Final Project:
Light Weight Model

Team 14

Outline

1. Baseline

- **Knowledge Distillation**
- **Model Pruning**

2. Our Method

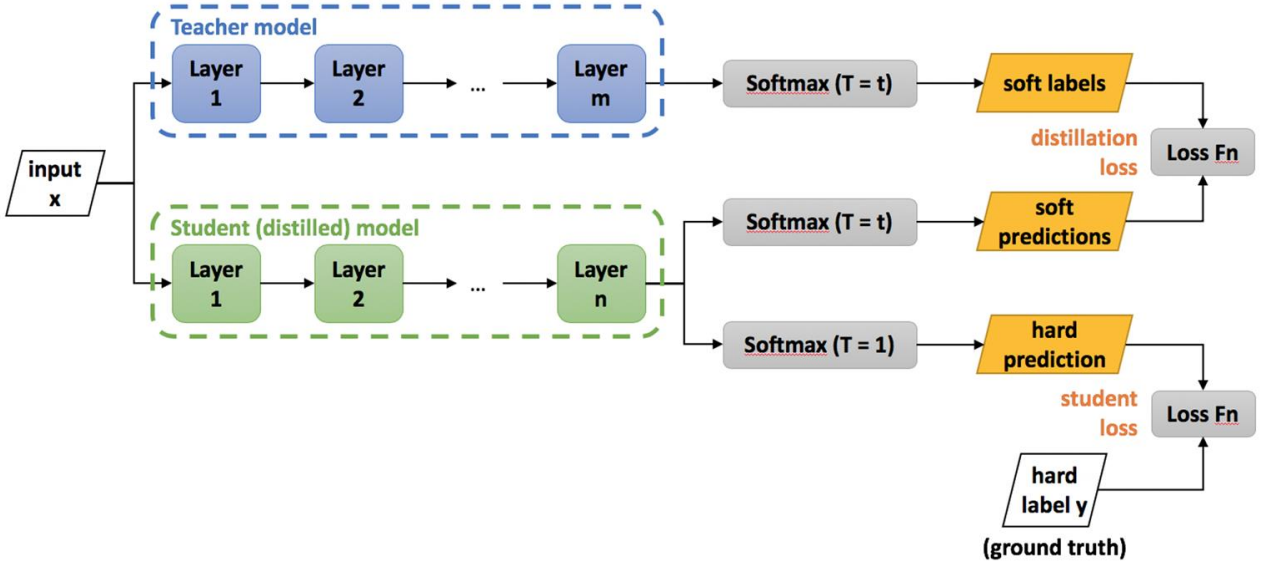
- **Training Pipeline**
- **Model Architecture**
- **Decreasing Temperature**
- **Training Details**
- **Learning Curve**

3. Experiment Result

- **Different Methods**
- **Student Model Architecture**

Knowledge Distillation

- Transfer knowledge from a complex model to a simplified one, typically by guiding the training of the simplified model using the predicted distribution of the complex model



Knowledge Distillation

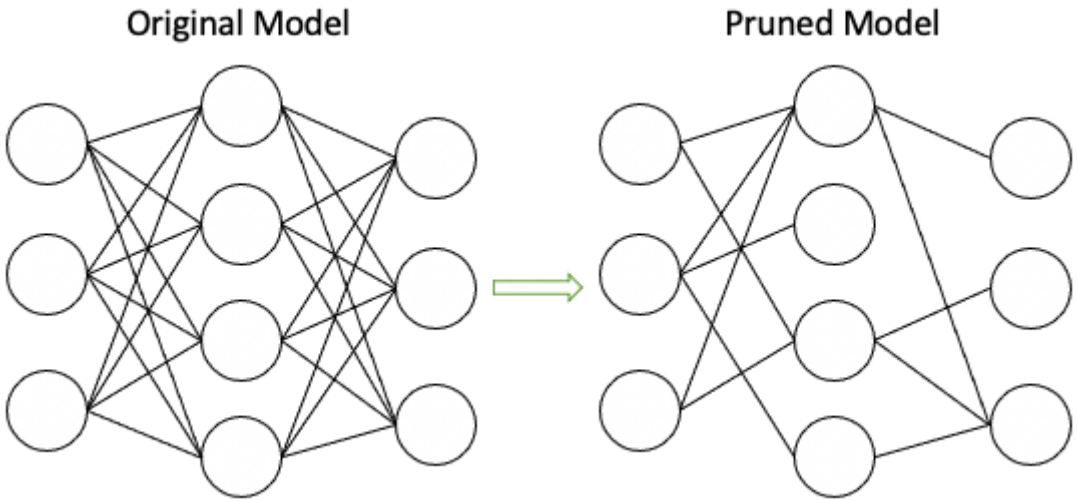
$$\mathcal{L} = (1 - \lambda)\, \mathcal{L}_{student} + \lambda\, \mathcal{L}_{distillation}$$

