# Fire Sprinkler Classifier with Deep Learning

*Justin Shin*

George Mason University

## 1. Introduction

Construction site surveys are often in need of classifying specific parts or materials in the field. Most of the time, the surveyor would have to classify the item using their domain expertise or they could use a computational tool that can automate such processes. But there is a lack in such technologies as this industry is notoriously infamous for adopting such new technologies at a much slower pace than others. However, in the realm of Computer Science, there have been exponential growth from research in artificial neural networks (ANN) that have allowed for such classifications to be automated and have grown more accurate and practical with every passing year. Such technologies as ResNet [1] and Vision Transformers [2] can aid a construction site surveyor in such classifications tasks with such convenience that it can change the industry forever.

## 2. Motivation

I currently work as a Construction Site Surveyor, specifically for Fire Sprinkler Systems. I go to construction sites and generate a map of the sprinkler system to inform the installers about the environment before they are deployed. I record the pipe sizes and elevations, take pictures, and most importantly, identify the sprinkler model. 80% of the time, I am able to identify the model with a quick glance, but for the other 20% of the time, I am unable to due to some of the following reasons: the sprinkler head may be 50 feet above and unreachable; the low visibility in a construction site due to dust or no power; I may encounter an antique sprinkler which may not have any marking to identify the model. These situations and more happen occasionally and it would be convenient to be able to identify the models with just a single snap of a picture, hence the idea and the primary motive for this project. The secondary motive is my interest in learning more about Transformers since the current hype around the performant model seen from examples in OpenAI's GPT models [3] Coincidentally, Google Researchers have invented a method to repurpose the Transformer, which was traditionally used to tackle Natural Language Processing (NLP) problems, to process images and perform image classification. It is called Vision Transformer [2] and the researchers claimed that it can be as performant as the state-of-the-art CNN models like ResNet and EfficientNet but also computationally more efficient. This project will allow me to study the architecture of the original Transformer as well as the newly proposed ViT in-depth as well as further my understanding of Attention [4] The objective of this project is to develop a model that can accurately predict sprinkler so to be used out in the field immediately. So the most performant model was to be deployed as soon as the model performs at an acceptable rate of 80%.

## 3. Approach

The first task was to collect data for the models to train on. The images were collected from my iPad, which I have used to take pictures of the construction site as well as a picture of the sprinkler head. I have worked as a surveyor for nearly 10 years but limited the dataset to 2.5 years of images due to time constraints. The images were then preprocessed. Almost all the images were taken in portrait mode using the iPad but for the few that were taken in landscape were forced into portrait mode by rotating the image 90 degrees. Then the images were resized and cropped to standardize the input size as well as meet the size requirements of the model (ViT requires 224x224). Then the models were trained using the finalized dataset. The first model was a modified version of the CNN model used in Homework 5. It was performed on the Fashion-MNIST [5] dataset and expected to work well on this dataset as well. The second model trained was a ResNet CNN model as it was often used as the state-of-the-art model that was compared to the performance of the Vision Transformer. The last model to train was the Vision Transformer.

## 4. Dataset

The dataset consists of 342 hand-labeled sprinkler images which fall under 12 classes. These images were collected over the course of 2.5 years since 2020 but not all were used in the dataset. Some images were too blurry to identify the correct model. And for some, there were only a handful of sample images for a particular model. This could hinder the model from training so if there were less than five images for a model, they were excluded. As shown in the distribution figure below, the model with the most samples is the V3802. This is because it is the most popular sprinkler used in the DC metropolitan area for commercial buildings.

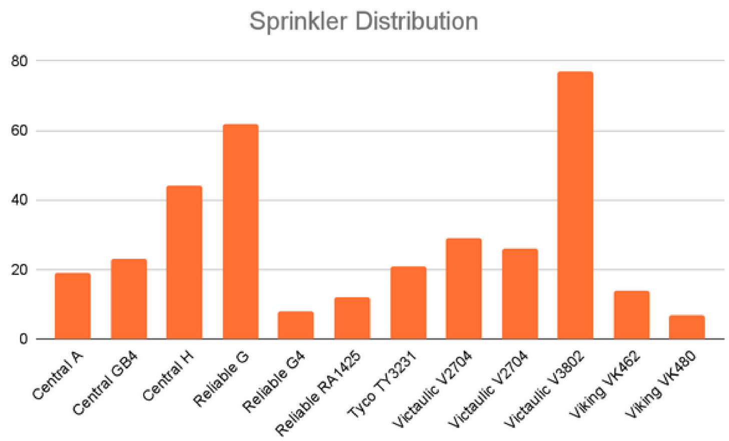

Figure 1. Distribution of Sprinkler Models

The dataset also consists of variations of a model. This includes blurry images, images taken in low light environments, images taken from a distance, and images of sprinkler heads with different paint colors. This will allow the models to be more robust against these real-world situations. These variations can be seen in the figure below.

