Name: Jason Zhou
Mentor: Dr. Dongjin Song
Status Report #: 6
Time Spent on Research This Week: 11.5
Cumulative Time Spent on Research: 33
Miles Traveled to/from Mentor This Week: 0
Cumulative Miles Traveled to/from Mentor: 0
==========

Monday, October 18th, 2021: (0.5 Hours)

On the first day of the new status report week, I had a Zoom meeting with my mentor. I was unsure of the scope of the project. For example, what types of sounds would we be analyzing (sounds can cover quite a large range from speech recognition to environmental sounds)? As such, I took this time to ask my mentor about the scope of the research project. He told me that we would just be focusing on the types of sounds that were provided in a machine learning community challenge, which is a task that inspired my project. For the most part, this includes the sound of machines like fans or cars.

Additionally, my mentor sent me some research articles related to my task for my literature search.

Wednesday, October 20th, 2021: (1.5)

On this day, I began looking through the four articles that my mentor gave me.

The first article is called *Anomaly Detection: A Survey*. This article aimed to take the different types of anomalous detection and categorize them based on the underlying algorithm that was used. For each method, it also lists the advantages, disadvantages, and computational complexity (how taxing the process is on the computer). However, it is worth noting that this article deals with anomalous detection as a whole, not just with sound. For example, anomaly detection could be used to detect suspicious credit card payments. This article is 72 pages. As such, I will not be reading all of it this week, but will most likely break up my reading over the span of several weeks.

After glancing at these articles, each of them appears to be well over 30 pages. Additionally, upon further investigation, these articles are actually textbooks to help me understand anomaly detection, how to do it, and its various applications to the real world. As such, I will read them at a later date and continue my literature search.

In order to gather more literature, I thought it would be a good idea to go to DCASE 2020 website, which is where the original challenge my research project is based on comes from. On this website, they have a list of all the people who participated and their corresponding research and technical reports. Additionally, the website also categorizes the articles based on the

method used. For the next hour, I went down the list of methods, searched for the corresponding article in Google Scholar, tagged it, and placed it in my library. I will have to read through these later and determine how useful they will be to my research project later.

Looking through all of these articles, I found a large variety of methods that were used to process audio data. These methods included:

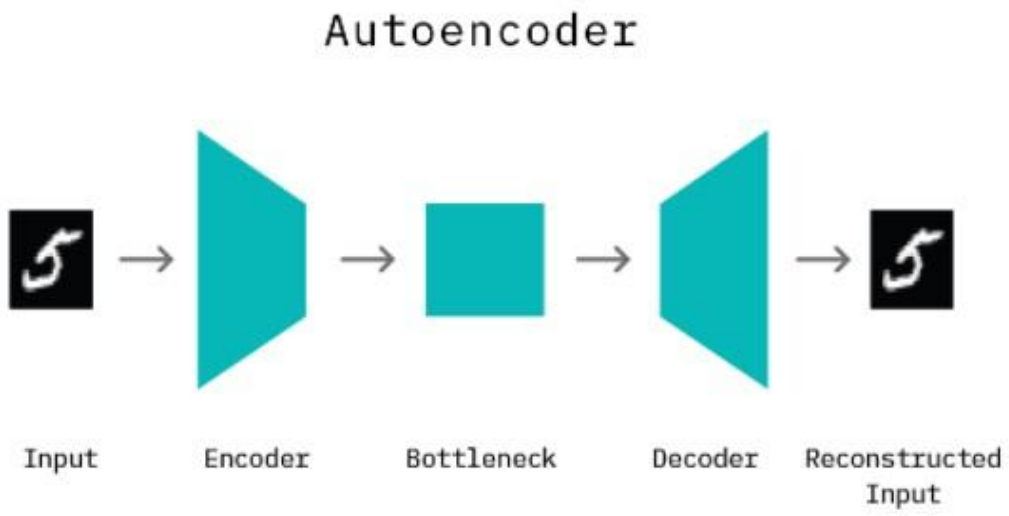
- Autoencoders
- Convolutional Neural Networks
- Conformers
- Dictionary Learning
- Gaussian Mixture Models
- Heteroskedastic Variational Autoencoders
- ID Embedding
- ID Regression
- Interpolation Deep Neural Networks
- L2-Softmax
- Large Margin Cosine Distance
- Local Outlier Factor
- Long Short Term Memory Autoencoders
- Mahalanobis Distance
- Matrix Normal Distribution
- MobileNetV2
- Normalizing Flow
- One Class Support Vector Machines
- Principal Component Analysis
- Phase Shift Prediction
- Probabilistic Linear Discriminant Analysis
- ProtoPNet
- ResNet50
- Regularized Linear Discriminant Analysis
- Semi-Supervised Autoencoders
- Stats Pooling
- Subspace Distance K-Nearest Neighbors
- Transformers
- U-Net Autoencoders
- Variational Autoencoders