$$\mathcal{L}_{student} = \mathbf{CE}(\mathrm{softmax}(a_s), y)$$

$$\mathcal{L}_{distillation} = T^2\, \mathbf{KL}(\mathrm{softmax}(\tfrac{a_s}{T}), \mathrm{softmax}(\tfrac{a_t}{T}))$$

Model Pruning

- Pruning eliminates the weights with low magnitude and makes the model sparser.



Model Pruning

- Using `torch.nn.utils.prune.global_unstructured` to prune our model.
- Global pruning
  - Prune the model all at once, by removing connections across the whole model, instead of in each layer.
- Unstructured pruning
  - Weights are pruned without considering any specific structure or pattern within a layer.

Outline

1. Baseline

- Knowledge Distillation
- Model Pruning

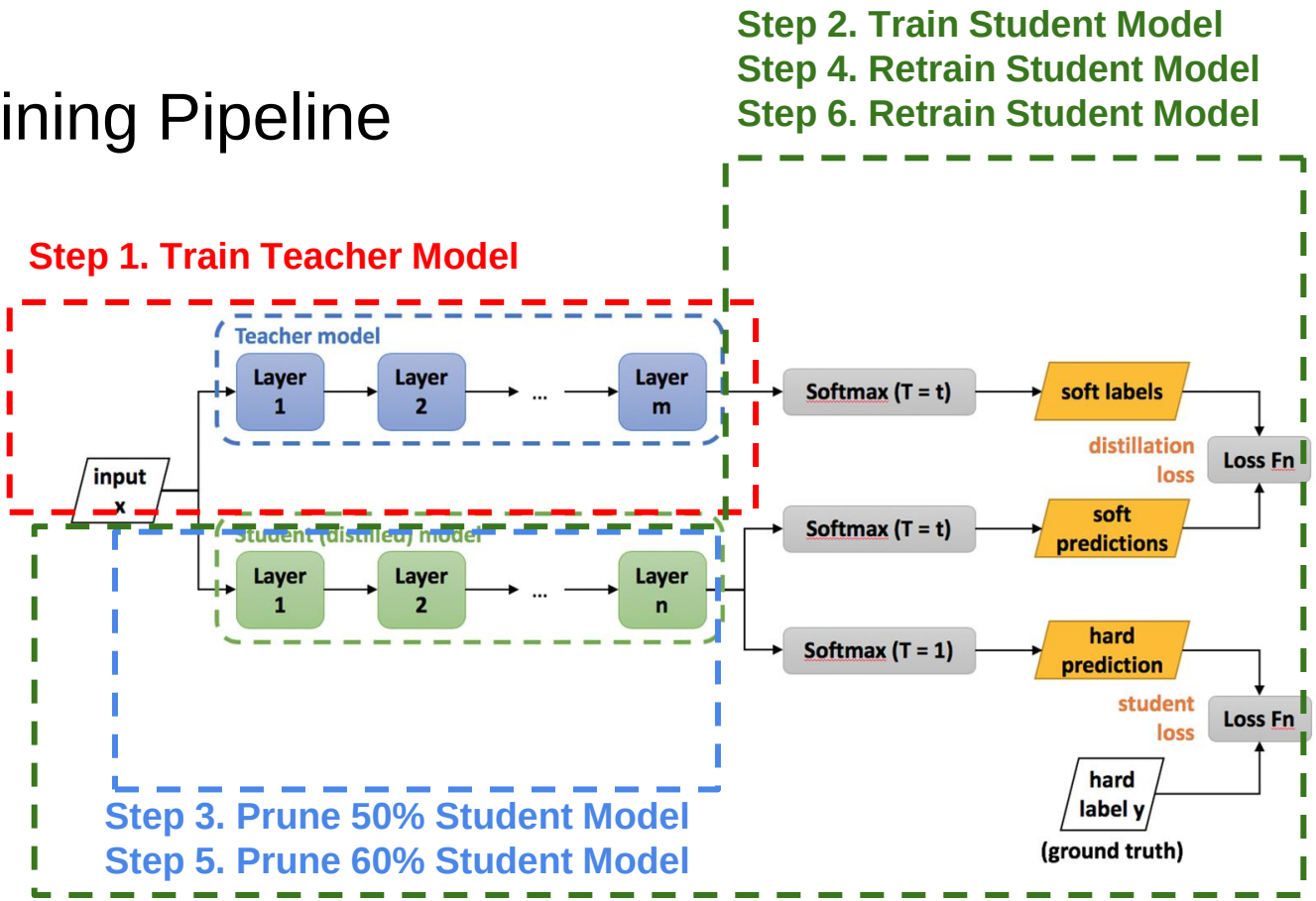
2. Our Method

- **Training Pipeline**
- **Model Architecture**
- **Decreasing Temperature**
- **Training Details**
- **Learning Curve**

3. Experiment Result

- Different Methods
- Student Model Architecture

Training Pipeline



Model Architecture

- Teacher model: RegNetX-1.6GF
  - modify the fc output neurons to 525
  - params: 8,756,461 (~8.8M)
- Student model: ShuffleNetV2-x0.5
  - modify the conv5 output channels to 64 and fc output neurons to 525
  - params: 189,677 (~0.19M)

Layer	Output size	KSize	Stride	Repeat	Output channels			
					0.5×	1×	1.5×	2×
Image	224×224				3	3	3	3
Conv1	112×112	3×3	2	1	24	24	24	24
MaxPool	56×56	3×3	2					
Stage2	28×28		2	1	48	116	176	244
	28×28		1	3				
Stage3	14×14		2	1	96	232	352	488
	14×14		1	7				
Stage4	7×7		2	1	192	464	704	976
	7×7		1	3				
Conv5	7×7	1×1	1	1 64	~~1024~~	1024	1024	2048
GlobalPool	1×1	7×7						
FC				525	~~1000~~	1000	1000	1000
FLOPs					41M	146M	299M	591M
# of Weights				0.19M	~~1.4M~~	2.3M	3.5M	7.4M

Table 5: Overall architecture of ShuffleNet v2, for four different levels of complexities.

