Weekly Report

Wangwon Lee, 2019/03/23

This week

- **New try in architecture**
  - Change the channel more steeply
  - What if the channel become smaller and smaller?
  - Summary

Next week

- **Audio Classification**
  - Understanding 1D-CNN model.
  - Mapping our brain data
  - Investigate 1D model more specific
- **Visualization**
  - Filter and Feature map
  - In frequency area
  - CAM(Class Activation Map)

Interesting and new finding

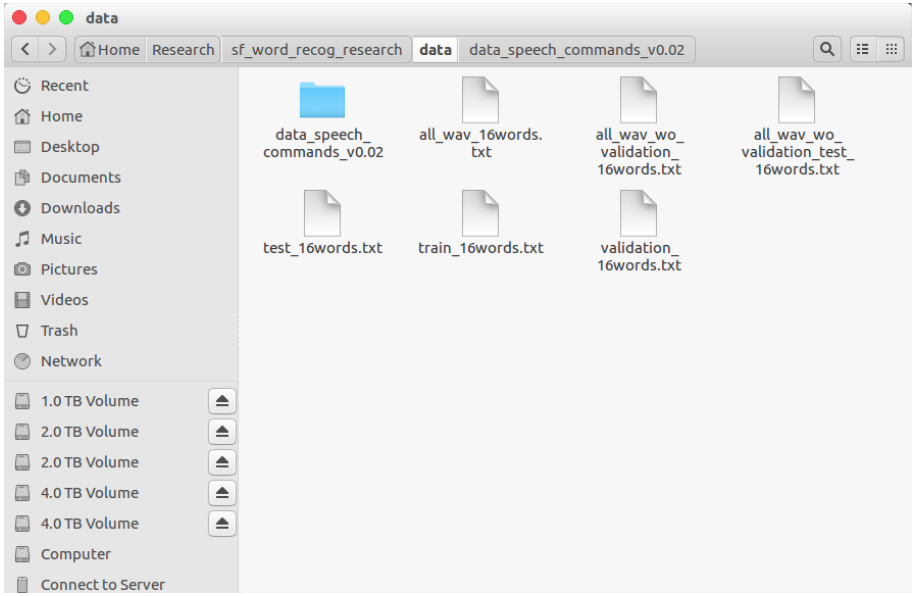
- Fine Tuning
- Feature Extraction

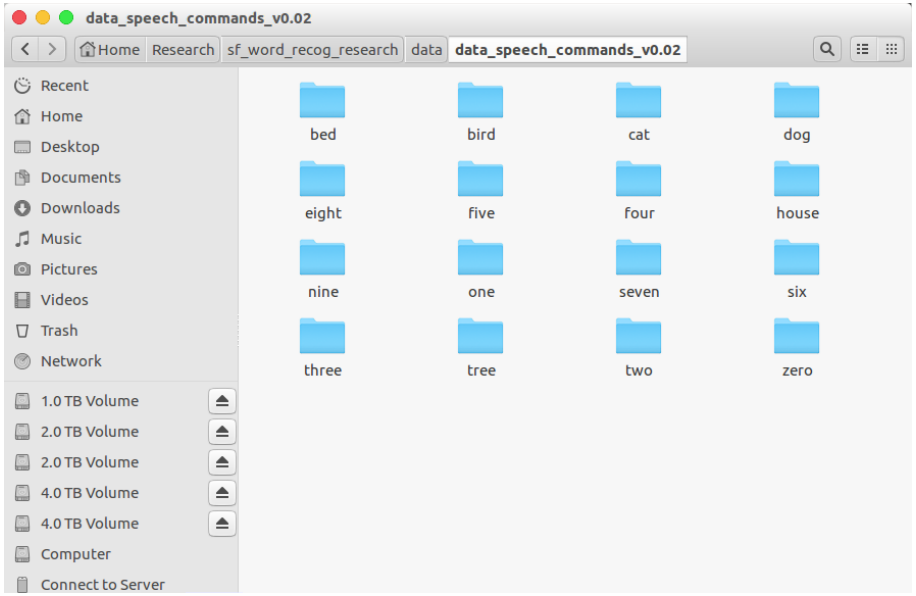
The aim of this month / Discussion

The aim of this month: To study the brain and GLM, To investigate about CNN.

Audio Classification - Previous Work

- Data is low-waveform.
  - sec: 1, sampling rate: 16000, type: float32, channel: mono
- 16 class data.
  - 'zero', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'bed', 'bird', 'tree', 'cat', 'house', 'dog'
- Train: 40851(≒80%), Validation: 4796(≒10%), Test: 5297(≒10%)

