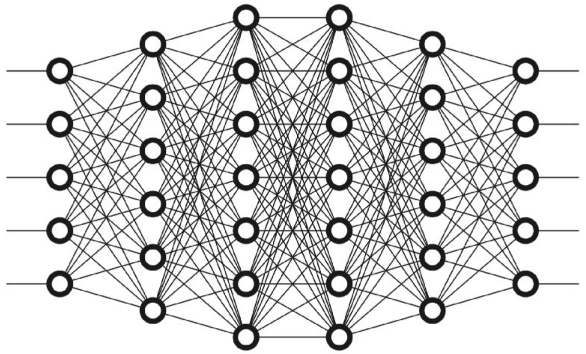
DNN Pruning for Acoustic Scene Classification

Tianyi Chen

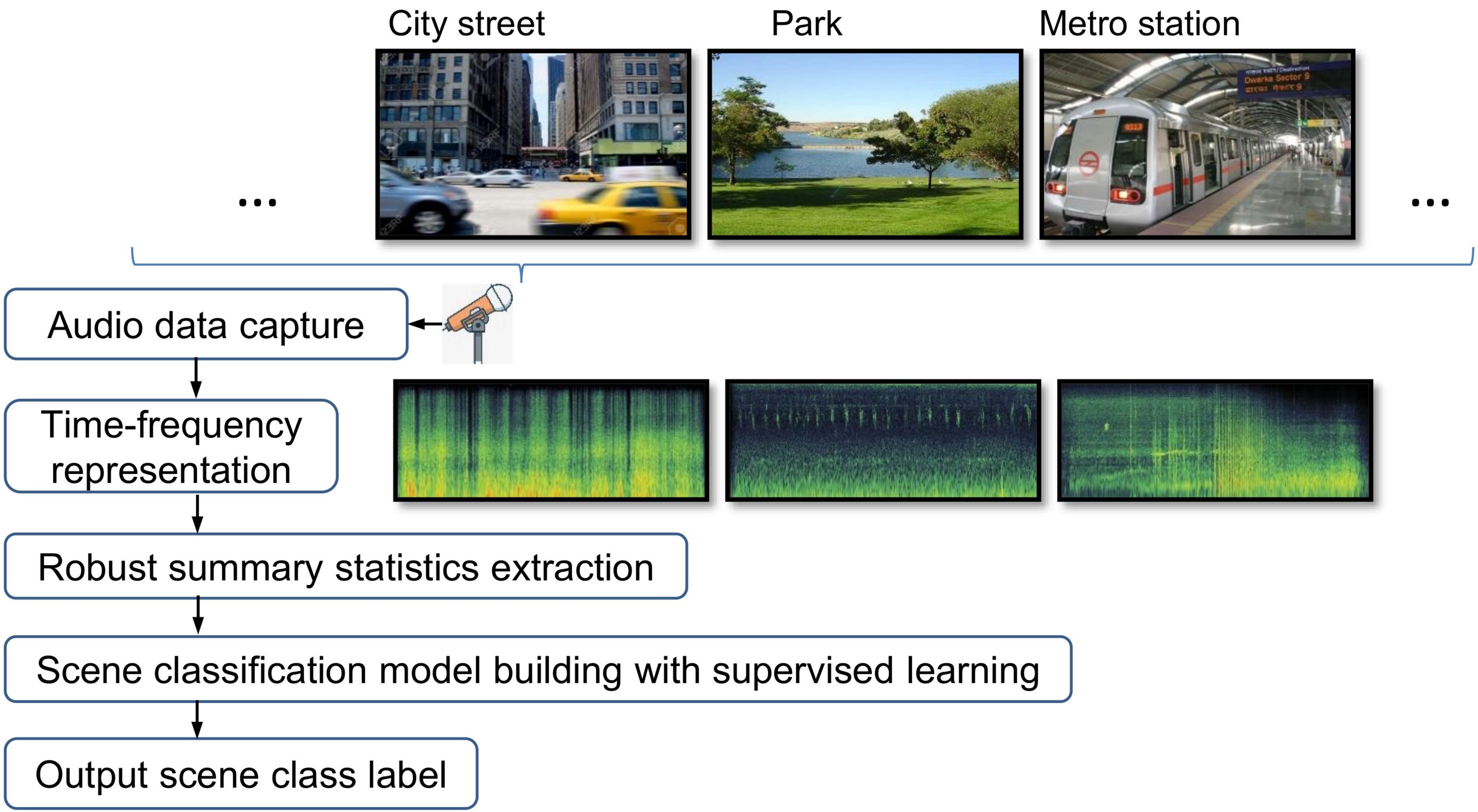
tianyic@uw.edu

Deep Neural Network in Audio Field.