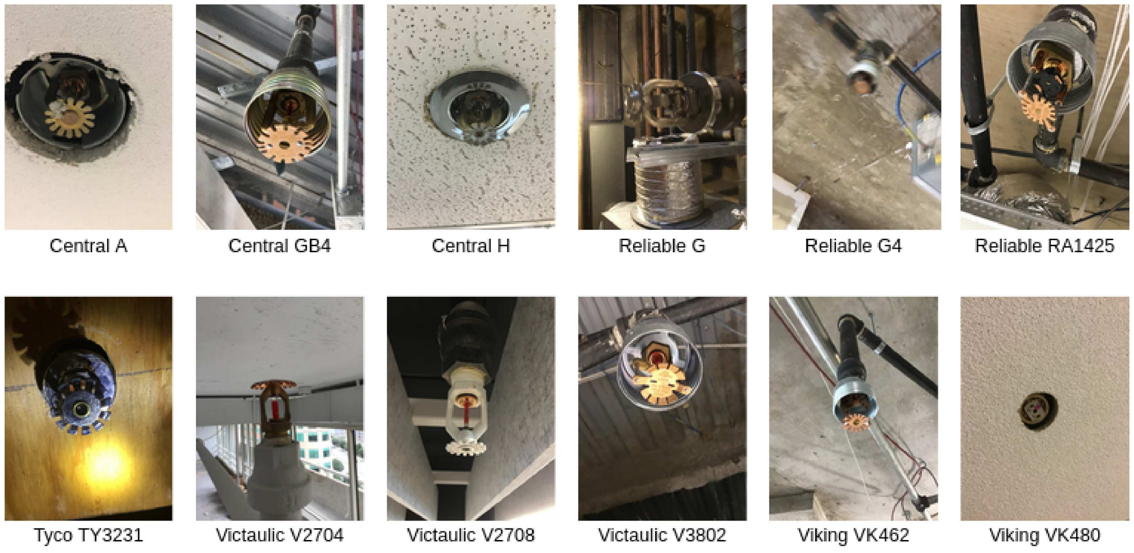

Figure 2. Sample of Sprinkler Dataset 5in

## 5. Models

### 5.1. From Scratch

The first model that was trained was a modified version of the Homework 5 CNN model that was trained on the Fashion-MNIST [5] to classify among 10 types of clothing. The architecture consists of 9 layers: 3 convolutional, 3 max pool, and 3 full-connected layers. This particular architecture was well-suited for the Fashion-MNIST classification problem which performed at 92% accuracy. It was thought that this performance could be carried over into a different domain with a slightly more difficult problem space since the increased fidelity of the quality of the sprinkler images.

### 5.2. ResNet

The Residual Neural Network [1] is a deep learning architecture that was proposed by a group of Microsoft researchers to improve the model training. It is implemented with skip connections called Residual Blocks which addresses the "vanishing gradient" problem found when increasing the number of layers in neural networks. The vanishing gradient problem occurs during backpropagation when the computed gradients shrink exponentially smaller, approaching zero, which can prevent the weights from updating and therefore slows the training or even completely halts. The residual block allows for data to pass not only to the immediately adjacent node but also a few steps deeper into the neural network. This effectively solves the vanishing gradient problem.

This particular model was chosen among many CNN architectures because it was common to pit it against the performance of the Vision Transformer. Specifically, the ResNet152 model was used. As its name suggests, this model has a total of 152 layers, 151 of which are convolutional layers and the last layer being the fully-connected layer.

### 5.3. Vision Transformer

The Vision Transformer (ViT) [2] is a rendition of the original Transformer proposed in "Attention is All You Need" [4] , that allows the encoding to be performed on image pixels rather than word tokens. The image is broken down into 16x16 patches as "image tokens" and passed into the encoder outputting into a Multilayer Perceptron Head which performs the image classification as illustrated in the figure below. One advantage to using a ViT over a CNN is that it may be computationally less expensive when trained on the same dataset. Another advantage is that while CNN's perform on localized portions of the image, the ViT can globally process the entire image which allows it to model the entire image as a whole where CNN's can only capture the relation with its immediate neighbors.

