the feature set generated by each convolutional layer. There are many ways in which a pooling layer can summarize the contribution from the previous layer. Various pooling layer types are discussed in Appendix G.

As evidenced by the common CNN configuration shown in Figure 2, the pattern of a convolutional layer followed by a pooling layer is repeated sequentially in the feature learning segment of a CNN. The learned features that are produced by the feature learning segment of the CNN are passed into the classification segment. This segment is a fully-connected neural network similar to the structure of the network defined in Figure 1. The number of convolutional layers, pooling layers, fully-connected layers, and the ordering of these layers defines the architecture of the CNN. The architecture of the CNN, in conjunction with the other factors discussed, contribute to the accuracy of the network after training.

### Modern CNN Architectures

In 1994, Yann LeCun’s LeNet5 [15] was released. The network was designed for the generic handwritten digit recognition task. The LeNet5 architecture is illustrated by the following Figure:

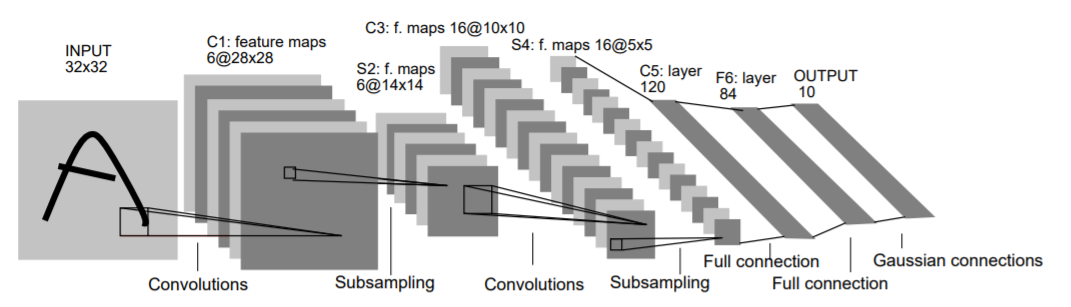


Figure : LeNet5 architecture [15]

The above architecture contains 60,000 parameters. This architecture rather simple in comparison to more modern architectures. This is partially due to the computational burden associated with training a network of this size as a result of the available computational power during the 1990s. In an age where manual heuristics were more common than learned approaches to image classification, LeCun argues that “better pattern recognition systems can be built by relying more on automatic learning and less on hand-designed heuristics” [15].

As GPUs became a general-purpose tool for computation, more complex architectures surfaced. In 2012, Alex Krizhevsky’s AlexNet [16] won the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [17]. Figure 4 details the AlexNet architecture:

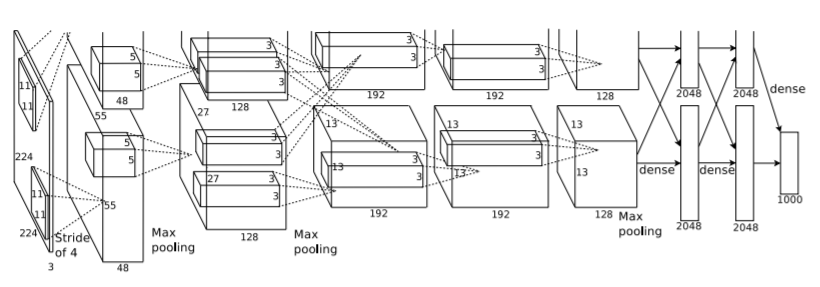


Figure : AlexNet architecture [16]

As illustrated in Figure 4, the architecture is much more complex. AlexNet has 60 million parameters and 650,000 neurons. Compared to the 60,000 parameters present in LeNet5, this is a tremendous increase in complexity. Consistent with the Figure above, the architecture consists of five convolutional layers and three fully-connected layers. Given the advances in GPU technology, the network was trained for a week on two NVIDIA GPUs.

In 2014, GoogLeNet [18], also known as Inception-v1, won the ILSVRC [19]. GoogLeNet boasted monumental accuracy improvements compared to the ILSVRC winners of 2012 and 2013. The fundamental factor responsible for the accuracy improvement yielded by GoogLeNet was the use of a novel neural network component called an inception module shown in Figure 5 below:

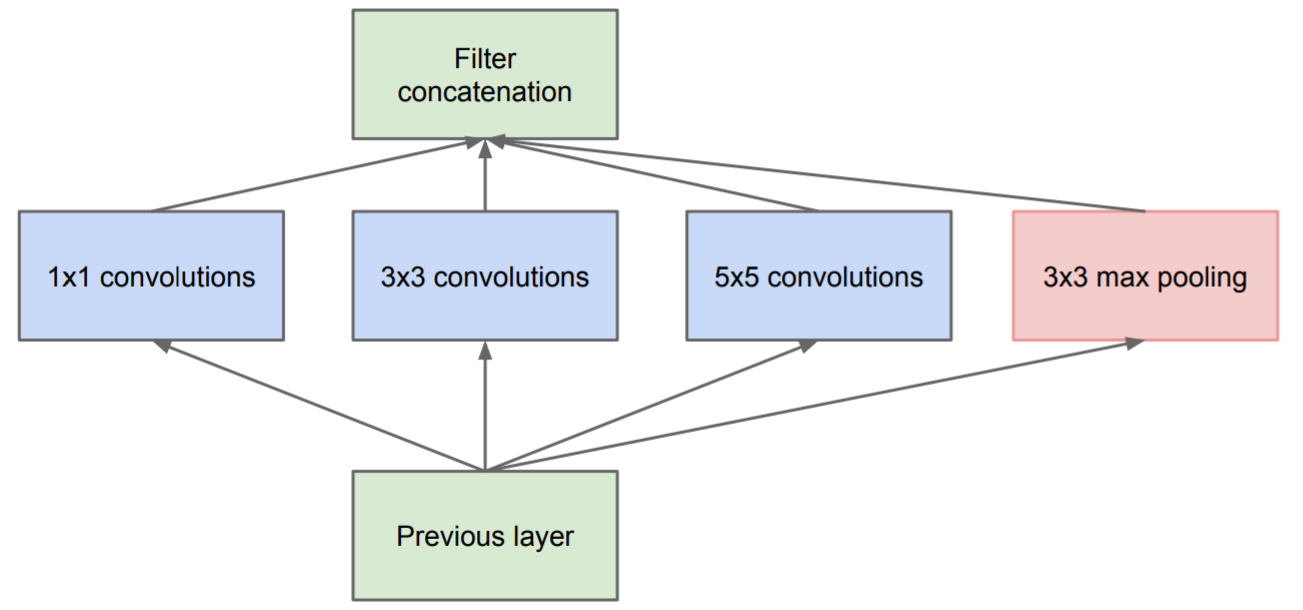


Figure : An illustration of the GoogLeNet inception module [18]

As delineated in Figure 5, the inception module synthesizes multiple filters of varying sizes stemming from the previous layer. The inception module “aligns with the intuition that visual information should be processed at various scales and then aggregated so that the next stage can abstract features from different scales simultaneously” [18]. Since Inception-v1, more modern architectures have presented minor accuracy improvements including Inception-v3 [20], ResNet-50 [21], Xception [22], Inception-v4 [23], Inception-ResNet-V2 [23], and ResNeXt-50 [24].