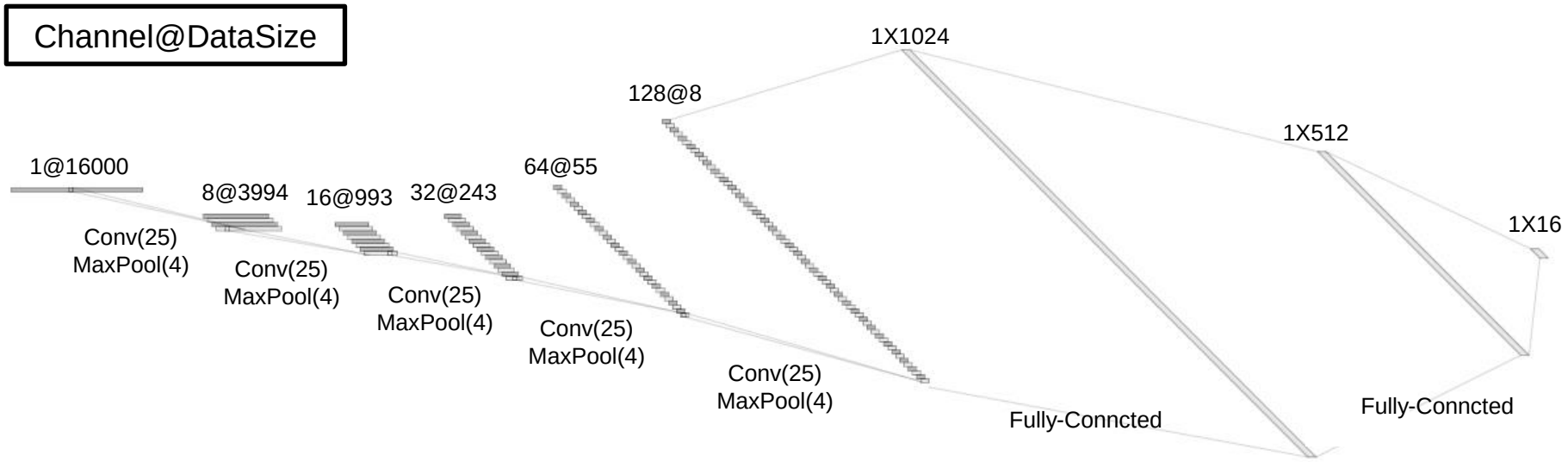




Audio Classification - Previous Work

- For example, ‘5Conv, 2FC’ baseline model’s detail.
- It just flatten 2D model. (5X5 filter->1X25 filter, 2X2 stride->1X4 stride)

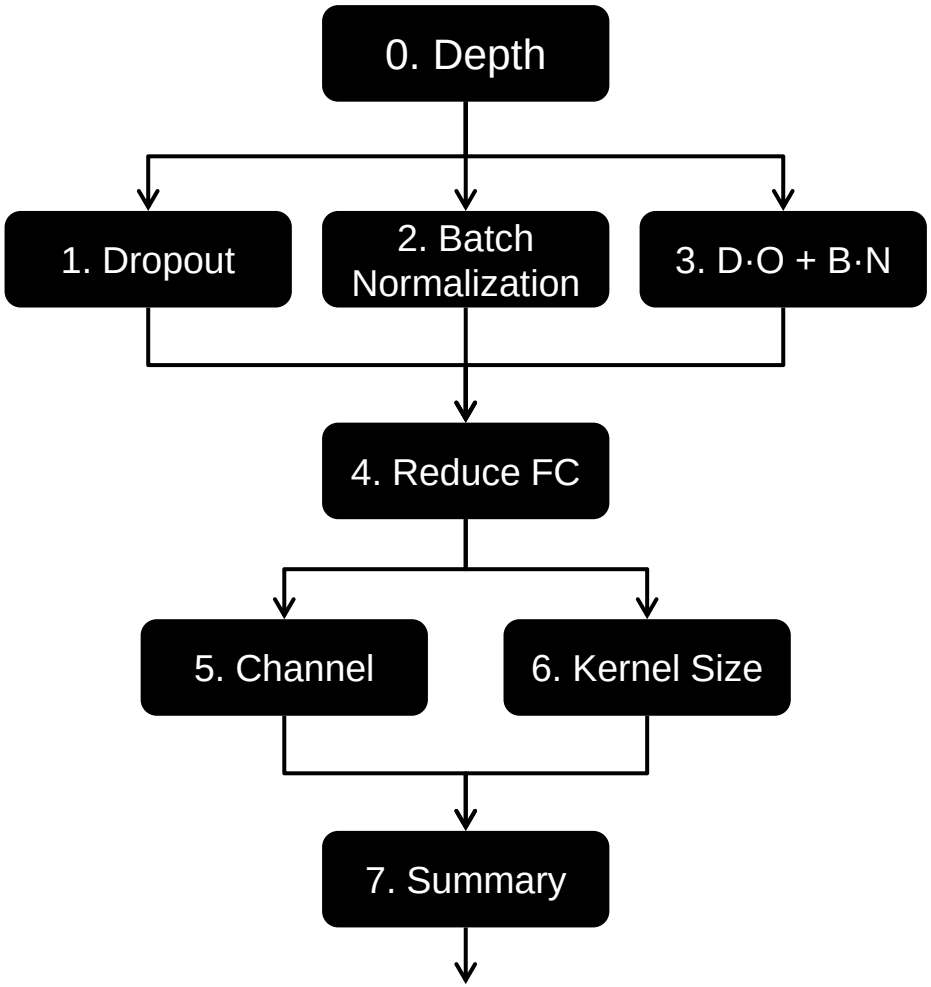
- Input: 16000X1 low waveform.
- Output:1x16 labeled one hot vector. ('zero', …, 'eight', …, 'house', 'dog')
- Loss: cross entropy loss
- Obtimizer: Adam

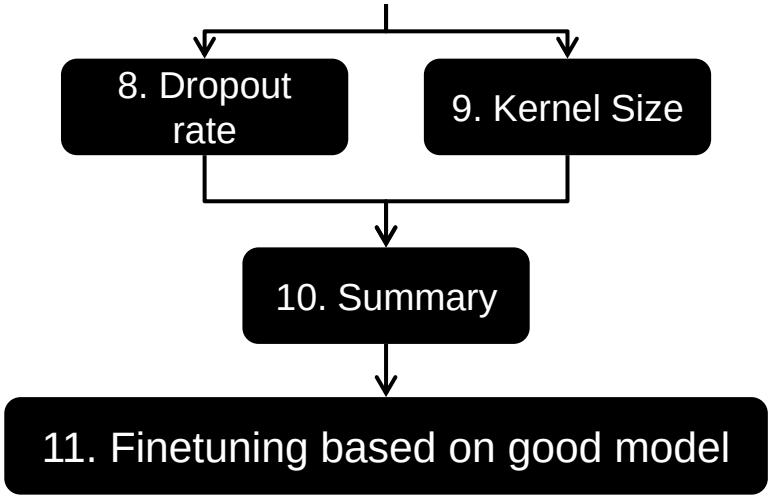


[Baseline Architecture]

Audio Classification - Previous Work

- Fine tuning task in 1D-CNN





Audio Classification

- This is SOTA(State Of The Art) in current research.