- DNN has achieved tremendous success in various fields.
- Deeper and heavier DNNs typically deliver better performance.
- However, for many audio processing tasks, opposite happened.
- Therefore, people typically use shallow DNN in many audio system.



Acoustic Scene Classification



Related Work

- In order to leverage deep and heavier model into acoustic tasks, there are some efforts.

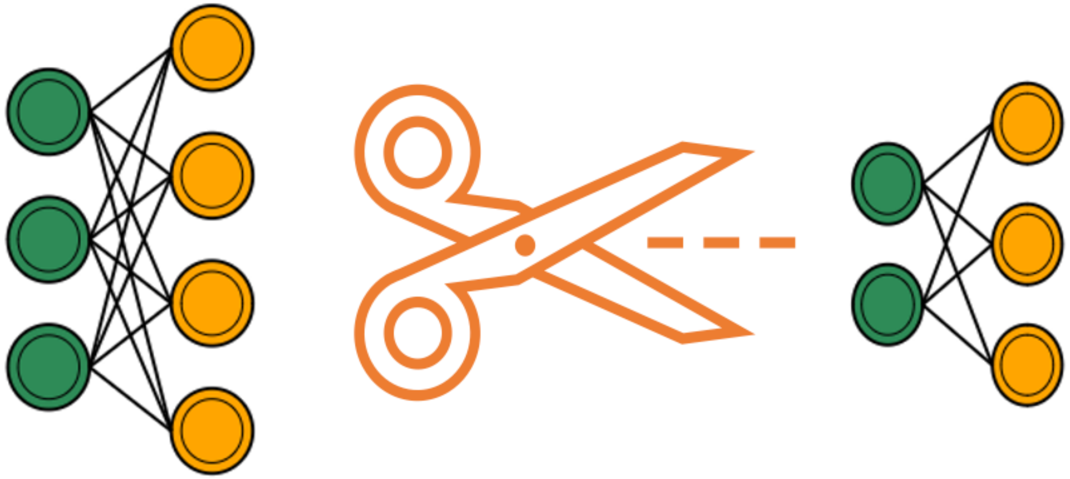
- Some works study how restrict the receptive fields of Deep Neural Network.

- Lower receptive field indicates DNN with lower learning capacity.

[1] The receptive field as a regularizer in deep convolutional neural net- works for acoustic scene classification.

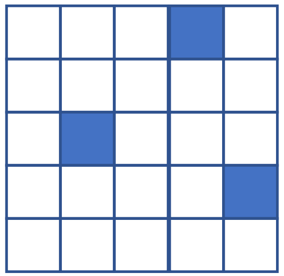
DNN Pruning

- Pruning projects elements onto DNN's variables onto zeros.
- Essentially introduce various **sparsity** into the DNN.
- Structured pruning can further speed up the DNN.

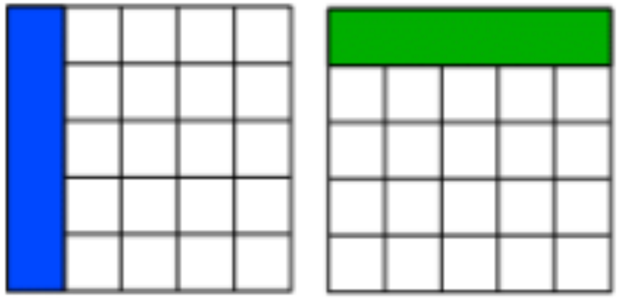


Prune redundancy.

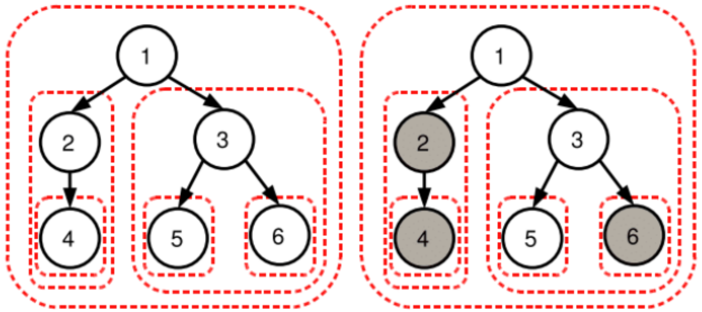
Sparsity Pattern



Fine-grained sparsity



Group sparsity



Hierarchy sparsity

Sparsity Inducing Optimization Problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ f(\boldsymbol{x}),\quad \text{s.t. Card}\{g\in\mathcal{G}\,|\,[\boldsymbol{x}]_g=0\}=K$$

- Card is cardinality. Cardinality of one set is refers to the number of elements in that set.
- K is the target sparsity level.
- G is a partition of index set of variables.
- X is the trainable variables.
- F is the loss function.