As I have only heard of a handful of these methods, I will have further research on the different ways of processing audio data.

Friday, October 22nd, 2021: (1.5 Hours)

Before reading any of the articles I collected on Wednesday, I need to be able to understand everything that is being written. As such, I began reading the online textbook my mentor sent me.

The first article I read was called CONFORMER-BASED ID-AWARE AUTOENCODER FOR UNSUPERVISED ANOMALOUS SOUND DETECTION. The authors and researchers of this article used a Transform and Conformer-based autoencoder for anomalous sound detection (ASD). Additionally, this article also proposed two methods for differentiating between ID information: ID embedding and ID regression. I am not quite sure what these are, so I will have to look into this further. Thus, I am not sure of the role ID embeddings and ID regression had on the process. However, as for the rest of the process, the researchers placed the data into an autoencoder. The data was deconstructed and reconstructed. Then an anomaly score was produced. The higher the anomaly score, the higher the likelihood that the audio file is an anomaly.



(Picture of an autoencoder. I have included it here just in case it was hard to visualize exactly what an autoencoder was.)

Saturday, October 23rd, 2021: (5 Hours)

On this day, I continued looking through my articles, learning the necessary concepts as I went. I began reading an article called UNSUPERVISED DETECTION OF ANOMALOUS SOUND FOR MACHINE CONDITION MONITORING USING DIFFERENT AUTO-ENCODER METHODS. Specifically, this paper focuses on four different kinds of deep learning methods: Long-short term memory autoencoder (LTSM AE), U-Net autoencoder, Interpolation deep neural network (IDNN), and a fully-connected autoencoder(FC AE). The article takes the Mel

Frequency Cepstral Coefficient, Short-Time Fourier Transform, Chroma Features, Mel Spectrogram, Spectral Contrast, and Tonnetz as features to create the average feature and then constructs a feature vector with them. Although I do not have a complete understanding of the above features I just mentioned, these are just characteristics of sound data. Also, a feature vector is simply just a list of numbers (the numbers represent features, for example, height or width). For the next feature, a log-mel spectrogram is taken. Both features will act as inputs for the model. Overall, based on the data, the models that were used were higher than the baseline (control). This means that their proposed models, IDNN, LSTM AE, FC AE, and U-Net AE all performed better than the control.

The next article I viewed was called TASK 2 DCASE 2020: ANOMALOUS SOUND DETECTION USING UNSUPERVISED AND SEMI-SUPERVISED AUTOENCODERS AND GAMMTONE AUDIO REPRESENTATION. This article uses a convolutional autoencoder[1] as the model. Furthermore, in the DCASE 2020 challenge, there are multiple types of machines (fan, toy car, valve, etc) included in the dataset. Normally, other articles have tried creating a specialized model for each machine because this would lead to more accurate results; however, the researchers of this article have made an overarching model that seems to work on all machines in the study. As for the results, all of the proposed models (unsupervised and semi-supervised models) appear to perform above the baseline

I then moved onto a research study titled UNSUPERVISED DETECTION OF ANOMALOUS SOUNDS BASED ON DICTIONARY LEARNING AND AUTOENCODER. This article proposes two methods for achieving the DASE 2020 task 2 challenge. Specifically, they are using auditory traditional features and dictionary learning (DL). Second, they are using auditory spectral features combined with an autoencoder.