# Objectives and Metrics

The objective of this thesis is to investigate the impact of various parameter designs in the context of encrypted neural networks. Activation function design, CNN structure design, layer depth adjustments, and the implementation of batch normalization is a non-exhaustive summary of potential performance improvement opportunities. Given the computational requirements, the performance of the experimental architectures will be evaluated using the CKKS encryption scheme.

The training accuracy of the unencrypted neural network is relevant because the expected value of the encrypted inferences is bounded by the training accuracy. To vet weak models, the training accuracy of the unencrypted version of each neural network model will be considered. Strong models maximize the performance in the encrypted realm. Performance, in the context of encrypted inference, corresponds to low computational time and high accuracy. Encrypted neural network applications prioritize low computational time and high accuracy differently. Therefore, alternative metrics must be employed to compare various models.

One such alternative metric is Pareto efficiency. In the context of CrypoNet performance measurement, the two resources that must be maximized are computational time efficiency and accuracy. Given a situation where multiple resources are present, an increase in one resource often results in a decrease in another resource. A Pareto improvement is an increase in one or more resources that does not result in a resultant decrease in any other resources. An allocation is considered optimal if no other allocations are Pareto improvements on that allocation. The set of all efficient allocations is the Pareto front. The models presented in this thesis will be formulated as a Pareto front. This will ensure that models that excel at efficiency, but not accuracy, are not discarded, and vice versa.

# Progress to Date

Microsoft SEAL is an open-source homomorphic encryption library that contains a variety of homomorphic encryption schemes including BFV and CKKS [25]. This repository was cloned and used in conjunction with Microsoft’s existing CryptoNet research [26] to fabricate a working CryptoNet environment. Recent work on CryptoNet optimization [27] [2] operates on the MNIST handwritten dataset [28]. To be consistent with the field standard, the MNIST dataset is used to characterize progress.

A simple candidate model has been designed and implemented. This model is illustrated by Figure 6:

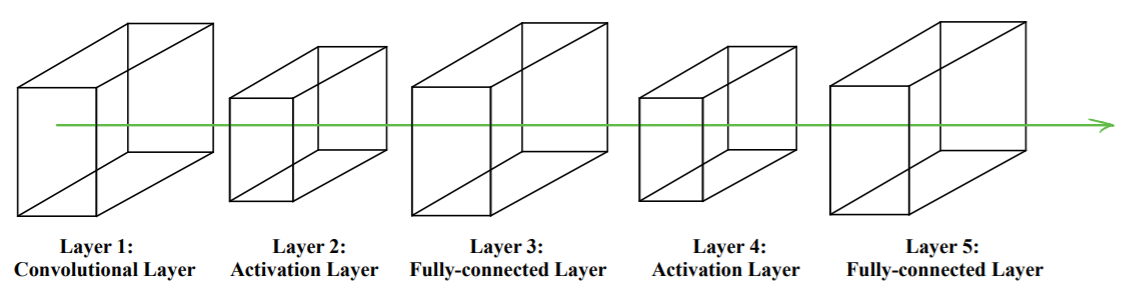


Figure : An illustration of the general structure of a candidate model

The input layer is 28x28x1 because each image has one grayscale channel and is 28 pixels x 28 pixels. The convolutional layer uses five 5x5 kernels, each with a stride of 2, to generate five feature maps. The activation function being applied in Layers 2 and 4 is the x-squared activation function defined below:

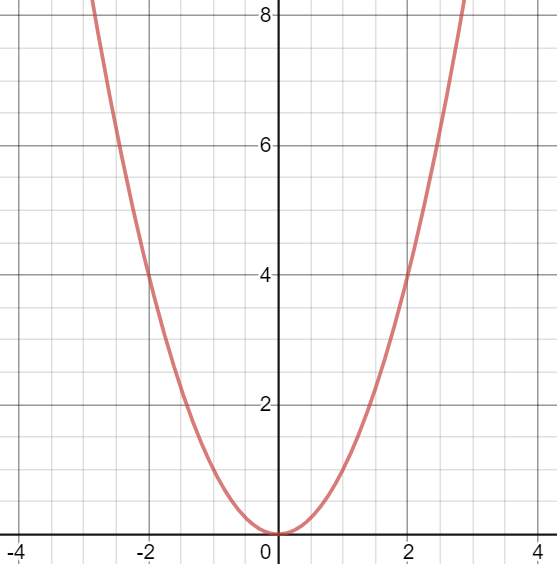


Figure : A graph [29] of the x-squared activation function

The fully-connected layer of Layer 3 receives an input from the activation layer of size 5x13x13 and downscales it to a size of 100. The output of Layer 3 is then passed back into the x-squared activation function, and through to the fully-connected output layer.

The network detailed above was translated into the encrypted realm. Using an Intel® Core™ i7-7500U CPU at 2.70GHz, the computational time at each layer was measured. These results are shown in Table 1:

Table : Computational time associated with each layer

|  |  |
| --- | --- |
| Layer: | Time(s) to infer on 8192 entries: |
| Layer 1: Convolutional Layer | 12.2713749s |
| Layer 2: Activation Layer | 35.2610618s |
| Layer 3: Fully-connected Layer | 42.8963216s |
| Layer 4: Activation Layer | 3.9026061s |
| Layer 5: Output Layer | 0.521606s |

From Table 1 above, there are two obvious computational bottlenecks. The activation function following the initial convolutional layer took 35.26s to compute. The multiplicative nature of the activation function may contribute to the computational congestion it creates. Significant time was also spent on the fully-connected layer. A smaller kernel size, less feature maps, and a large stride are potential changes that could facilitate a speedup in the third layer. Regarding accuracy, this model achieved an accuracy of 98.95% in the encrypted realm. Thus, there is room for computational efficiency trade-offs.

# Future Work

A successful contribution to the field will be a result of relentless experimentation and a thorough analysis of many possible models. Only one model has been explored. Therefore, significant work remains. To accomplish the previously outlined objectives, a rigorous timeline has been established. The final presentations start 56 days after January 27, 2020, on March 23, 2020. The final report is due 73 days after January 27, 2020, on April 9, 2020. A Gantt chart delineating tasks that must be accomplished in the days leading up to the final evaluations is detailed in Figure 8, below:

Figure : Gantt chart detailing a procedural timeline leading up to the final evaluations

The first three weeks are dedicated to researching the effects of altering various factors that impact the performance of a neural network in the encrypted space. From the initial result discussed in the previous section, the two modifications that seem the most influential are the modification of the activation function and the modification of the general network structure. Three weeks of consistent research is ample time to thoroughly investigate a wide variety of models. During this initial sprint, performance will be closely tracked. By the end of the three weeks, it is expected that the results of the thesis will be completely compiled.

The weeks that follow are committed to the production of a final report draft. It is important to develop a draft well in advance because this draft will serve as a basis for the final presentation slides. These slides will be finished a week before the presentations begin. In the week leading up to the final presentations, time will be set aside to practice the in-person segment of the presentation. During the final two weeks leading up to the due date of the final report, trusted sources will be selected to review the final report draft, and the final report will be completed.