	Architecture (i = 0,1,2…)	1D DO(0.5)	1D BN	1D DO+BN	Params
	baseline				
base model	5 Conv(25, $8*2^i$), 5 Pool(4), 2 FC	0.9090	0.9072	0.9240	1,855,056
	Accuracy and Number of parameters				
Custom channel 32 DO(0.75)	8 CONV(5, 64)	0.9533	0.9285	0.9391	94,768
Custom channel 64 DO(0.75)	8 CONV(5, 128)	0.9589	X	0.9497	363,600
Custom VGG style DO(0.75)	16 CONV(3, 128) , 8 Pool	0.9620	0.9423	0.9136	470,736
	Only Accuracy				
Custom channel 128 DO(0.25)+BN	9 CONV(5, 512)	0.9535	X	0.9701	2,071,184

Audio Classification

- Confusion matrix
- Compare two best model.
- Not much different, but the right model is a little better.

Actual class

Predict Class

	Zero	One	Two	Three	Fore	Five	Six	Seven	Eight	Nine	Bed	Bird	Cat	Dog	House	Tree
Zero	369	0	5	1	3	0	1	4	0	1	0	0	1	0	0	0
One	1	347	0	0	4	1	1	0	0	7	2	0	0	1	0	0
Two	8	0	369	2	0	0	0	0	1	0	0	0	2	2	0	0
Three	1	0	2	356	0	1	1	2	6	0	0	0	0	0	0	8
Fore	2	2	1	1	359	2	0	0	0	0	0	0	1	0	0	0
Five	0	6	0	3	4	386	0	2	1	3	0	1	2	0	0	0
Six	0	0	0	3	1	0	366	1	2	0	1	0	0	0	0	0
Seven	3	0	0	0	1	0	2	367	0	0	3	0	0	0	0	0
Eight	1	0	1	2	1	0	0	0	367	0	2	0	1	0	0	1
Nine	0	6	0	0	0	3	0	0	0	364	3	1	0	0	0	0
Bed	1	0	2	0	0	0	0	0	6	0	166	5	2	1	0	0
Bird	0	1	1	1	0	0	2	0	0	2	7	137	0	2	0	0
Cat	0	0	0	0	0	0	0	1	0	0	1	0	161	4	1	0
Dog	0	1	5	0	0	0	0	0	0	1	1	0	2	182	0	0
House	0	0	2	0	0	1	1	0	0	1	0	0	5	1	156	0
Tree	2	0	2	15	0	0	1	0	4	1	0	0	0	0	0	138