Afterward, I viewed an article titled AN ENSEMBLE APPROACH FOR DETECTING MACHINE FAILURE FROM SOUND. This paper aimed to use an ensemble[2], which consists of autoencoders, a self-predictive heuristic[3], and a Gaussian mixture model[4] (GMM). The

---

[1] A convolutional autoencoder is simply an encoder that uses convolutional layers. Convolutional layers are layers in a neural network that apply a filter to the input data to generate an output feature map. Typically, convolutional layers are used to process images; however, this article has applied them to audio files instead to yield interesting results.

[2] A combination of models/algorithms

[3] A heuristic is a method of doing something that is not always the most optimal method, but is guaranteed to be a fast solution to the problem at hand. For example, if I am at a fast food restaurant trying to decide what to buy and there are a line of people behind me waiting, then I might decide to go with what I always get to save time and not keep people waiting.

[4] Gaussian Mixture Models (GMM) are just a combination of Gaussian distributions. Gaussian distributions are the bell-shaped curves that one might see when calculating the probability of something or taking the standard deviation of a dataset. As such, if each Gaussian distribution represents a cluster of data (a dataset), then a GMM simply gives the probability that a selected point will fall under one of the given clusters. To give an example, imagine two circles that merge together to make a venn diagram-like shape. If the selected point is in the left circle (the part that is not overlapping with the right circle), then the probability of it being in the left circle is 100%. However, if the selected point is within the overlap, the probability of it belonging to the left or right circle could be 50% for both circles.

autoencoder is based off of a multilayer perceptron [5]. It uses ReLU[6] as the activation function and Adam as the optimization function[7]. The classifier uses a 6-way classifier[8] and a convolutional ResNet classifier to predict machine type. The GMM uses 34 short-term audio features to calculate the various possibilities of a sound being anomalous or not. Additionally, all of these models performed above the baseline.

The next article was called ENSEMBLE OF AUTO-ENCODER BASED AND WAVENET LIKE SYSTEMS FOR UNSUPERVISED ANOMALY DETECTION. First, it uses a U-Net to remove the background noise on training samples. Then, this data is inputted into an autoencoder to process and classify the data. Also, these authors attempted to simplify the learning process by using an ID conditioned autoencoder combined with machine IDs. Next, using an autoencoder, the data is inputted and features extracted. Basically, the data was turned into a smaller version of its original size. Afterward, it is reconstructed. The reconstructed data is compared to the original input and a reconstruction value is calculated. The higher the reconstruction value the more likely it is that the sound is anomalous.

After viewing that article, I moved onto another one named AUTO-ENCODER AND METRIC-LEARNING FOR ANOMALOUS SOUND DETECTION TASK. These researchers tried an autoencoder based approach but found that tuning the hyper parameters led to minimal performance increases. However, using the encoders as a foundation, they applied an algorithm called metric learning[9] to extract features. Afterword, the extracted features are passed through an L2-softmax function[10]. Finally, a method called local outlier factor[11] is used to obtain the