# References

|  |  |
| --- | --- |
| [1] | C. Gentry, "Fully homomorphic encryption using ideal lattices," *In Proc. STOC,* pp. 169-178, 2009. |
| [2] | N. Dowli, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig and J. Wernsing, "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," Microsoft Research, 2016. |
| [3] | C. Gentry and S. Halevi, "Implementing Gentry's fully-homomorphic encryption scheme," *EUROCRYPT'11: Proceedings of the 30th Annual international conference on Theory and applications of cryptographic techniques: advances in cryptology 2011,* p. 129–148, 2011. |
| [4] | M. van Dijk, C. Gentry, S. Halevi and V. Vaikuntanathan, "Fully Homomorphic Encryption over the Integers," in *Advances in Cryptology -- EUROCRYPT 2010*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2010, pp. 24-43. |
| [5] | Z. Brakerski, C. Gentry and V. Vaikuntanathan, "(Leveled) Fully homomorphic encryption without bootstrapping," in *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012. |
| [6] | A. Lopez-Alt, E. Tromer and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," *STOC '12: Proceedings of the forty-fourth annual ACM symposium on Theory of computing,* p. 1219–1234, 2012. |
| [7] | Z. Brakerski, "Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP," *Proceedings of the 32nd Annual Cryptology Conference on Advances in Cryptology --- CRYPTO 2012,* vol. 7417, p. 868–886, 2012. |
| [8] | J. Fan and F. Vercauteren, "Somewhat Practical Fully Homomorphic Encryption," *IACR Cryptology ePrint Archive 2012/144,* 2012. |
| [9] | J. W. Bos, K. Lauter, J. Loftus and M. Naehrig, "Improved Security for a Ring-Based Fully Homomorphic Encryption Scheme," in *Cryptography and Coding*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2013, pp. 45-64. |
| [10] | J. H. Cheon, A. Kim, M. Kim and Y. Song, "Homomorphic Encryption for Arithmetic of Approximate Numbers," *Cryptology ePrint Archive, Report 2016/421,* 2016. |
| [11] | Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *Journal of Analysis and Testing,* vol. 2, 2018. |
| [12] | R. Parmar, "Deep network architecture with multiple layers.," Towards Data Science, 2018. |
| [13] | B. C. Csáji, "Approximation with Artificial Neural Networks," *MSc Thesis,* p. 22, 2001. |
| [14] | S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way," Towards Data Science, 2018. |
| [15] | Y. Lecun, L. Bottou, Y. Benigo and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998. |
| [16] | A. Krizhevsky, G. Hinton and I. Sutskever, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25,* pp. 1097-1105, 2012. |
| [17] | "ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)," [Online]. Available: http://image-net.org/challenges/LSVRC/2012/results.html. [Accessed 26 January 2020]. |
| [18] | C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1-9, 2015. |
| [19] | "ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014)," [Online]. Available: http://image-net.org/challenges/LSVRC/2014/results. [Accessed 26 January 2020]. |
| [20] | C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2818-2826, 2016. |
| [21] | K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, 2016. |
| [22] | F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1800-1807, 2017. |
| [23] | C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,* p. 4278–4284, 2017. |
| [24] | S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 5987-5995, 2017. |
| [25] | "Microsoft SEAL (release 3.2)," Microsoft Research, Redmond, WA., February 2019. [Online]. Available: https://github.com/Microsoft/SEAL. |
| [26] | A. Brutzkus, O. Elisha and R. Gilad-Bachrach, "Low Latency Privacy Preserving Inference," *CoRR,* vol. abs/1812.10659, 2018 . |
| [27] | E. Chou, J. Beal, D. Levy, S. Yeung, A. Haque and L. Fei-Fei, "Faster CryptoNets: Leveraging Sparsity for Real-World Encrypted Inference," *CoRR,* vol. abs/1811.09953, 2018. |
| [28] | Y. LeCun, C. Cortes and C. Burges, "THE MNIST DATABASE of handwritten digits," 1999. [Online]. Available: http://yann.lecun.com/exdb/mnist/. [Accessed 27 January 2020]. |
| [29] | "Desmos Graphing Calculator," Desmos, 2020. [Online]. Available: https://www.desmos.com/calculator. [Accessed 27 01 2020]. |
| [30] | I. Goodfellow, Y. Bengio and A. Courville, in *Deep Learning (Adaptive Computation and Machine Learning series)*, MIT Press, 2016, p. 226. |

# Appendix A: Public-Key Encryption

Suppose Bob wishes to send an encrypted to message to his friend, Alice. To do so, Alice would create two keys: a private key and a public key. The public key can encrypt messages into ciphertexts, and the private key can decrypt those messages back into plaintext. After creating this pair of keys, Alice would send the public key to Bob as in Figure 9, below:

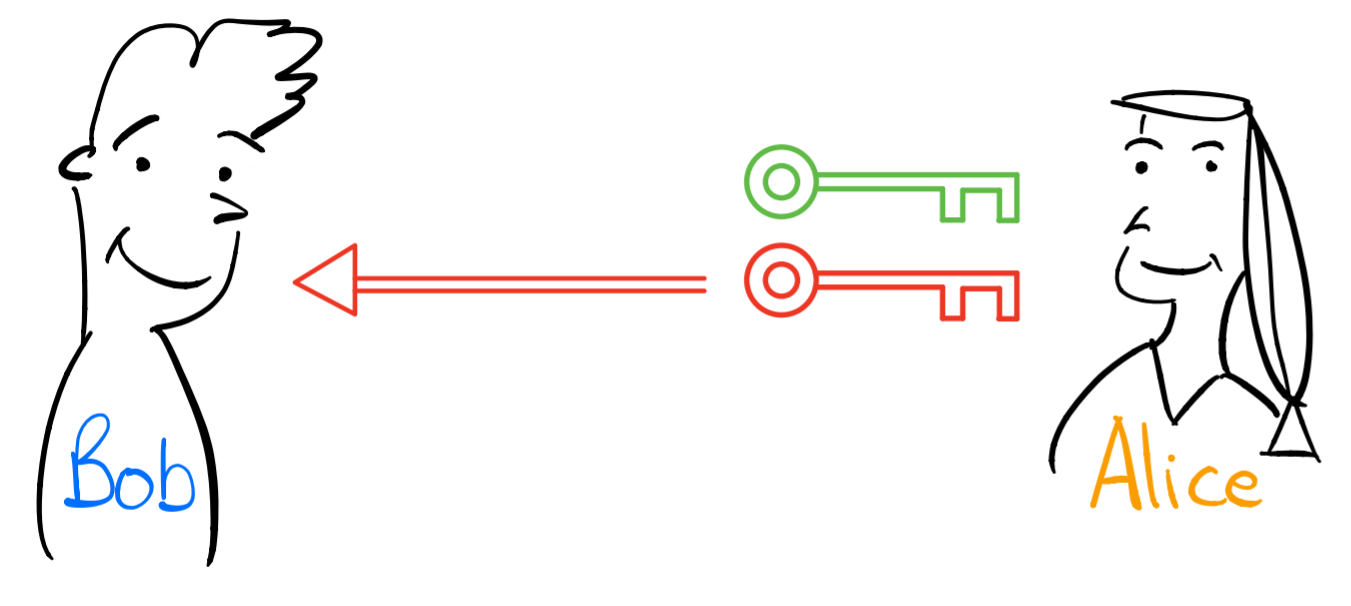


Figure : Alice sends Bob a public key (red), keeping her private key (green) to herself