Custom channel 32 DO(0.75) **(Acc: 0.9533)**

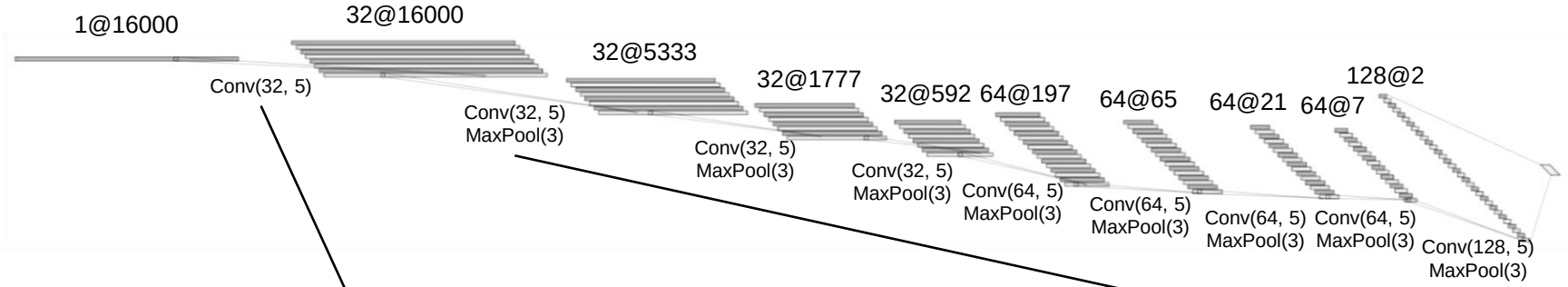
Actual class

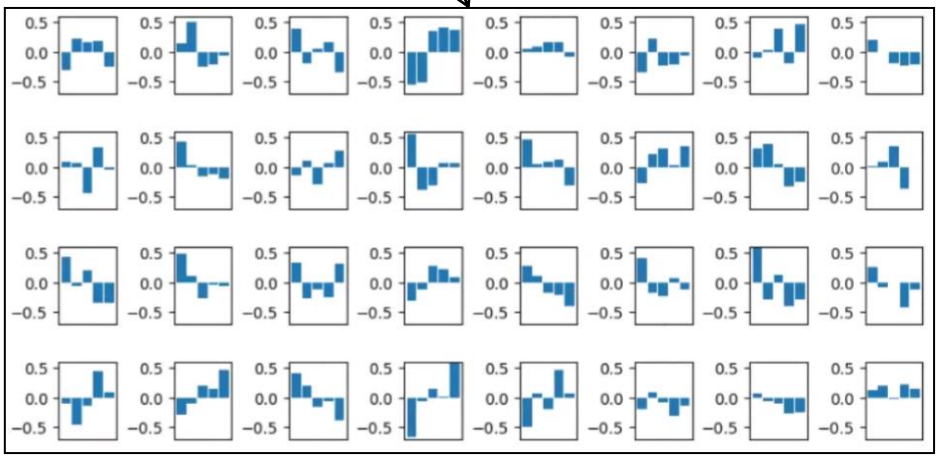
	Zero	One	Two	Three	Fore	Five	Six	Seven	Eight	Nine	Bed	Bird	Cat	Dog	House	Tree
Zero	375	0	3	0	2	0	0	3	0	1	0	0	0	0	0	1
One	0	348	0	0	4	1	0	1	1	6	1	0	0	0	1	1
Two	1	0	375	2	1	0	0	0	0	0	0	0	0	2	1	2
Three	1	0	5	356	0	0	2	3	1	0	1	0	0	0	0	8
Fore	0	0	0	0	364	1	0	0	0	0	0	1	1	0	0	1
Five	0	0	0	0	5	399	0	0	1	2	0	0	1	0	0	0
Six	0	0	0	1	0	0	371	1	0	0	1	0	0	0	0	0
Seven	1	0	0	0	0	0	1	373	1	0	0	0	0	0	0	0
Eight	0	0	4	2	0	0	0	0	362	4	0	2	1	0	0	1
Nine	0	5	0	0	0	3	0	1	0	363	2	3	0	0	0	0
Bed	0	0	1	0	0	0	1	0	1	0	178	1	0	0	0	1
Bird	0	0	0	0	0	0	0	0	0	1	3	147	0	1	0	1
Cat	0	0	0	0	0	0	0	0	0	0	0	0	166	1	1	0
Dog	0	0	2	0	0	0	0	0	0	0	1	1	0	186	0	2
House	0	0	0	0	0	0	0	0	2	0	0	1	3	0	161	0
Tree	0	0	1	10	0	0	0	0	4	1	0	0	0	0	0	147

Custom channel 128 DO(0.25)+BN **(Acc: 0.9701)**

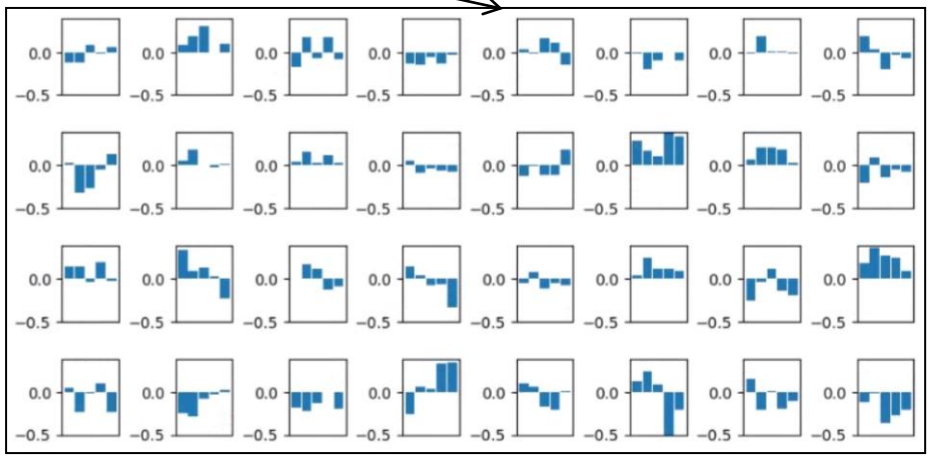
Audio Classification

- Visualize the filter map. (Custom channel 32 DO(0.75) Model)
- Of course, Small number of parameters is easy to analyze
- There was a shape to know, but most of the shape was hard to understand.
- Next time, I will prepare the feature map and analyze it more detail.