---

[5] A multilayer perceptron, from what I understand, is a way to classify data in a non-linear way.
[6] ReLU is an activation function that simply places the values of the neural network into a range that can be easily understood or used by the model.
[7] Optimization functions help to optimize the model to obtain a smaller loss function.
[8] A classifier that classifies data into one of six different categories. Hence it is called a 6-way classifier.
[9] From what I understand, metric learning is a type of machine learning that aims to create new metrics such that similar values are placed next to each other while different values are placed far away. Ultimately, the purpose of creating a new metric is to create a unit of measure that is useful for the model. Specifically, a good metric should allot the model to easily find and recognize patterns within the data. In machine learning, this is especially important because in higher dimensional data (8D, 9D, 10D, etc), the Euclidean metrics (inches, meters, etc) that people usually use become useless. In simpler terms, it becomes hard to determine distance using metrics that people would usually use.
[10] L2 refers to normalizing or standardizing the data. It is important to standardize the data so that all features are on a common scale and can, thus, be weighted accordingly. For example, if one is measuring the number of rooms and the surface area of an apartment room to determine price, the number of rooms would typically range between 1-4, while the surface area would be somewhere in the hundreds. Because the surface area will typically be in the hundreds, in a neural network, this value would have a larger influence on the output due to its magnitude. However, what if the number of rooms actually has a larger influence on the price than the number of rooms. This is why one would need to normalize the data: it gives the data a chance to be looked at equally (one is not inherently more important than the other just because it's value is larger). Besides L2, a softmax is a function that calculates the probability of a data point being one of three or more classes. As such, an L2-softmax function normalizes the data and then calculates a probability distribution.
[11] The local outlier factor is an algorithm used to calculate the outlier in a cluster of data. This is relevant for anomalous sound detection because the anomalous sound will almost definitely be an outlier within the data.

anomalous score. They have outlined more to their proposal such as the mahalanobis distance algorithm; however, because I am not sure how this works, I did not include it in the summary.

Because I did not understand much from the last article, I started on a new article, which was called ENSEMBLE OF AUTO-ENCODER BASED AND WAVENET LIKE SYSTEMS FOR UNSUPERVISED ANOMALY DETECTION. Overall, this paper covers various methods like variational autoencoders (VAE), ID Conditioned autoencoders (IDCAE), and a WaveNet like network. The variational auto-encoder is used to find errors during backpropagation. The IDCAE uses the machine IDs as labels during the classification task. If the audio file does not match the expected machine IDs audio file, then it is likely an anomaly. As for the WaveNet-like network, this algorithm uses layered convolutional networks to predict the next frame of sound (interval of sound). The predicted frame is compared with the actual frame. An anomaly score is calculated. If the predicted frame is different from the actual frame, then the score will be high.

Upon finishing the last article, I began reading another, which was named ANOMALY CALCULATION FOR EACH COMPONENTS OF SOUND DATA AND ITS INTEGRATION FOR DCASE 2020 CHALLENGE TASK2. THis paper focuses on mahalanobis distance and k-nearest neighbors (KNN). Although I am unsure what mahalanobis distance is, it seems like the paper mainly uses KNN as the main algorithm. The data is used to create mel-spectrograms. Then, these features are extracted and plotted on graphs to use KNN. By clustering the data points, an anomaly score can be calculated for a point based upon its distance from a cluster (indicating that it is most likely an outlier).

As the final research paper of the day, I read UNSUPERVISED DETECTION OF ANOMALOUS SOUNDS VIA PROTOPNET. As the name suggests, the paper focuses on using a ProtoPNet for sound anomaly detection. Although ProtoPNet is mainly used for image classification, the authors of this paper have attempted to apply the algorithm to sound anomaly detection. In general, the data was transformed into mel-spectrograms and inputted into multilayer convolutional layers (which is the ProtoPNet). There seemed to be an abundance of complex math in this paper that I did not understand. As such, besides the general process I was not able to comprehend anything else. I will have to look further into this article at a later date.