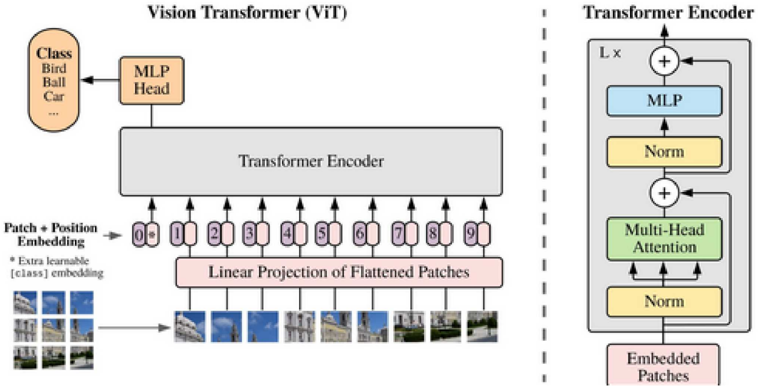

Figure 4. Vision Transformer Architecture

For this project, a pretrained model was fine-tuned to perform image classification on the sprinkler dataset. The ViT was trained on ImageNet-21k dataset which consisted of 14 million images among 12 classes. While the pretraining consisted of processing 16x16 patches of an image, the fine-tuned model takes 224x224 pixel images whole for training So required the sprinkler dataset to be resized and cropped to 224x224.

### 5.4. Hyperparameters

**Table 1. Hyperparameters**

| Model | Learning Rate | Optimizer | Epochs |
|---|---|---|---|
| CNN (Scratch) | 1e-3 | Adam | 20 |
| CNN (ResNet) | 1e-2 | SGD | 20 |
| Vision Transformer | 2e-4 | AdamW | 10 |

The table above shows the respective hyperparameters specifically set to fine-tune the models. One thing to note is that the Vision Transformer did not require many epochs of training to converge unlike the other two tested models. This may be the case because the pretrained ViT model was trained on the ImageNet-21k dataset, which consisted of 14 times more images and 20 times more classes. This possibly resulted in a pretrained model that is capable of learning with only a few training examples.

## 6. Results

**Table 2. Result Comparison**

| Model | Accuracy |
|---|---|
| CNN (Scratch) | 23.47% |
| CNN (ResNet) | 91.09% |
| Vision Transformer | 91.26% |

### 6.1. From Scratch

The model that was trained from scratch performed the weakest among the three models and as shown in the figure below, suffers from severe overfitting. This is likely the case because the architecture was not sophisticated enough to model more complex images. The architecture was modified from the Homework 5 CNN model and it was trained on a simpler image dataset. All the Fashion-MNIST images were black-and-white, had a black background, and all items were in a fixed position. In contrast, the sprinkler images were in full color, included the noisy background, and were taken at a convenient angle. This model would perform better if the network was deeper, introducing more layers as well as including residual blocks so as to avoid the vanishing gradient problem.

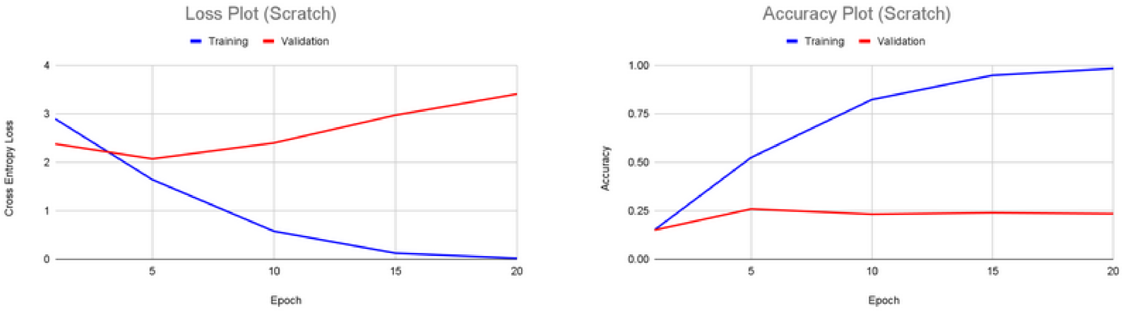

Figure 5. Performance of Model Trained from Scratch

### 6.2. ResNet

The ResNet performed well at around 91% accuracy and comparable to the Vision Transformer's accuracy. This model may have performed well possibly from easily finding patterns in the dataset. One such pattern is that although the sprinkler images contain noisy backgrounds, they mostly consist of two types: white ceiling or construction debris. Finding this pattern allows the model to isolate the model in question and more accurately perform classification. Another pattern the model may have captured is that some sprinkler models are manufactured in mostly one color, brass. The V2704 is an upright sprinkler head that is used in warehouse and empty spaces so tenants do not have a preference for color and so purchase these sprinklers at the cheapest price with the default brass color. The ability to recognize this pattern means that the model can more easily differentiate between models that only come in certain colors improving the classification accuracy. It exceeded the expected 80% accuracy and will be tested out in the field using real-world data.