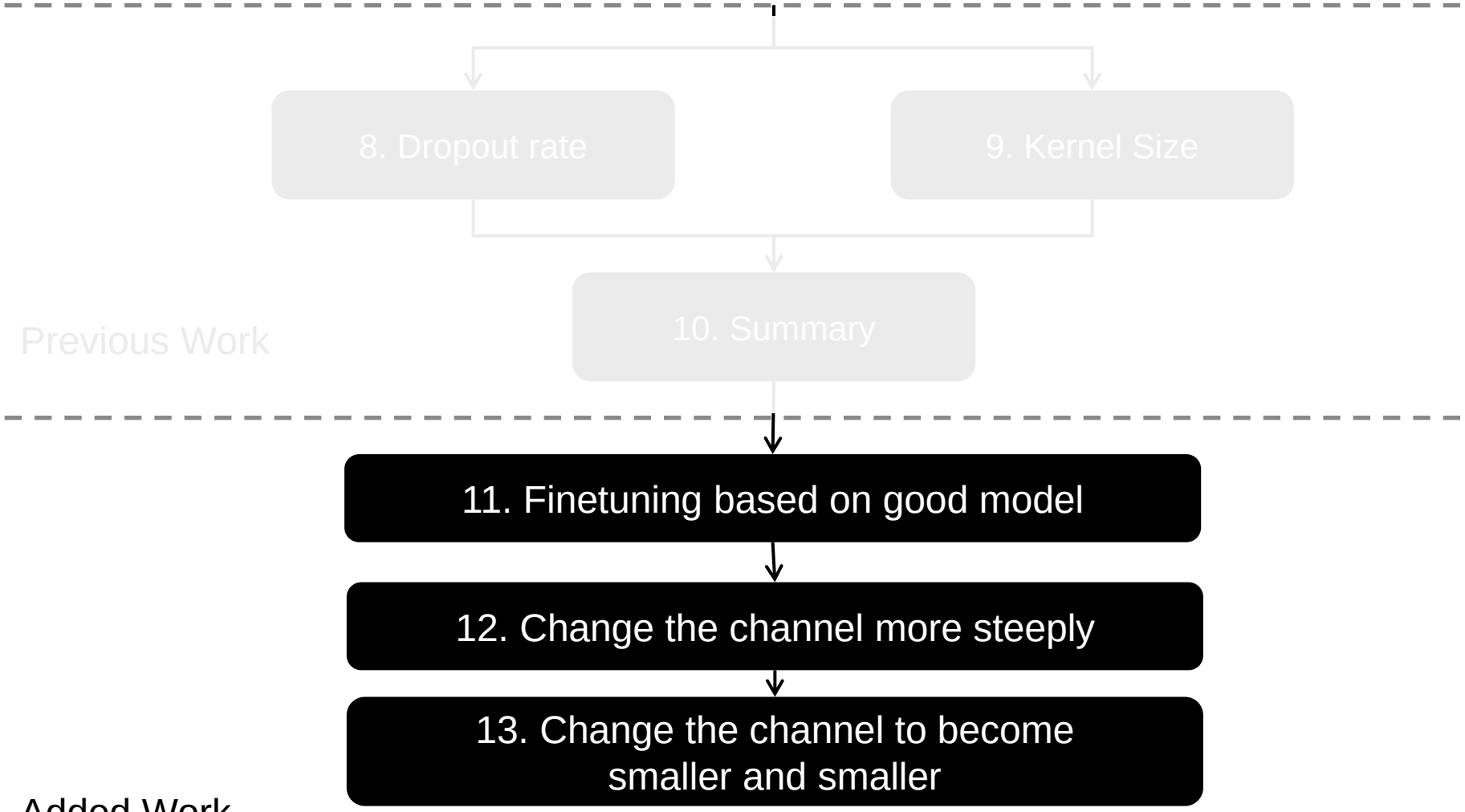
First layer's filter



Second layer's filter

Audio Classification - Added Work

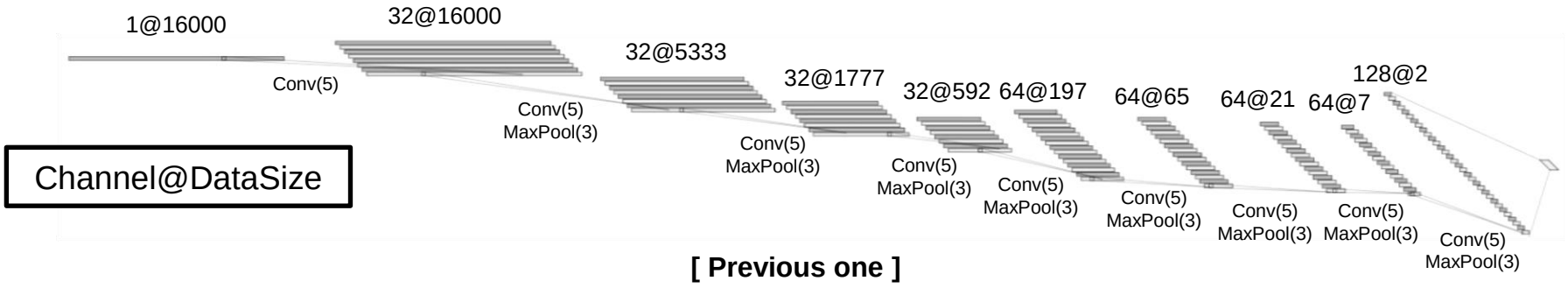
- Fine tuning task in 1D-CNN



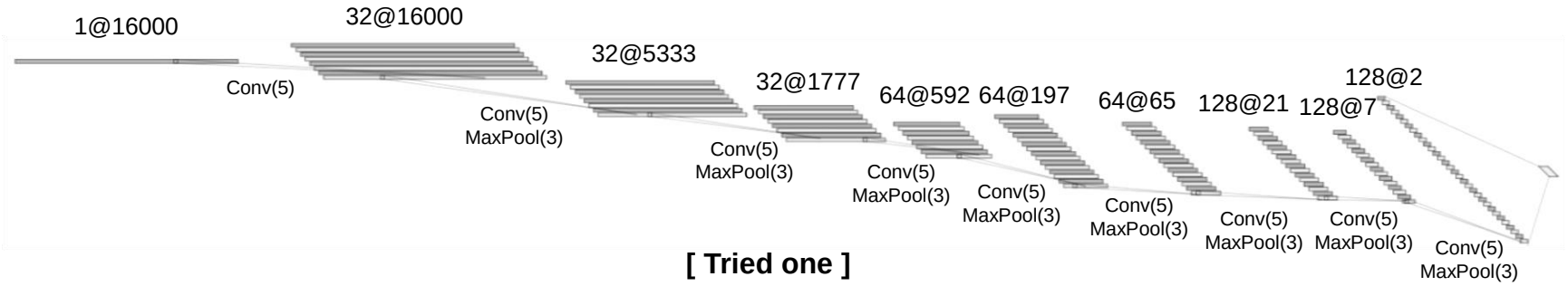
Audio Classification

- Change the channel more steeply
- The upper is ‘Custom channel 32 DO(0.75) 9 Conv’ model’s detail.