Sunday, October 24th, 2021: (3 Hours)

Continuing from yesterday, I started with an article called UNSUPERVISED ANOMALOUS SOUND DETECTION USING SELF-SUPERVISED CLASSIFICATION AND GROUP MASKED AUTOENCODER FOR DENSITY ESTIMATION. This paper uses an ensemble (a collection) of algorithms that rely on calculating density to discern anomalies within the data such as Group Masked Autoencoder for Density Estimation (GMADE) and a self-supervised classification anomaly detector. For GMADE, the data is run through a negative

log-likelihood function[12]. As for the self-supervised classifier, two different types of architecture (algorithms) were used: MobileNetV2 and ResNet50. Then, a cross-entropy softmax[13] layer is used to calculate the loss. The negative of this value is taken as the anomaly score.

Afterward, I read a paper called A SPEAKER RECOGNITION APPROACH TO ANOMALY DETECTION. Just like previous researchers, this group of researchers used an autoencoder to process the data and make predictions. However, unlike other projects, this one combines a deep neural network (which is an autoencoder) (DNN) with a speech recognition algorithm to calculate anomaly scores with sound. As for the architecture of the DNN, it has a spectrogram layer[14],an encoder, and ends with a fully connected margin softmax layer. Additionally, it may include an x-vector model; however, the authors have said that this layer is optional. Overall, this model succeeded in scoring above the baseline in 5 out of the 6 types of machines. When it came to the toy car (the last type of machine), the authors say that their proposed model did not achieve good results.

The next article I read was titled ANOMALOUS SOUND DETECTION BY USING LOCAL OUTLIER FACTOR AND GAUSSIAN MIXTURE MODEL. As the name suggests, the authors of this paper attempted to use local outlier factor (LOF) and gaussian mixture models (GMM) to detect sound anomalies. Out of all the papers I have read so far, this one is particularly interesting because it attempts to achieve anomaly detection without the use of deep learning. This means that methods like autoencoders, which are the most commonly used method for anomaly detection in general, are not used within this paper. The model they have outlined splits the audio file into different parts and derives a feature from each part. Then, an anomaly score is calculated for each section of the audio file using LOF or GMM. Afterward, the various anomaly scores are brought together to create one singular, overarching, anomaly score. Overall, their methods performed above the baseline. This is significant because it shows that deep learning methods, which have been widely considered more efficient, are not always better.

Moving on, I began reading ANOMALOUS SOUND DETECTION WITH MASKED AUTOREGRESSIVE FLOWS AND MACHINE TYPE DEPENDENT POSTPROCESSING. In this research, a Masked Autoregressive Flow (MAF). Essentially, it is an algorithm that works off of density-based estimations similar to algorithms that I have seen in the past like LOF or GMADE. The audio is split into multiple parts and turned into 10-second long spectrograms. Then, the spectrograms are normalized and passed into the flow-based model[15].

---

[12] A negative log likelihood function is simply a loss/cost function. Essentially, it calculates the error of the model. The higher the loss function value, the more errors there are. Thus, one should try to minimize the loss function value to get the most accurate model possible.

[13] Just like negative log likelihood, cross entropy is a type of loss/cost function.

[14] This is a layer that converts the waveform into a spectrogram, a visual representation of sound. This transformation of data is important because spectrograms allow later layers of a DNN to process data and extract features.

[15] A flow-based model is a model that relies on normalizing flow as a way to process the data and make predictions. At the most basic level, normalizing flow takes math equations and continuously takes the inverses of those equations to look at data through different transformations and dimensions.My guess is that this acts as a kind of data augmentation to train the model to be more accurate. Data augmentation is