This problem is hard to solve since the constraint is non-convex, non-smooth.
So people relaxes it to some regularizer r(x). The problem becomes

$$\min_{x\in R^n} f(x)+\lambda r(x)$$

Choices of $\mathrm{r}(x)$

- Different $\Omega(x)$ results in different pattern of sparsity of model parameters.
- **l1 norm of x**:

$$\mathrm{r}(x) = \|x\|_1$$

each element in x is individually set as zero.

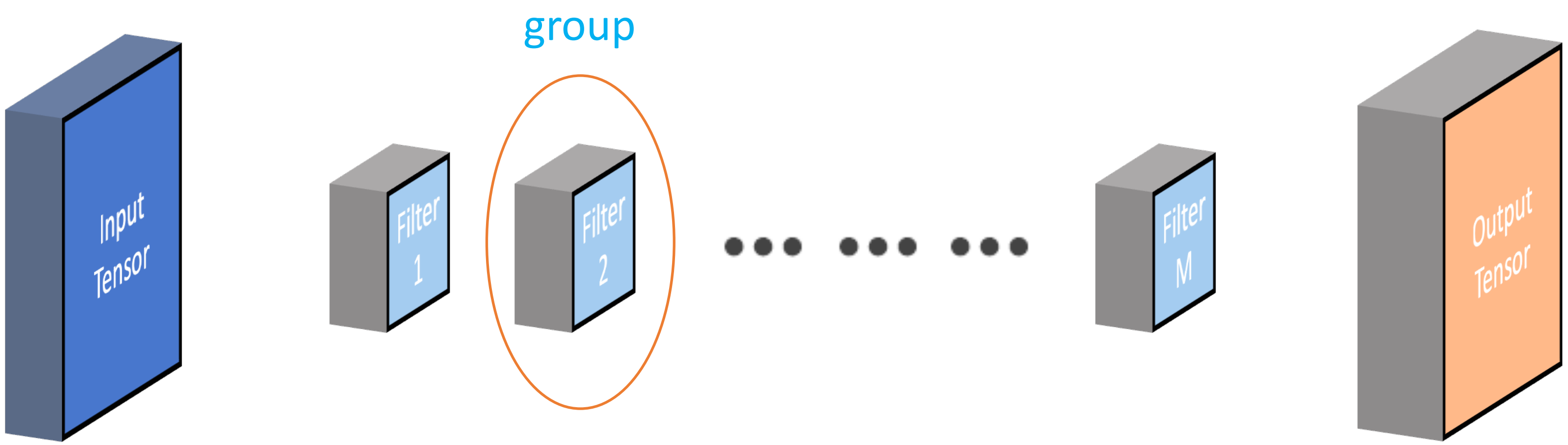
- **Mixed l1/lp norm of x**:

$$\mathrm{r}(x) = \sum_{g \in G} \left\| [x]_g \right\|_p$$

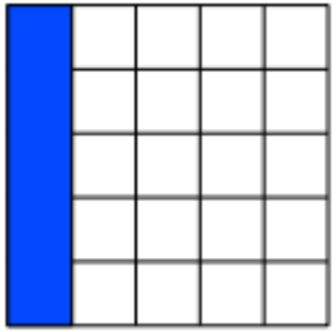
where G is a partition of variable indices, such norm can promote a group of elements as zero, referred **group sparsity**.

Mixed l1/lp norm of x in Neural Models

- For CNN, a group of variables can be defined as a filter in ConvLayer.



- For RNN, the row or column of weight matrix can be selected a group.



Sparse Optimizer

- Proximal Method.
- ADMM.
- HSPG family. (Ours, the best so far.)

Two key metrics: a) **low objective function value**, and b) **high group sparsity.**

ADMM is somewhat equivalent to proximal method, but is **unnecessarily complicated**.

In optimization area, proximal method is the main trend, and appears on top-tier confs every year, e.g., Prox-SGD, SAGA, Spider.

However, ADMM merely appears in application papers.