- Previous one: 1 → 32 → 32 → 32 → 32 → 64 → 64 → 64 → 64 → 128
  Tried one : 1 → 32 → 32 → 32 → 64 → 64 → 64 → 128 →128 → 128



[Previous one]



[Tried one]

Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 32

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 32)	X	X	X	X
2 CONV(5, 32)	X	X	X	X
3 CONV(5, 32)	X	X	X	X
4 CONV(5, 64)	0.6339	0.6021	0.6982	627,024
5 CONV(5, 64)	0.7701	0.7020	0.8027	243,088
6 CONV(5, 64)	0.8856	0.8224	0.8739	128,464
7 CONV(5, 128)	0.9319	0.9034	0.9242	146,000
8 CONV(5, 128)	0.9479	0.9267	0.9452	199,376
9 CONV(5, 128)	0.9406	0.9294	0.9458	271,184

Audio Classification

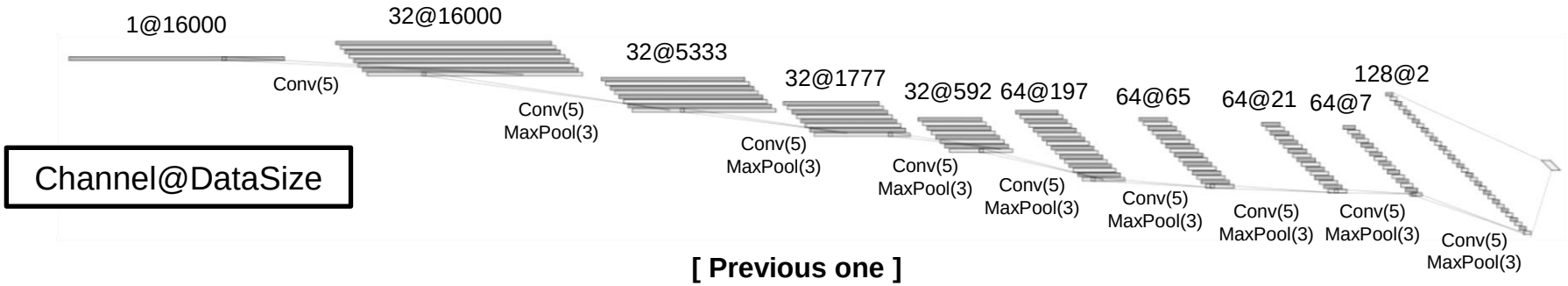
- Start channel size: 64
- The performance is not getting better with only a lot of channel.
- Smooth expansion of the layer is better for feature extraction.

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 64)	X	X	X	X
2 CONV(5, 64)	X	X	X	X
3 CONV(5, 64)	X	X	X	X
4 CONV(5, 128)	0.6540	0.6108	0.4974	1,294,992
5 CONV(5, 128)	0.7724	0.7007	0.8108	568,080
6 CONV(5, 128)	0.8862	0.8336	0.8719	379,792
7 CONV(5, 256)	0.9327	0.9259	0.9250	496,784
8 CONV(5, 256)	0.9470	0.9504	0.9516	767,376
9 CONV(5, 256)	0.9402	0.9551	0.9626	1,074,832

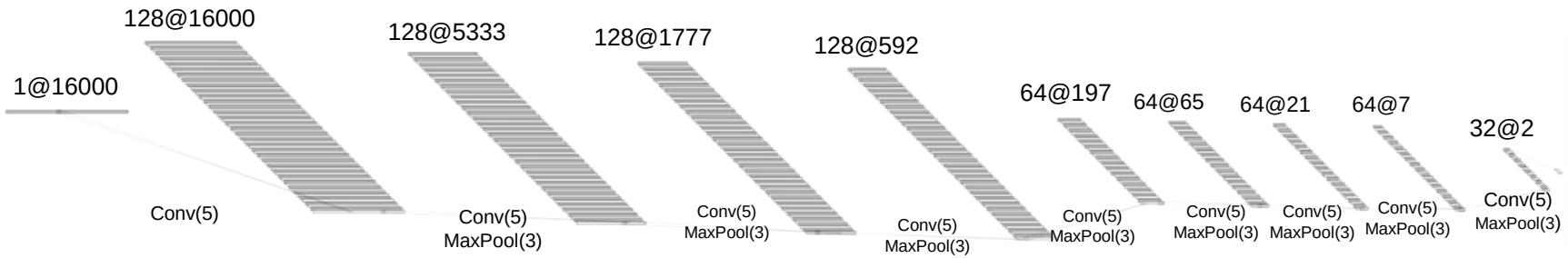
Audio Classification

- Change the channel to become smaller and smaller.
- The upper is ‘Custom channel 32 DO(0.75) 9 Conv’ model’s detail.