- Weights initialized with pre-trained weights on ImageNetV2

Model Architecture

- Teacher model

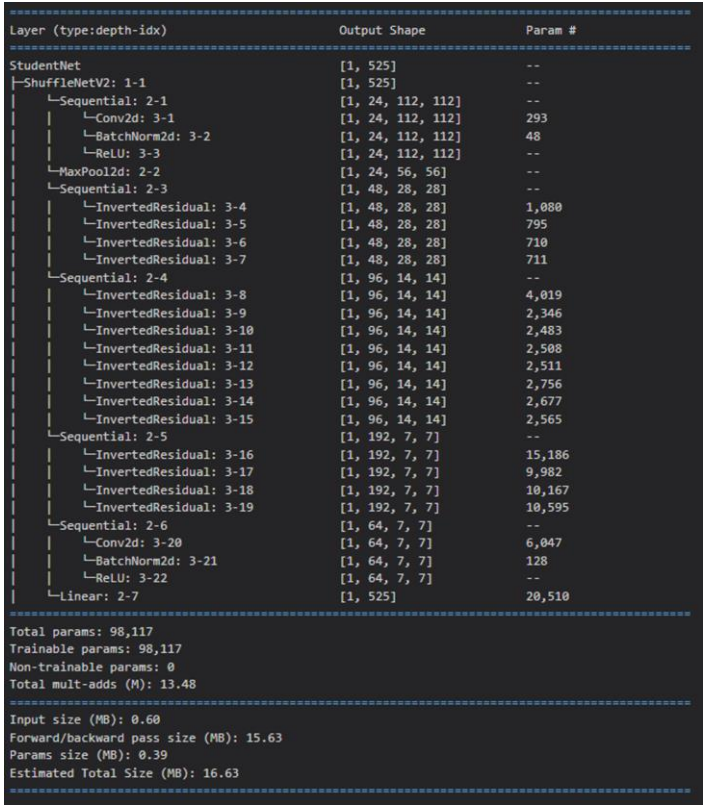
```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
TeacherNet                                              [1, 525]                  --
├─RegNet: 1-1                                           [1, 525]                  --
│    └─SimpleStemIN: 2-1                                [1, 32, 112, 112]         --
│    │    └─Conv2d: 3-1                                 [1, 32, 112, 112]         864
│    │    └─BatchNorm2d: 3-2                            [1, 32, 112, 112]         64
│    │    └─ReLU: 3-3                                   [1, 32, 112, 112]         --
│    └─Sequential: 2-2                                  [1, 912, 7, 7]            --
│    │    └─AnyStage: 3-4                               [1, 72, 56, 56]           52,272
│    │    └─AnyStage: 3-5                               [1, 168, 28, 28]          371,280
│    │    └─AnyStage: 3-6                               [1, 408, 14, 14]          4,206,480
│    │    └─AnyStage: 3-7                               [1, 912, 7, 7]            3,646,176
│    └─AdaptiveAvgPool2d: 2-3                           [1, 912, 1, 1]            --
│    └─Linear: 2-4                                      [1, 525]                  479,325
==========================================================================================
Total params: 8,756,461
Trainable params: 8,756,461
Non-trainable params: 0
Total mult-adds (G): 1.60
==========================================================================================
Input size (MB): 0.60
Forward/backward pass size (MB): 126.92
Params size (MB): 35.03
Estimated Total Size (MB): 162.55
==========================================================================================
```