Sparsity Optimizer Comparison

Effectively solve the following problem in stochastic setting:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ f(\boldsymbol{x}),\quad \text{s.t. Card}\{g\in\mathcal{G}|[\boldsymbol{x}]_g=0\}=K,$$

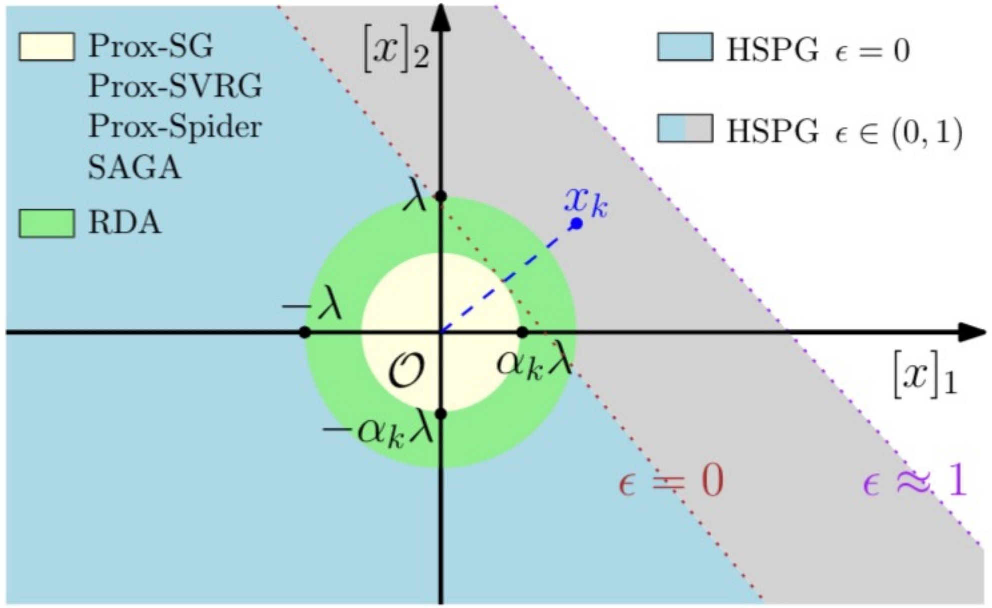
Metric	Proximal Method	ADMM	OBProxSG	HSPG
Convergence (final objective value)	**#1**	**#1**	**#1**	**#1**
Group-Sparsity	Poor	Depends	**#1**	**#1**
Runtime	Fast	Slow	**Fast**	**Fast**
One-shot	Yes	Depends	**Yes**	**Yes**

Why existing stochastic optimizers failed?

- Existing methods, such as proximal gradient method rarely generates group sparse solution. The solution is even fully dense.

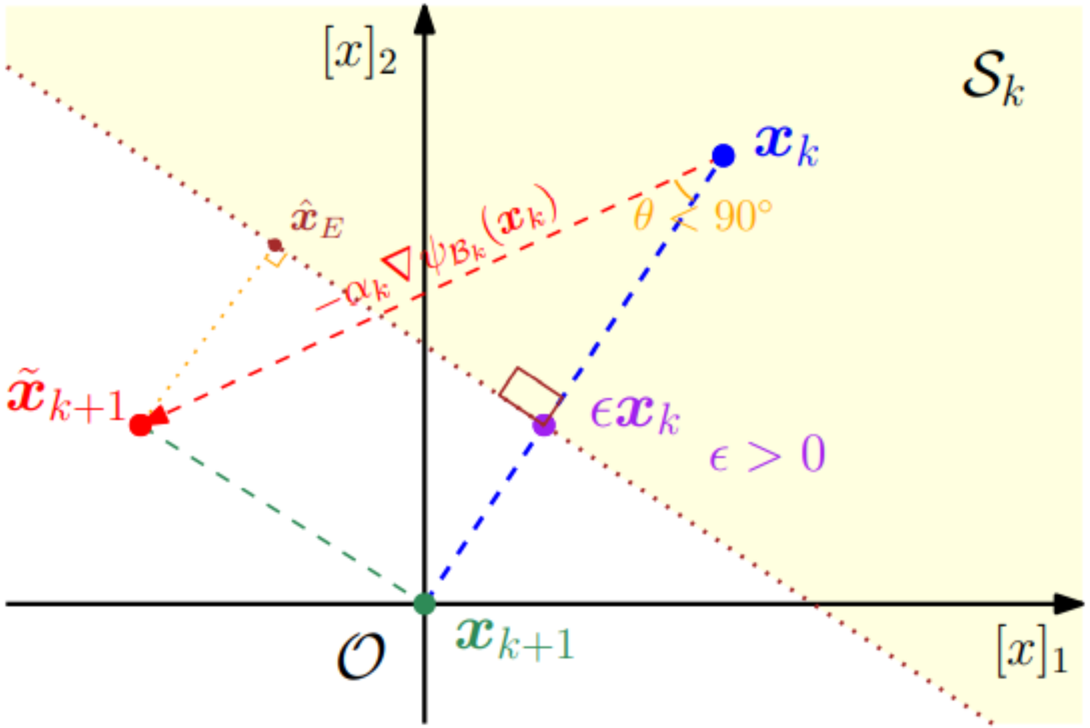
$$[\boldsymbol{x}_{k+1}]_g = \max\left\{0, 1 - \alpha_k\lambda / \left\|[\widehat{\boldsymbol{x}}_{k+1}]_g\right\|\right\} \cdot [\widehat{\boldsymbol{x}}_{k+1}]_g.$$