Upon receiving the public key from Alice, Bob can use the public key to encrypt a message:

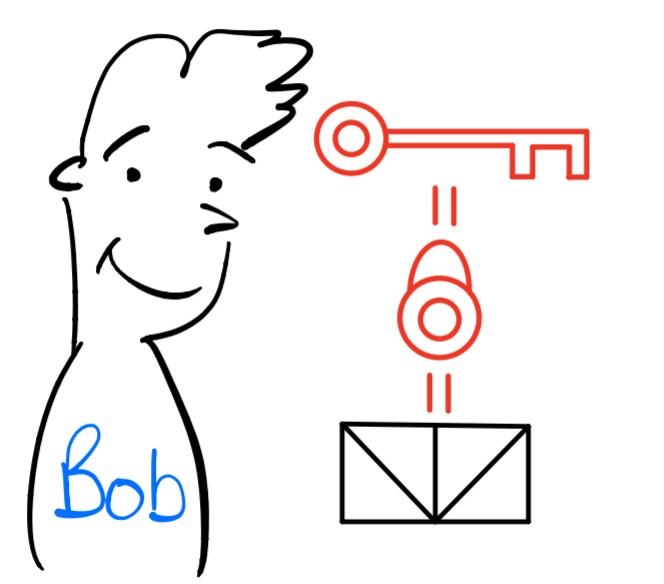


Figure : Bob uses Alice’s public key to encrypt a message

Then, Bob can send his encrypted message to Alice:

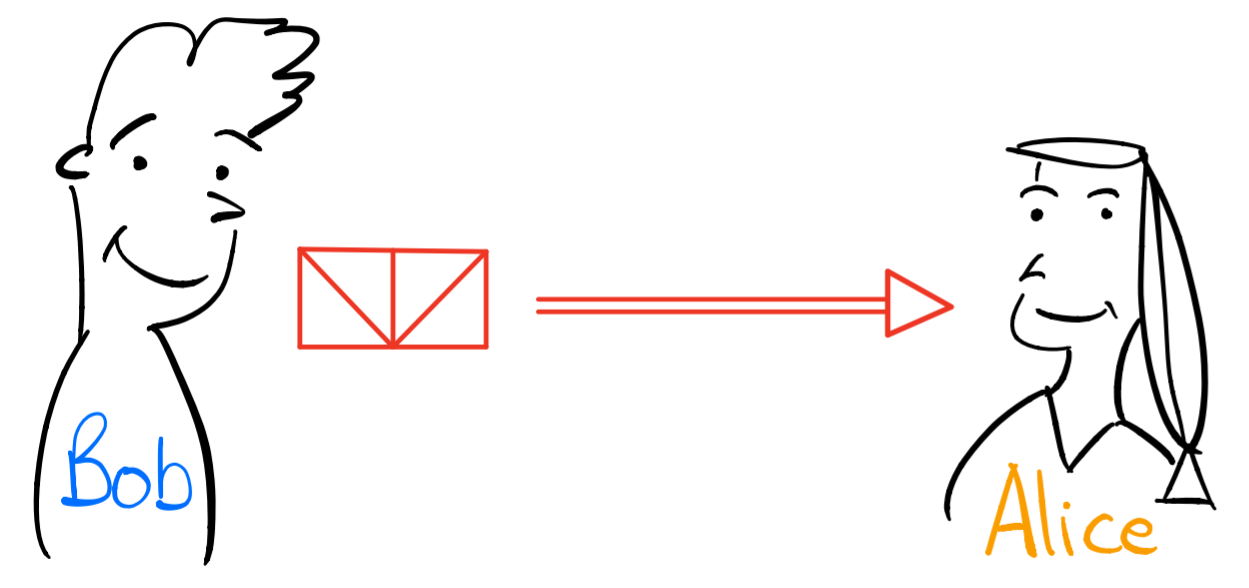


Figure : Bob sends his encrypted message to Alice

Using her private key, Alice can then decrypt Bob’s encrypted message back into readable plaintext, as in Figures 12 and 13:

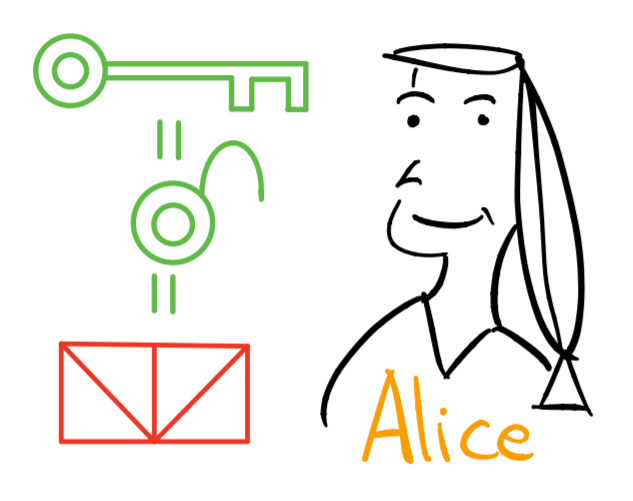


Figure : Alice uses her private key to decrypt Bob’s encrypted message

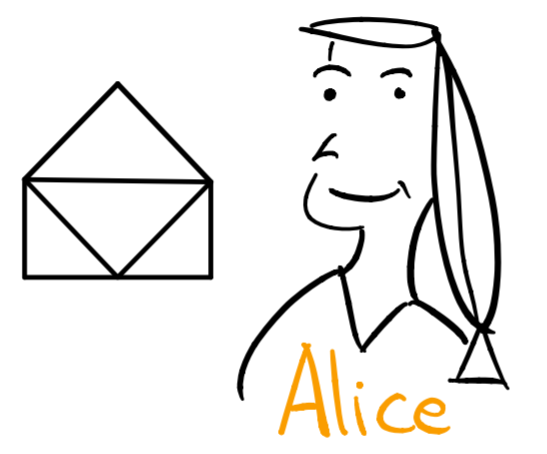


Figure : Alice can now view Bob’s message

During this entire process, Alice has kept her private key secure. A third party listening in on Alice and Bob’s conversation could have captured the contents of the public key, but would not be privy to Alice’s private key, the essential element to decrypting Bob’s messages.