- Student model

```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
StudentNet                                              [1, 525]                  --
├─ShuffleNetV2: 1-1                                     [1, 525]                  --
│    └─Sequential: 2-1                                  [1, 24, 112, 112]         --
│    │    └─Conv2d: 3-1                                 [1, 24, 112, 112]         648
│    │    └─BatchNorm2d: 3-2                            [1, 24, 112, 112]         48
│    │    └─ReLU: 3-3                                   [1, 24, 112, 112]         --
│    └─MaxPool2d: 2-2                                   [1, 24, 56, 56]           --
│    └─Sequential: 2-3                                  [1, 48, 28, 28]           --
│    │    └─InvertedResidual: 3-4                       [1, 48, 28, 28]           2,400
│    │    └─InvertedResidual: 3-5                       [1, 48, 28, 28]           1,512
│    │    └─InvertedResidual: 3-6                       [1, 48, 28, 28]           1,512
│    │    └─InvertedResidual: 3-7                       [1, 48, 28, 28]           1,512
│    └─Sequential: 2-4                                  [1, 96, 14, 14]           --
│    │    └─InvertedResidual: 3-8                       [1, 96, 14, 14]           8,256
│    │    └─InvertedResidual: 3-9                       [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-10                      [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-11                      [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-12                      [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-13                      [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-14                      [1, 96, 14, 14]           5,328
│    │    └─InvertedResidual: 3-15                      [1, 96, 14, 14]           5,328
│    └─Sequential: 2-5                                  [1, 192, 7, 7]            --
│    │    └─InvertedResidual: 3-16                      [1, 192, 7, 7]            30,336
│    │    └─InvertedResidual: 3-17                      [1, 192, 7, 7]            19,872
│    │    └─InvertedResidual: 3-18                      [1, 192, 7, 7]            19,872
│    │    └─InvertedResidual: 3-19                      [1, 192, 7, 7]            19,872
│    └─Sequential: 2-6                                  [1, 64, 7, 7]             --
│    │    └─Conv2d: 3-20                                [1, 64, 7, 7]             12,288
│    │    └─BatchNorm2d: 3-21                           [1, 64, 7, 7]             128
│    │    └─ReLU: 3-22                                  [1, 64, 7, 7]             --
│    └─Linear: 2-7                                      [1, 525]                  34,125
==========================================================================================
Total params: 189,677
Trainable params: 189,677
Non-trainable params: 0
Total mult-adds (M): 30.46
==========================================================================================
Input size (MB): 0.60
Forward/backward pass size (MB): 15.63
Params size (MB): 0.76
Estimated Total Size (MB): 16.99
==========================================================================================
```