- **Projection Region is too small.**
  - In DNN, learning rate is typically less than 1e-3.
  - Lambda is much less than 1.
- Randomness.

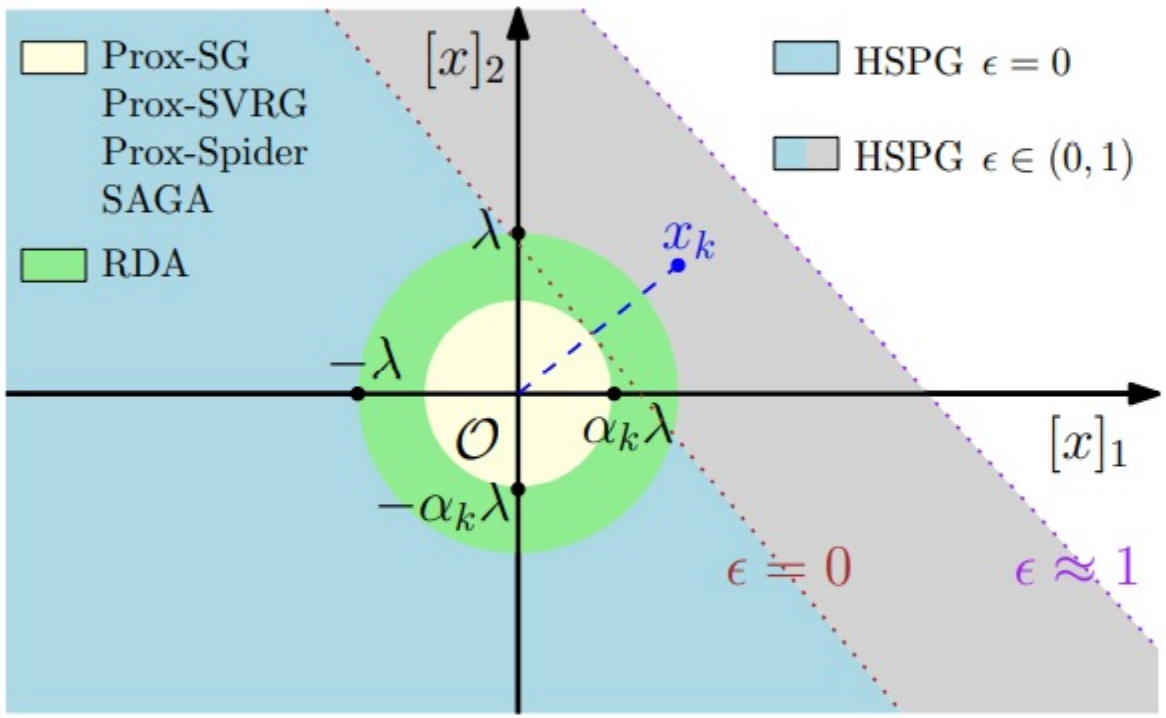


HSPG

- Resolve the poor capacity of sparsity exploration.
- **OBProxSG** is for fine-grained sparsity.
- **HSPG** is for group sparsity.



(a) Half-Space Projection



(b) Projection Region For Mixed ℓ_1/ℓ_2 Regularization

Experiment: Datasets

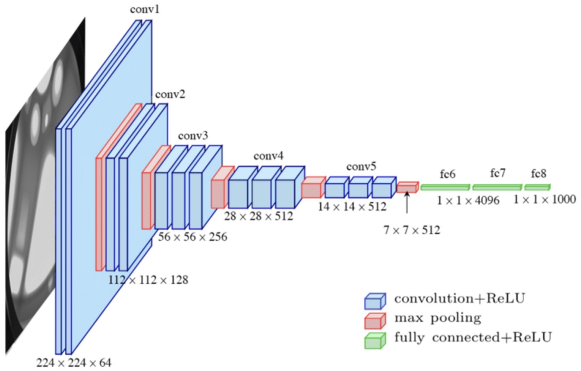
- DCASE2017.