# Appendix B: Forward Propagation

The input to each neural network is vector of numbers of a constant dimension. During forward propagation, each node is assigned a calculated value based on the parameters of the network and the values of the nodes in the previous layer. Consider the following illustration of a simple neural network:

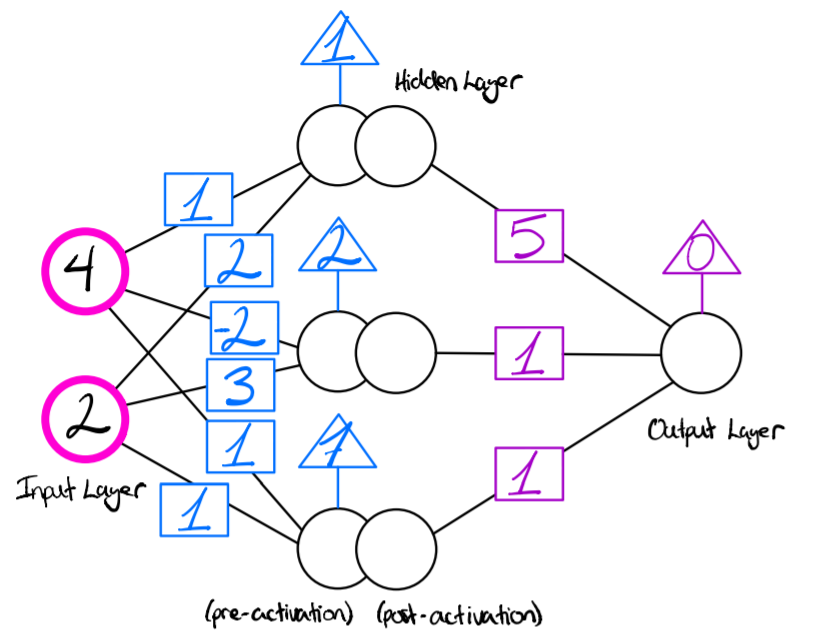


Figure : An example of a fully-connected neural network prior to forward propagation

The neural network detailed by Figure 14 has an input layer of dimension 2, one hidden layer of dimension 3, and an output layer of dimension 1. The weights connecting the input layer to the hidden layer are the numbers in the blue boxes. The biases associated with the hidden layer are the numbers in the blue triangles. Similarly, the weights connecting the hidden layer to the output layer and the bias term associated with the output layer are the numbers in the purple boxes and purple triangles, respectively. To illustrate the effects of the activation function, the hidden layer has been split into a pre-activation phase and a post activation phase.

During forward propagation, the input vector is propagated forward to the hidden layer. This propagation is the result of a multi-step process. First, the weighted sum of the input vector and the parameters connecting the input layer and hidden layer is calculated for each node. These calculations are illustrated below:

After calculating the weighted sum for each node, the bias terms associated with each node are added to their respective nodes. Figure 12 illustrates the result of these calculations:

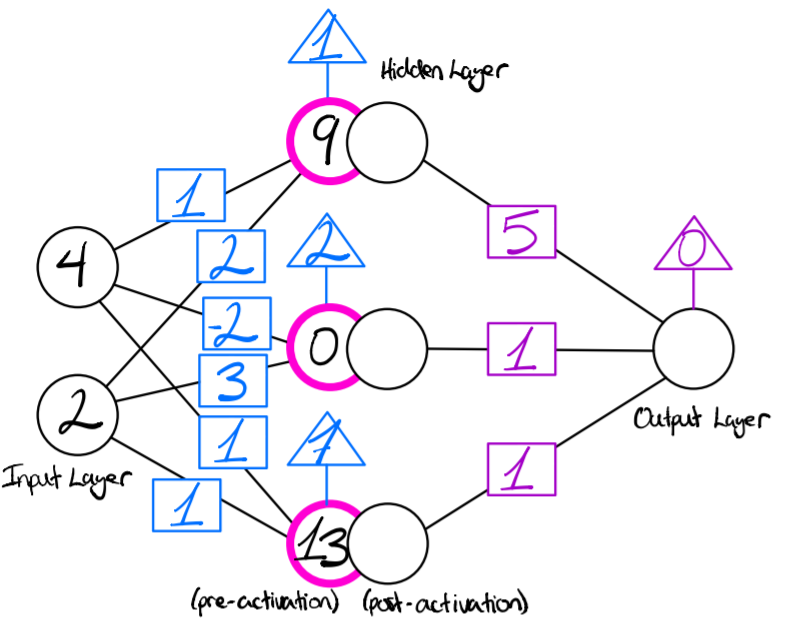


Figure : Illustration of the forward propagation process prior to the application of the activation function

Now, the values corresponding to each node are passed through an *activation function.* For a detailed explanation on activation functions, see Appendix D. The activation function used in this example is the Sigmoid activation function.

The values of the nodes in the hidden layer prior to being passed through the Sigmoid activation function are 9, 0, and 13, respectively. As such, we expect the Sigmoid activation function to activate the nodes with values of 9 and 13 and to mediate the contribution from the node with a value of 0. Figure 16, below, illustrates the result of applying the Sigmoid activation function to the hidden layer:

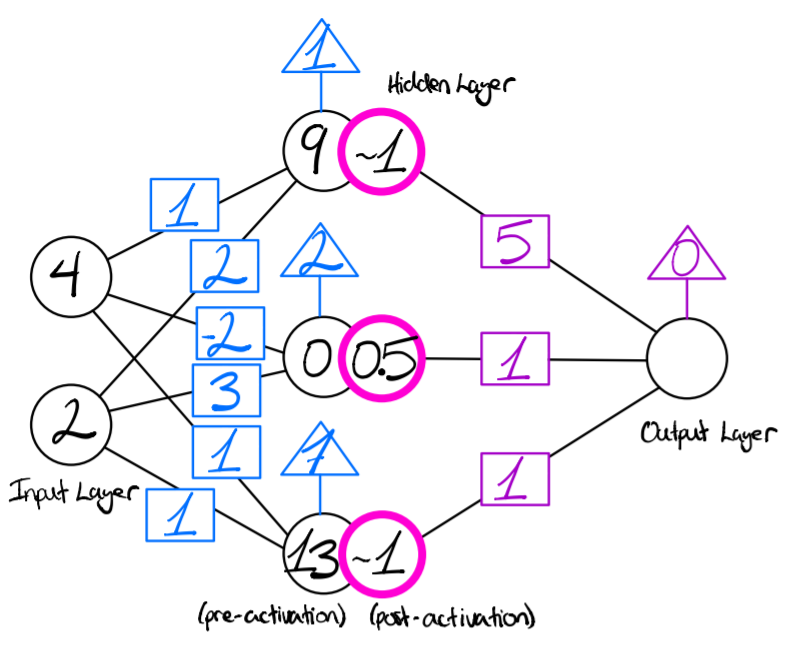


Figure : The forward propagation process after applying the activation function to the nodes in the hidden layer

As illustrated by Figure 16, the nodes in the hidden layer are transformed from values of 9, 0, and 13, to values of approximately 1, 0.5, and 1.