- Previous one: 1 → 32 → 32 → 32 → 32 → 64 → 64 → 64 → 64 → 128
  Tried one : 1 → 128 → 128 → 128 → 128 → 64 → 64 → 64 → 64 → 32



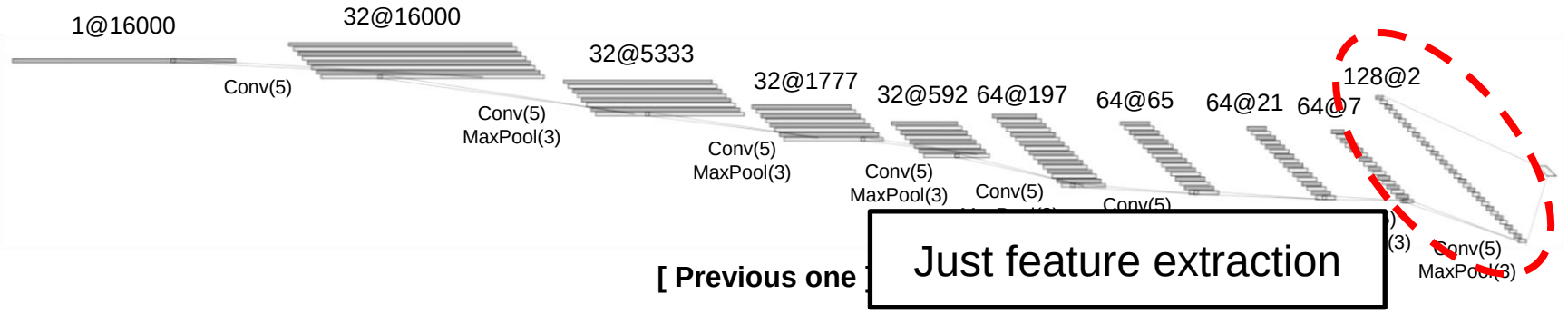
[Previous one]

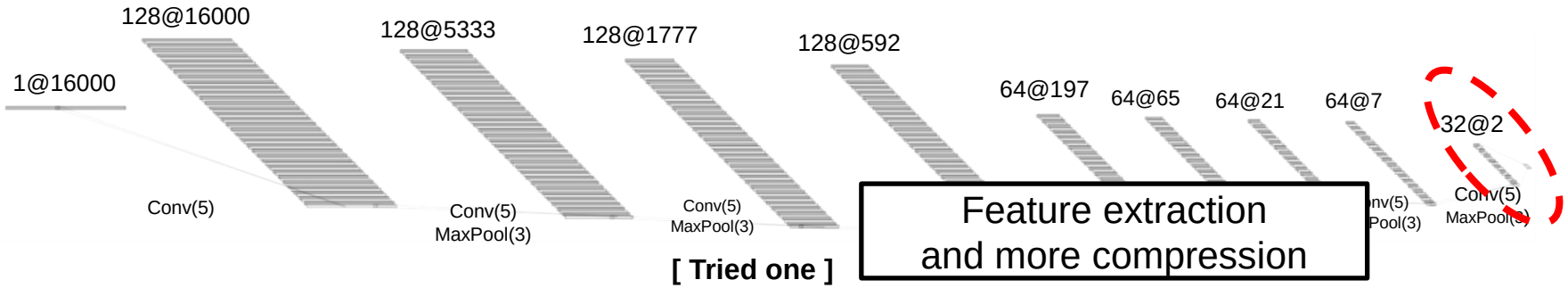


[Tried one]

Audio Classification

- I think this try is very meaningful.
- Because this model focus on the feature compression than previous model.
- Like autoencoder. (16000 → 64(=2*32channel))
- This model is better when use other classifiers after extract the feature from CNN.





Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 64

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 64)	X	X	X	18,432,528
2 CONV(5, 64)	X	X	X	6,164,816
3 CONV(5, 64)	X	X	X	2,088,976
4 CONV(5, 64)	X	X	X	744,528
5 CONV(5, 32)	X	X	X	186,352
6 CONV(5, 32)	0.8897	0.8085	0.8667	115,536
7 CONV(5, 32)	0.9259	0.8854	0.9121	95,408
8 CONV(5, 32)	0.9464	0.9205	0.9321	92,560
9 CONV(5, 16)	0.9238	0.9074	0.9327	91,712

Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 128

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 128)	X	X	X	32,768,784
2 CONV(5, 128)	X	X	X	11,004,816
3 CONV(5, 128)	X	X	X	3,804,176
4 CONV(5, 128)	X	X	X	1,459,344
5 CONV(5, 64)	X	X	X	489,680
6 CONV(5, 64)	0.8995	0.8474	0.8829	384,656
7 CONV(5, 64)	0.9421	0.9043	0.9202	354,640
8 CONV(5, 64)	0.9516	0.9360	0.9450	359,184
9 CONV(5, 32)	0.9425	0.9379	0.9458	362,608

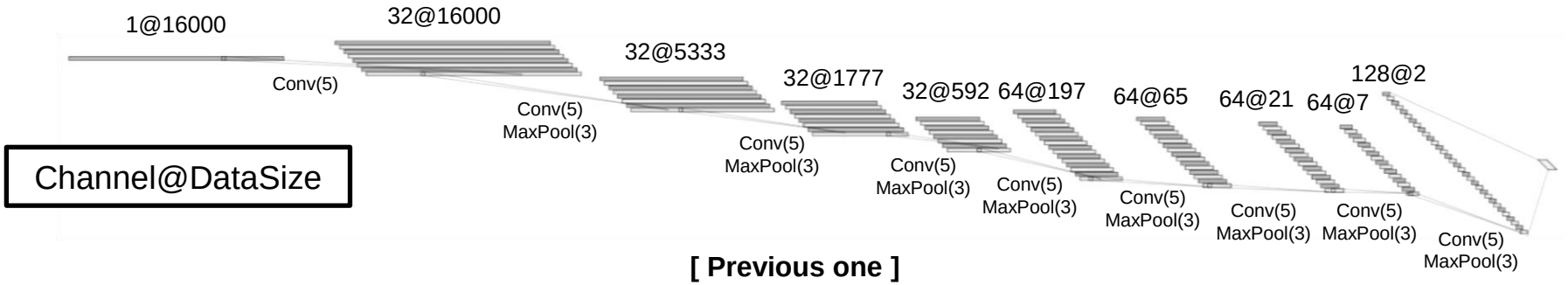
Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 256