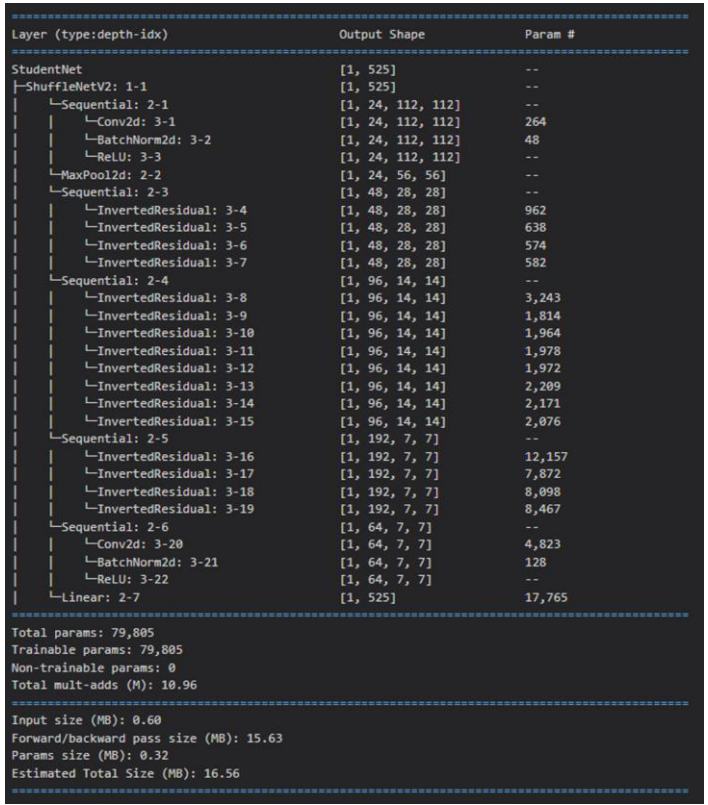
Model Architecture

- Student model after 50% pruning



```
==========================================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================================
StudentNet                               [1, 525]                  --
├─ShuffleNetV2: 1-1                      [1, 525]                  --
│    └─Sequential: 2-1                   [1, 24, 112, 112]         --
│    │    └─Conv2d: 3-1                  [1, 24, 112, 112]         293
│    │    └─BatchNorm2d: 3-2             [1, 24, 112, 112]         48
│    │    └─ReLU: 3-3                    [1, 24, 112, 112]         --
│    └─MaxPool2d: 2-2                    [1, 24, 56, 56]           --
│    └─Sequential: 2-3                   [1, 48, 28, 28]           --
│    │    └─InvertedResidual: 3-4        [1, 48, 28, 28]           1,080
│    │    └─InvertedResidual: 3-5        [1, 48, 28, 28]           795
│    │    └─InvertedResidual: 3-6        [1, 48, 28, 28]           710
│    │    └─InvertedResidual: 3-7        [1, 48, 28, 28]           711
│    └─Sequential: 2-4                   [1, 96, 14, 14]           --
│    │    └─InvertedResidual: 3-8        [1, 96, 14, 14]           4,019
│    │    └─InvertedResidual: 3-9        [1, 96, 14, 14]           2,346
│    │    └─InvertedResidual: 3-10       [1, 96, 14, 14]           2,483
│    │    └─InvertedResidual: 3-11       [1, 96, 14, 14]           2,508
│    │    └─InvertedResidual: 3-12       [1, 96, 14, 14]           2,511
│    │    └─InvertedResidual: 3-13       [1, 96, 14, 14]           2,756
│    │    └─InvertedResidual: 3-14       [1, 96, 14, 14]           2,677
│    │    └─InvertedResidual: 3-15       [1, 96, 14, 14]           2,565
│    └─Sequential: 2-5                   [1, 192, 7, 7]            --
│    │    └─InvertedResidual: 3-16       [1, 192, 7, 7]            15,186
│    │    └─InvertedResidual: 3-17       [1, 192, 7, 7]            9,982
│    │    └─InvertedResidual: 3-18       [1, 192, 7, 7]            10,167
│    │    └─InvertedResidual: 3-19       [1, 192, 7, 7]            10,595
│    └─Sequential: 2-6                   [1, 64, 7, 7]             --
│    │    └─Conv2d: 3-20                 [1, 64, 7, 7]             6,047
│    │    └─BatchNorm2d: 3-21            [1, 64, 7, 7]             128
│    │    └─ReLU: 3-22                   [1, 64, 7, 7]             --
│    └─Linear: 2-7                       [1, 525]                  20,510
==========================================================================================
Total params: 98,117
Trainable params: 98,117
Non-trainable params: 0
Total mult-adds (M): 13.48
==========================================================================================
Input size (MB): 0.60
Forward/backward pass size (MB): 15.63
Params size (MB): 0.39
Estimated Total Size (MB): 16.63
==========================================================================================
```