The values from the input layer have now propagated to the hidden layer. These values will continue to propagate in a similar fashion to the output layer. The output layer, in this case, uses a linear activation function. This is because there is a single output and the problem that is implicitly being solved by this simplified example is a regression problem. In a multi-class classification problem, however, one might use a Softmax activation function. For more information on common activation functions, see Appendix C. The final output of the neural network that was previously Figure 14 is delineated by Figure 17, below:

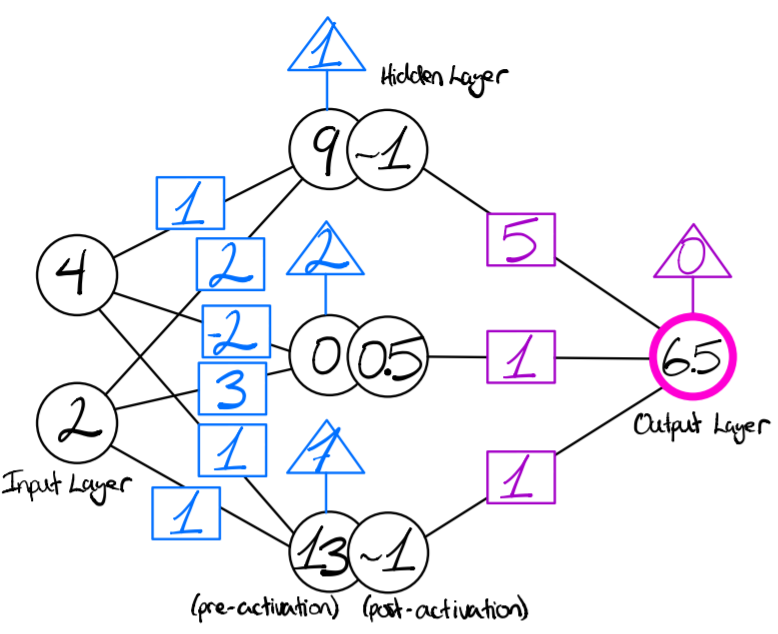


Figure : An illustration of the output of the neural network after forward propagation has occurred

The network has fully propagated the values from the input vector through to the output layer. The network has made an inference on the input vector and has completed the forward propagation process.

# Appendix C: Activation Functions

A wide variety of activation functions are commonly applied to the output of a fully-connected neural network. Traditional activation functions include the Sigmoid function, the tanh function, and the arctan function. The equations and graphs corresponding to the functions listed above are included below:

## Sigmoid

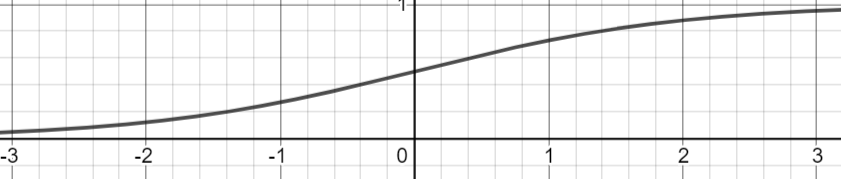


Figure : A graph [29] of the Sigmoid activation function

The Sigmoid activation function truncates large positive values to an output value close to 1 and truncates large negative values to an output value close to 0. The Sigmoid activation function activates nodes that yield a large positive value and suppressesnodes that yield a large negative value.

## Tanh

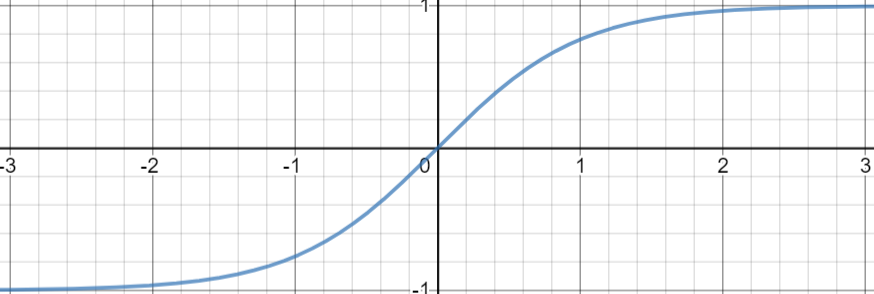


Figure : A graph [29] of the tanh activation function

Similar to the Sigmoid function, the tanh activation function truncates large positive values to an output value close to 1. Large negative values, however, are truncated to a value of -1 as opposed to 0.

## Arctan

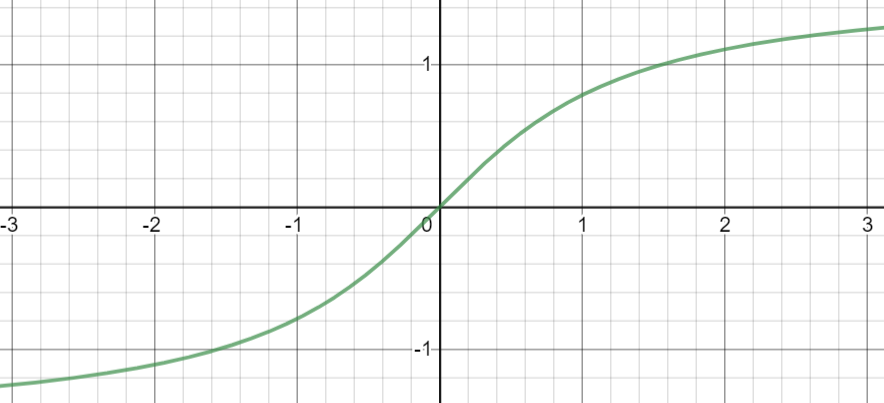


Figure : A graph [29] of the arctan activation function

The arctan function, as illustrated above, decays large positive and negative values as they grow. This ensures that the gradient of the activation function does not explode during backpropagation.

## ReLU

In recent years, the rectified linear unit (ReLU) activation function has increased in popularity due to the observed performance increase: “[A] major algorithmic change that has greatly improved the performance of feedforward networks was the replacement of sigmoid hidden units with piecewise linear hidden units, such as rectified linear units” [30]. The following Equation and Figure illustrate the nature of the ReLU activation function:

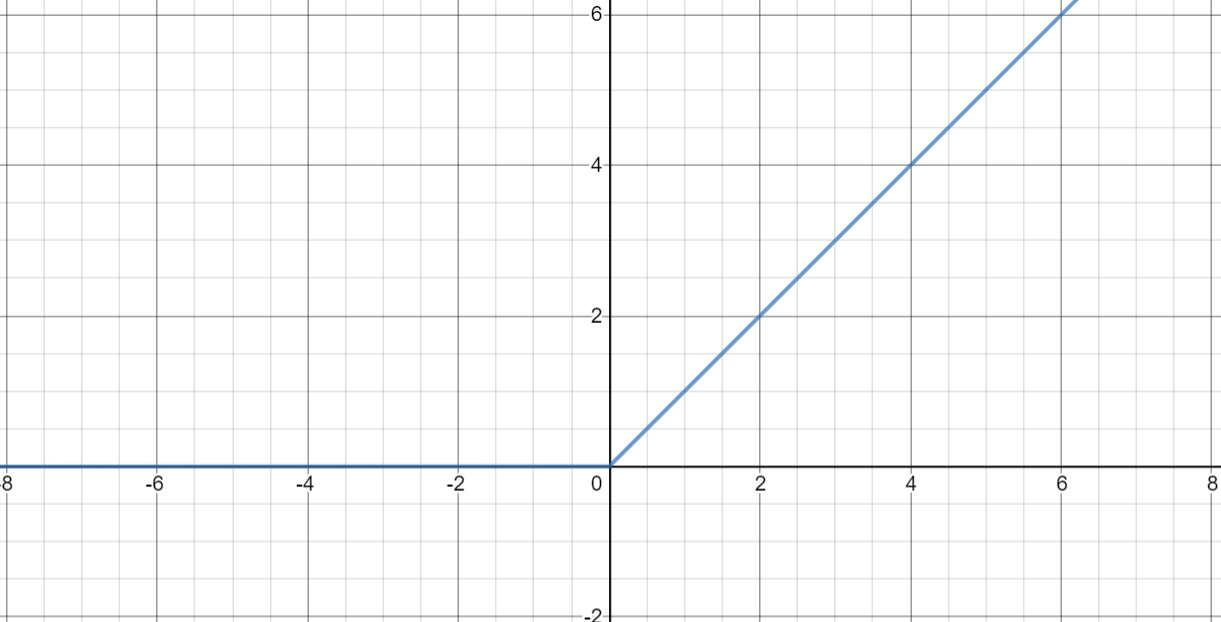


Figure : A graph [29] of the ReLU activation function

Unlike the above activation functions, the ReLU function truncates negative numbers and simply applies a linear function to the positive numbers. CNNs commonly use the ReLU activation function due to the aforementioned performance advantages.

## Leaky ReLU

One problem that arises with the ReLU activation function is that due to the truncation of negative neurons, certain neurons may never contribute to the functionality of the network because the gradients of those neurons will always be zero. These neurons are called *dead neurons*. To address this issue, the Leaky ReLU function was conceived:

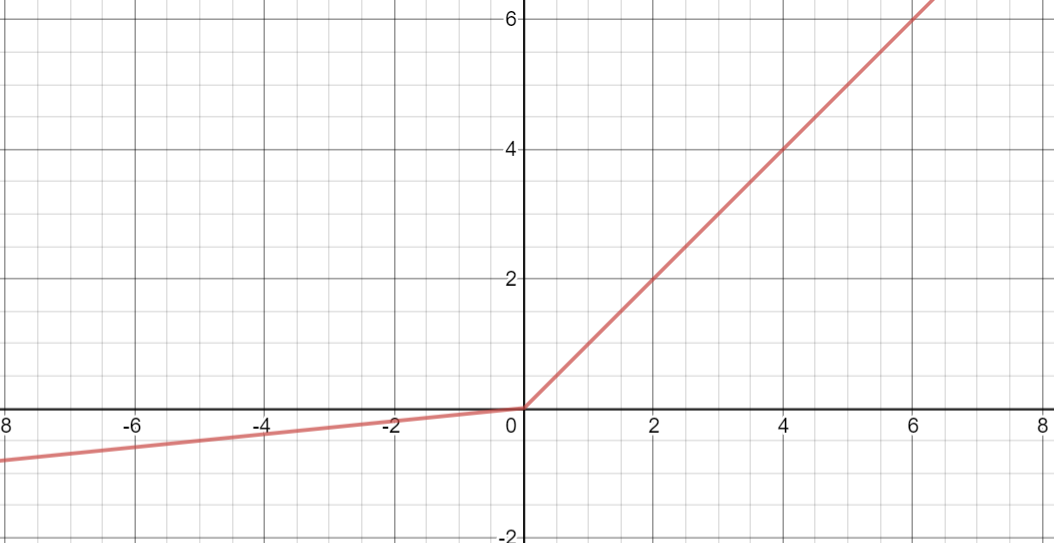


Figure : A graph [29] of the Leaky ReLU function

The Leaky ReLU function facilitates a small positive linear function for numbers less than 0. This effectively revives the dead neurons by allowing negative neurons to contribute to the predictive power of the network.

# Appendix D: Loss Functions for Binary and Multi-Class Classification

For binary classification problems, a common loss function is the binary cross-entropy (BCE) loss function. This loss function is defined below:

In the above equation, the function *y(c)* is either 0 or 1, depending on which class the current data entry belongs to. The function *p(c)* is a discrete function containing the predicted probabilities of the two classes.

For multi-class classification problems, the BCE loss function can be extended to include multiple classes. This loss function is called the cross-entropy loss function and it is also defined below:

The classification problem pertaining to the defined loss function above has M classes. The function *y(c)* is 1 if the current data entry has class *c*. The discrete function *p(c)* contains the predicted probabilities of each class.

The binary cross-entropy and cross-entropy loss functions attempt to accurately define a loss metric that delineates how accurate each prediction with respect to the ground truth label. Figure 23 is a graph of the negative natural logarithm:

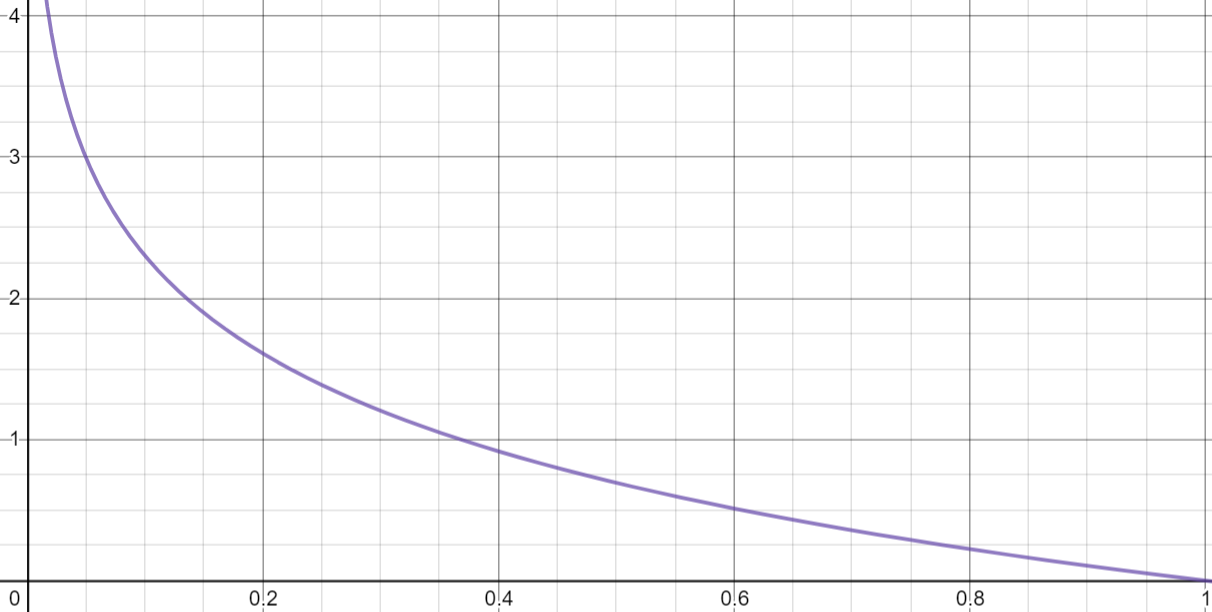


Figure : A graph [29] of the negative natural logarithm, y = -ln(x)

Recall that maximizing the cross-entropy loss is the same thing as minimizing the logarithmic loss. Thus, Figure 23 illustrates the impact of a prediction on a given ground truth label. If the network predicts the ground truth label with a low confidence, then the negative natural logarithm of Figure 23 will approach a very large number. If the network predicts the correct ground truth label with high confidence, however, the negative natural logarithm will be very low, as desired. This is the intuition behind using the negative natural logarithm as a loss function.