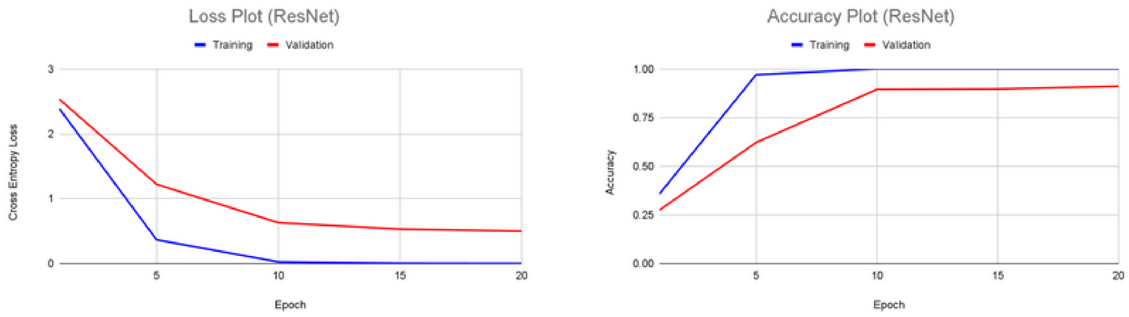

Figure 6. Performance of ResNet152

## 6.3. Vision Transformer

The Vision Transformer performed slightly better than the ResNet but the difference is miniscule and not meaningful. The only noticeable difference between the two is that it was trained on 10 epochs as opposed to the ResNet's 20. Though, this observation is not very meaningful because the ResNet also converges at around 10 epochs similar to the Vision Transformer's which converges at around 8 epochs. The same possible reasons for the ResNet's good performance can also be applied to the Vision Transformer. This model also exceeded the expected 80% accuracy and will be tested out in the field.

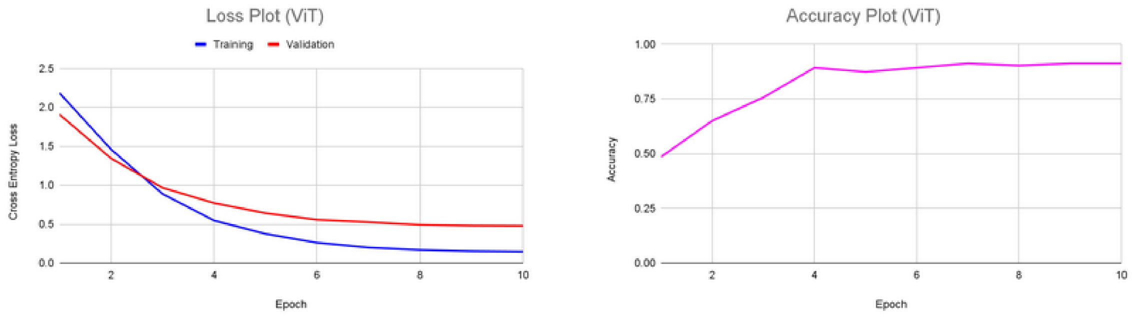

Figure 7. Performance of Vision Transformer

## 7. Future Work

Unlike other one-off projects that are quickly abandoned after submission, this project's development will continue well after the deadline since it will assist me in my job as well as offer assistance to other sprinkler surveyors who require a convenient sprinkler classification tool. The first way to improve the models is to increase the size of the dataset. Although the models trained well despite the small dataset, it would still improve the models accuracy with edge-cases found in real-world situations. For example, the dataset does not contain many images taken in dark environments. This can affect the accuracy of the classification model when performed on the RA1425 model in a dark room because the dataset only contains 8 images, none of which were taken in such dark environments. The second method to improve the models is to expand the number classes by supporting more sprinkler models. As mentioned before, during the process creating the dataset, if a particular model contained less than five samples, then they were discarded. Since the models were not trained on that particular model, it will not be able to classify when tested in the field. The solution is to add the class regardless of the number of samples and progressively increase the sample size for that model by including new samples with each new encounter. Lastly, an interface needs to be developed in order to communicate with the model in the field. Currently, the trained models can only be interfaced with using the commandline and so requires a laptop with a keyboard as well as a camera to take a picture and feed the model the image for classification.

## 8. Conclusion

The objective of this project was to train a few fire sprinkler classifiers with models that have different architectures and use the best performing model at work to help with the survey process. The results of this project exceeded expectations and produced two models that were highly performant and ready for testing in the field. Though the model trained from scratch suffered from overfitting and performed poorly, the ResNet and Vision Transformers models performed at 91.09% and 91.26% respectively. Expectations were exceeded and the performant models will be tested out with real-world data immediately.

## References

1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep Residual Learning for Image Recognition,” *CVPR* (2015).
2. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR* (2021).
3. Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya, *Improving language understanding by generative pre-training* (2018). OpenAI.
4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need,” *NIPS* (2017).
5. Han Xiao, Kashif Rasul, and Roland Vollgraf, *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms* (2017).