The audio clips are decomposed into 10 seconds samples forming 4680 training samples (13 hours) and 1620 testing samples (4.5 hours).

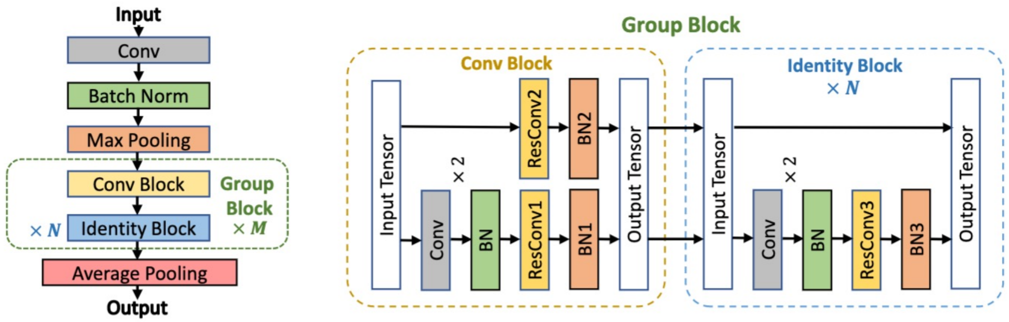
- DCASE 2018.

17 hours of audio for training (6122 10-second clips) and 7 hours for evaluation (2518 10-second clips).

Experiments: DNN Architectures



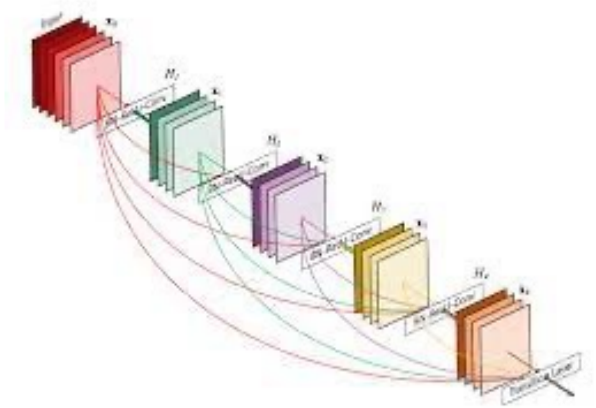
VGG



(a) ResNet50.

(b) Group block.

ResNet



DenseNet

Experiment Setting

- Train 350 epochs.
- First 50 epochs for warm up training.
- Learning rate starts at 1e-4.
- Decay learning rate linearly till 5e-6 after 50 epochs.

Experiment Result

Table 3: Accuracy (%) / Group Sparsity (%) on DCASE17

	VGG	DenseNet	ResNet
Baseline	67.90 ± 1.31	63.48 ± 4.96	67.19 ± 1.72
RN1	–	–	71.11 ± 1.19
RN2	–	–	72.41 ± 0.96
RN3	–	–	71.74 ± 0.85
DN1	–	72.24 ± 1.00	–
HSPG ($\lambda = 10^{-3}$)	71.34 / 50.12	70.36 / 65.47	72.28 / 69.38
HSPG ($\lambda = 10^{-4}$)	69.22 / 10.49	68.26 / 8.08	70.32 / 15.27

- Larger lambda higher group sparsity.
- Higher group sparsity higher accuracy.

Experiment Result

Table 4: Accuracy (%) / Group Sparsity (%) on DCASE18

	VGG	DenseNet	ResNet
Baseline	74.56± 1.01	71.55 ± 0.85	71.05± 0.87
RN1	–	–	77.34 ± 1.53
RN2	–	–	75.71 ± 0.70
RN3	–	–	77.61 ± 0.22
DN1	–	76.39 ± 0.14	–
HSPG ($\lambda = 10^{-3}$)	75.29 / 53.42	75.33 / 61.48	77.46 / 71.25
HSPG ($\lambda = 10^{-4}$)	73.24 / 6.21	72.20 / 3.55	72.98 / 17.29

- Larger lambda higher group sparsity.
- Higher group sparsity higher accuracy.

Thank you very much!