# Appendix E: Loss Curves and Neural Network Training

The performance of a neural network during training is measured using a *loss curve*. The loss curve measures the pre-defined loss of a neural network as a function of the number of update iterations the network has experienced. The following is an example of a neural network validation loss curve on top of a training loss curve. In this example, the loss function being measured is the negative natural logarithm loss function.

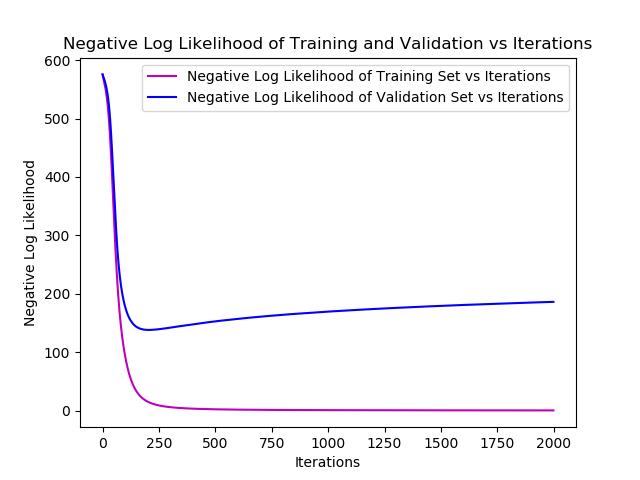


Figure : A typical graph of the validation and training loss of a neural network as a function of the update iterations

The Figure above is a strong representation of what typical validation and training loss curves look like. Initially, the validation loss decreases rapidly. Then, as the neural network continues to train on the training dataset, the validation loss increasesas the training loss continues to decrease. This occurs because the neural network is starting to memorize the training data. Models that are able to make accurate predictions on unseen data are valuable. Models that can learn a certain set of data and only predict on seen data are completely useless. As such, validation and training loss curves are used to determine which version of a model should be selected. In this example, the version of the model at iteration 200 would be selected because it minimizes the validation loss.

# Appendix F: Linear Image Filtering and Convolutions

CNNs use simple linear image filters to extract information from the input image. The following Figure is a grayscale image of Jenga blocks:



Figure : Grayscale image of Jenga blocks

The above image can be represented by a matrix of pixel values corresponding to the pixel intensities at each location in the image. Linear image filters can be convolved over this image to produce new versions of the above image that are more meaningful.

One such linear filter is the Sobel filter. The Sobel filter calculates the gradient of an input image in a desired direction. The 3x3 right Sobel, illustrated by Figure 26 below, computes the horizontal derivative of each pixel with respect to the adjacent pixels.

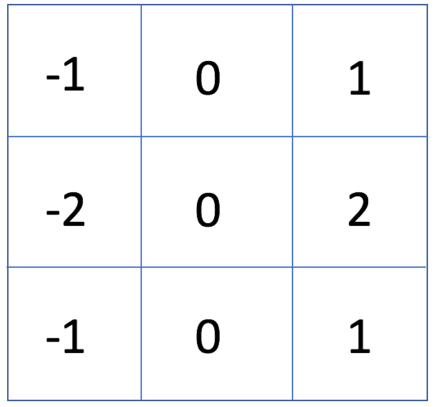


Figure : Illustration of the 3x3 right Sobel filter

By applying the right Sobel filter to the grayscale image of Jenga blocks in Figure 25, the following image is obtained:



Figure : Result of the application of the right Sobel filter

The image above is smooth in the areas that do not have well-defined horizontal edges. The right Sobel filter accentuates the locations in the image that have significant horizontal derivates. In other words, the right Sobel filter captures the locations and magnitudes of the horizontal edges present in the original image. These edges may be useful in the context of image classification because the locations and magnitude of the edges could help indicate which class that image belongs to.

CNNs are powerful because they learn which image features matter with respect to the classification task at hand. Each convolutional layer uses a series of kernels of a pre-defined dimension. The dimension of a kernel is defined by the height and width of that particular kernel. The Sobel filter in Figure 26 has dimension 3, for example. The adjustable weights of a convolutional layer are the matrix entries of the filters associated with that layer. During backpropagation, the weights of each convolutional filter in a convolutional layer are updated to minimize the pre-defined loss function. In this way, each convolutional layer learns which input image characteristics are relevant to the meaning of the image as a whole.

# Appendix G: Types of Pooling Layers

There are two categories of pooling layers: local pooling layers and global pooling layers. Local pooling layers divide the input matrix into windowed chunks. Then, the contribution from each windowed chunk is extracted. The two most common local pooling types are average pooling and max pooling. Global pooling layers compute a feature of a significantly smaller dimension using a global mathematical operation that captures the meaning of each feature map. Global average pooling is an example of a type of a global pooling layer.

Consider the following matrix of integers:

|  |  |  |  |
| --- | --- | --- | --- |
| 5 | 4 | 2 | 5 |
| 0 | 4 | 2 | 9 |
| 2 | 1 | 1 | 7 |
| 2 | 1 | 4 | 3 |

Figure : 4x4 matrix of integers representing the input to a pooling layer

Let the above matrix serve as an input to a pooling layer. The colors in the above Figure define the windowed chunks that the input matrix has been divided into. Average-pooling uses the average of each windowed chunk to define a smaller matrix that attempts to capture the contribution from each individual window. The result of applying average-pooling to the above Figure is a 2x2 output matrix shown below:

|  |  |
| --- | --- |
| 3.25 | 4.5 |
| 1.5 | 3.75 |

Figure : Result of applying average pooling to the input matrix in Figure 25

This matrix is able to retain some of the information from the input matrix. Furthermore, this matrix is half the size of the input matrix.

Max-pooling uses the maximum value from each windowed chunk to perform the same dimensionality reduction. The result of applying max pooling to the input matrix of Figure 28 is illustrated below:

|  |  |
| --- | --- |
| 5 | 9 |
| 2 | 7 |

Figure : Result of applying max-pooling to the input matrix of Figure 25

Max-pooling and average pooling each reduce the dimensionality of the input while attempting to retain the information conveyed by that input.

Average global pooling is a completely different approach. Instead of splitting the input matrix into windowed chunks, the algorithm takes the average of the entire input matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| 5 | 4 | 2 | 5 |
| 0 | 4 | 2 | 9 |
| 2 | 1 | 1 | 7 |
| 2 | 1 | 4 | 3 |

|  |
| --- |
| 3.25 |

Figure : Result of applying global average pooling to the input matrix

The original 4x4 input matrix is reduced to a 1x1 matrix. Surprisingly, this pooling method can be very effective and is used as a fundamental component of the ResNet50 architecture [21].