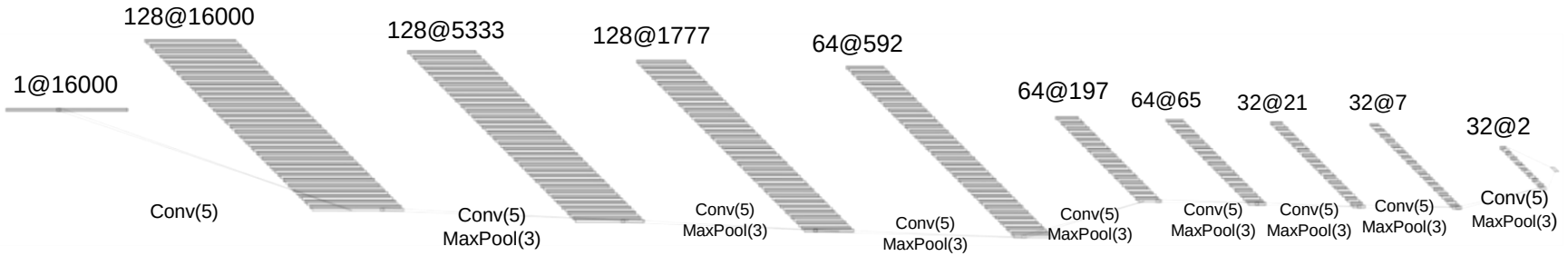
Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 256)	X	X	X	65,537,552
2 CONV(5, 256)	X	X	X	22,173,456
3 CONV(5, 256)	X	X	X	7,936,016
4 CONV(5, 256)	X	X	X	3,410,192
5 CONV(5, 128)	X	X	X	1,552,784
6 CONV(5, 128)	0.8866	0.8517	0.8964	1,364,496
7 CONV(5, 128)	0.9423	0.9315	0.9346	1,356,432
8 CONV(5, 128)	0.9583	0.9502	0.9605	1,409,808
9 CONV(5, 64)	0.9603	0.9572	0.9628	1,438,544

Audio Classification

- So I summarise previous try.
- Change the channel more steeply and to become more smaller and smaller
- The upper is ‘Custom channel 32 DO(0.75) 9 Conv’ model’s detail.



[Previous one]



[Tried one]

Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 64

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 64)	X	X	X	16,384,400
2 CONV(5, 64)	X	X	X	5,481,936
3 CONV(5, 64)	X	X	X	1,861,136
4 CONV(5, 32)	0.6816	0.6426	0.7277	354,864
5 CONV(5, 32)	0.8019	0.7373	0.8075	157,776
6 CONV(5, 32)	0.8933	0.8363	0.8766	95,344
7 CONV(5, 16)	0.9190	0.8654	0.9128	70,016
8 CONV(5, 16)	0.9211	0.9022	0.9223	67,728
9 CONV(5, 16)	0.9053	0.9067	0.9094	67,744

Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 128

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 128)	X	X	X	32,768,784
2 CONV(5, 128)	X	X	X	11,004,816
3 CONV(5, 128)	X	X	X	3,804,176
4 CONV(5, 64)	0.6814	0.6557	0.7242	812,112
5 CONV(5, 64)	0.8012	0.7497	0.8010	428,176
6 CONV(5, 64)	0.8887	0.8355	0.8725	313,552
7 CONV(5, 32)	0.9385	0.9007	0.9161	268,016
8 CONV(5, 32)	0.9402	0.9240	0.9373	266,000
9 CONV(5, 32)	0.9381	0.9259	0.9408	268,592

Audio Classification

- I didn’t try early case because the results were probably not good.
- Start channel size: 256

Architecture	DO(0.5)	BN	DO+BN	Params
1 CONV(5, 256)	X	X	X	65,537,552
2 CONV(5, 256)	X	X	X	22,173,456
3 CONV(5, 256)	X	X	X	7,936,016
4 CONV(5, 128)	0.6982	0.6505	0.7425	2,033,808
5 CONV(5, 128)	0.7859	0.7065	0.8201	1,306,896
6 CONV(5, 128)	0.8947	0.8467	0.8721	1,118,608
7 CONV(5, 64)	0.9381	0.9047	0.9196	1,048,016
8 CONV(5, 64)	0.9572	0.9522	0.9458	1,054,224
9 CONV(5, 64)	0.9472	0.9431	0.9589	1,069,648

Audio Classification

- In summary, This is SOTA(State Of The Art) in this research.

- Though the performance is lower than the previous model, I think it is meaningful.
- The original data size, 16000, was compressed to 32 and the performance was 0.9327.
- And the other model compress to 128 and the performance was 0.9628.

- What if we use a feature from CNN as input to another model, this model will be very useful.

	Architecture (i = 0,1,2…)	1D DO(0.5)	1D BN	1D DO+BN	Params
	16000(Input data size) → 2(Length) * 16(Channel) = 32 (Feature size)				
Feature extraction ch 64	9 CONV(5, 16)	0.9238	0.9074	0.9327	91,712
	16000(Input data size) → 2(Length) * 64(Channel) = 128 (Feature Size)				
Feature extraction ch 256	9 CONV(5, 64)	0.9603	0.9572	0.9628	1,438,544

Any Question?

Thank you