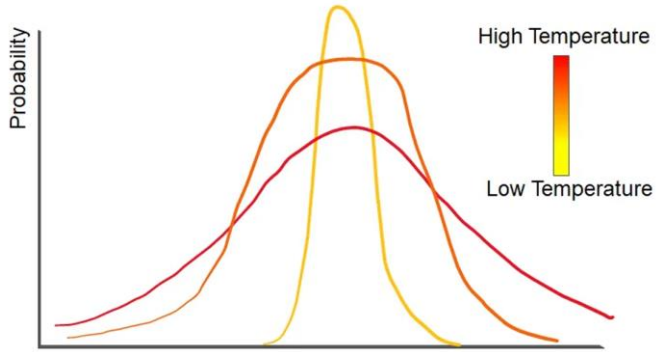
- Student model after 60% pruning



```
==========================================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================================
StudentNet                               [1, 525]                  --
├─ShuffleNetV2: 1-1                      [1, 525]                  --
│    └─Sequential: 2-1                   [1, 24, 112, 112]         --
│    │    └─Conv2d: 3-1                  [1, 24, 112, 112]         264
│    │    └─BatchNorm2d: 3-2             [1, 24, 112, 112]         48
│    │    └─ReLU: 3-3                    [1, 24, 112, 112]         --
│    └─MaxPool2d: 2-2                    [1, 24, 56, 56]           --
│    └─Sequential: 2-3                   [1, 48, 28, 28]           --
│    │    └─InvertedResidual: 3-4        [1, 48, 28, 28]           962
│    │    └─InvertedResidual: 3-5        [1, 48, 28, 28]           638
│    │    └─InvertedResidual: 3-6        [1, 48, 28, 28]           574
│    │    └─InvertedResidual: 3-7        [1, 48, 28, 28]           582
│    └─Sequential: 2-4                   [1, 96, 14, 14]           --
│    │    └─InvertedResidual: 3-8        [1, 96, 14, 14]           3,243
│    │    └─InvertedResidual: 3-9        [1, 96, 14, 14]           1,814
│    │    └─InvertedResidual: 3-10       [1, 96, 14, 14]           1,964
│    │    └─InvertedResidual: 3-11       [1, 96, 14, 14]           1,978
│    │    └─InvertedResidual: 3-12       [1, 96, 14, 14]           1,972
│    │    └─InvertedResidual: 3-13       [1, 96, 14, 14]           2,209
│    │    └─InvertedResidual: 3-14       [1, 96, 14, 14]           2,171
│    │    └─InvertedResidual: 3-15       [1, 96, 14, 14]           2,076
│    └─Sequential: 2-5                   [1, 192, 7, 7]            --
│    │    └─InvertedResidual: 3-16       [1, 192, 7, 7]            12,157
│    │    └─InvertedResidual: 3-17       [1, 192, 7, 7]            7,872
│    │    └─InvertedResidual: 3-18       [1, 192, 7, 7]            8,098
│    │    └─InvertedResidual: 3-19       [1, 192, 7, 7]            8,467
│    └─Sequential: 2-6                   [1, 64, 7, 7]             --
│    │    └─Conv2d: 3-20                 [1, 64, 7, 7]             4,823
│    │    └─BatchNorm2d: 3-21            [1, 64, 7, 7]             128
│    │    └─ReLU: 3-22                   [1, 64, 7, 7]             --
│    └─Linear: 2-7                       [1, 525]                  17,765
==========================================================================================
Total params: 79,805
Trainable params: 79,805
Non-trainable params: 0
Total mult-adds (M): 10.96
==========================================================================================
Input size (MB): 0.60
Forward/backward pass size (MB): 15.63
Params size (MB): 0.32
Estimated Total Size (MB): 16.56
==========================================================================================
```

Decreasing Temperature

- Temperature: control the smoothness of the model's predicted distribution
- Linearly decrease from 10 to 5
- Early training stages: high temperature → flatter distribution → help the student model to learn the teacher model’s behavior faster and easier
- Later training stages: low temperature → sharper distribution → help the student model to concentrate on the most likely classes