The final article I looked at today was called MODULATION SPECTRAL SIGNAL REPRESENTATION AND I-VECTORS FOR ANOMALOUS SOUND DETECTION. This paper proposes two methods for anomaly detection: K-Nearest Neighbors (KNN)[16] and the extraction of i-vectors[17] to train a Gaussian Mixture Model (GMM). Overall, the algorithms performed around 6% better than the baseline, and the authors found that an ensemble (a collection or combination) of the two methods yielded even better results. First, the audio goes through preprocessing. Essentially, the noise[18] within multiple audio clips are taken and are averaged together. Next, a fast Fourier transform[19] (FFT) is calculated using this average noise clip. The information gathered from the FFT is analyzed and used to create a threshold for a frequency band[20]. This means that an acceptable range of frequencies is created based on the FFT. Afterward, another FFT is used on the audio files (which have their noise removed). This new FFT is then compared to the threshold value and a mask [21]is applied to the signal. Finally, the signal is inverted, which ends the pre-processing step. The actual processing step is fairly similar to other articles I have read. The data is given to a model (in this case, it is GMM or KNN) and a reconstruction error value is calculated, which is synonymous with the anomaly score.

References

Conv2D layer. (n.d.). Keras. Retrieved October 25, 2021, from https://keras.io/api/layers/convolution_layers/convolution2d/.

Dense Layer. (n.d.). Keras. Retrieved October 25, 2021, from https://keras.io/api/layers/core_layers/dense/

Favaits, M. (2020, November 9). *Tutorial | anomaly detection | local outlier factor | LOF algorithm* [Video]. YouTube. https://www.youtube.com/watch?v=CePgbdVdLvg

when more data is created by modifying existing pieces of data. For example, a picture might become 4 different photos by rotating it or reflecting the original image. Ultimately, I will have to look into this further.

[16] K Nearest Neighbors is essentially the same as K-means. The algorithm relies on clustering in order to determine an anomaly, which would usually be an outlier or a point in the data that is not within a cluster.

[17] I-vectors are fixed-length vectors (vectors whose lengths do not change) that encapsulate an abundance of data into a low-dimensional representation. This makes it easy for models to process the data.

[18] Noise is a general term for background noise or any sound that is not the focus of the audio clip/data.

[19] A fast Fourier transform is an algorithm that breaks down an audio file into its components. This means that the audio files are broken down into the frequencies that make them up.

[20] In layman's terms, a frequency band is a range of frequencies. For example, it could have a range of 3-30 kHz or 30-300 kHz.

[21] A mask or "masking" is a way to smooth the processing of data. Oftentimes, in data, certain inputs are missing. Essentially, a mask tells the model to skip those inputs and continue processing data regardless.

Frequency domain. (n.d.). DeepAI. Retrieved October 25, 2021, from https://deepai.org/machine-learning-glossary-and-terms/frequency-domain

Jayaswal, V. (2020, August 30). *Local outlier factor (LOF) -- algorithm for outlier identification*. Towards Data Science. Retrieved October 25, 2021, from https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843

KAYA, M., & BILGE, H. S. (2021, August 21). *Deep metric learning: A survey*. MDPI. Retrieved October 25, 2021, from https://www.mdpi.com/2073-8994/11/9/1066/htm

Mandal, M. K. (2017, September 13). *Implementing l2-constrained softmax loss function on a convolutional neural network using tensorflow*. Manash's Blog. Retrieved October 25, 2021, from https://blog.manash.io/implementing-l2-constrained-softmax-loss-function-on-a-convolutional-neural-network-using-1bb7c0aab7b1

Metric learning for image similarity search. (n.d.). Keras. Retrieved October 25, 2021, from https://keras.io/examples/vision/metric_learning/

Normalizing flows. (n.d.). Papers With Code. Retrieved October 25, 2021, from https://paperswithcode.com/method/normalizing-flows#:~:text=Normalizing%20Flows%20are%20a%20method,the%20sequence%20of%20invertible%20mappings.

NTS. (2012, October 10). *FFT basic concepts* [Video]. YouTube. https://www.youtube.com/watch?v=z7X6jgFnB6Y

What is metric learning? (n.d.). Metric Learn. Retrieved October 25, 2021, from http://contrib.scikit-learn.org/metric-learn/introduction.html

Why do we normalize the data. (n.d.). Quora. Retrieved October 25, 2021, from https://www.quora.com/Why-do-we-normalize-the-data