Training Details - Data Preprocessing

- Resize images to 224x224
- Apply data augmentation, including random horizontal flips and random adjustments to brightness and contrast
- Normalize the pixel values of the images using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]

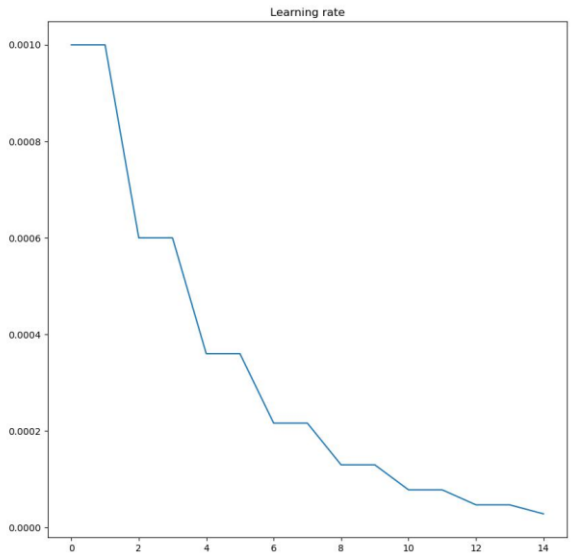
Training Details - Hyperparameters

	Teacher	Student
Epoch	15	30
Batch Size	64	64
Optimizer	Adam	Adam
Learning Rate	Initial learning rate is set to 1e-3, decreasing by a factor of 0.6 every 2 epochs	Cycle learning rate schedule between 1e-3 and 1e-2
Data Augmentation	- Random horizontal flipping with a 50% probability - Brightness and contrast jitter within ±10% ranges	
Temperature	x	Linearly decrease from 10 to 5

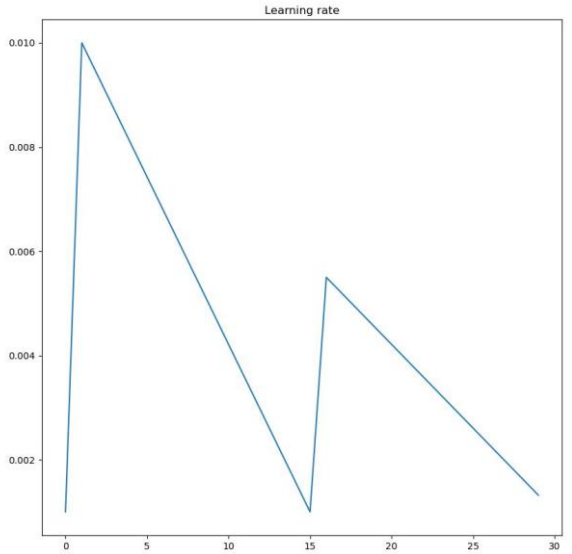
Training Details - Hyperparameters

- Learning rate

Teacher



Student (cycle learning rate)

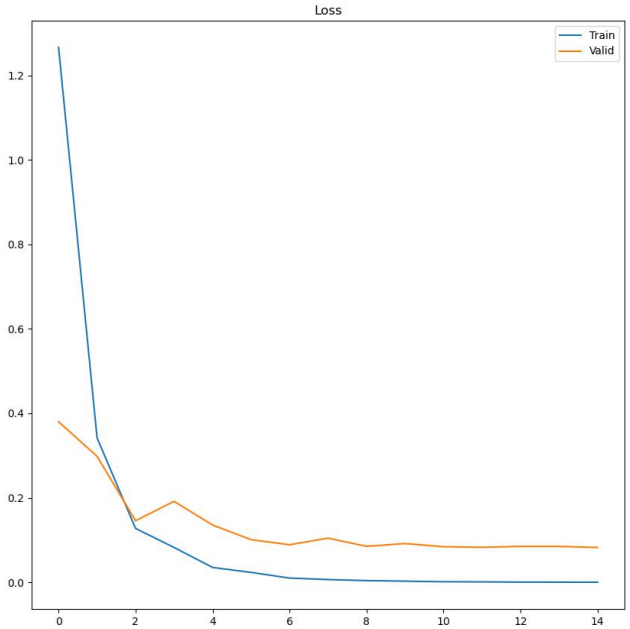


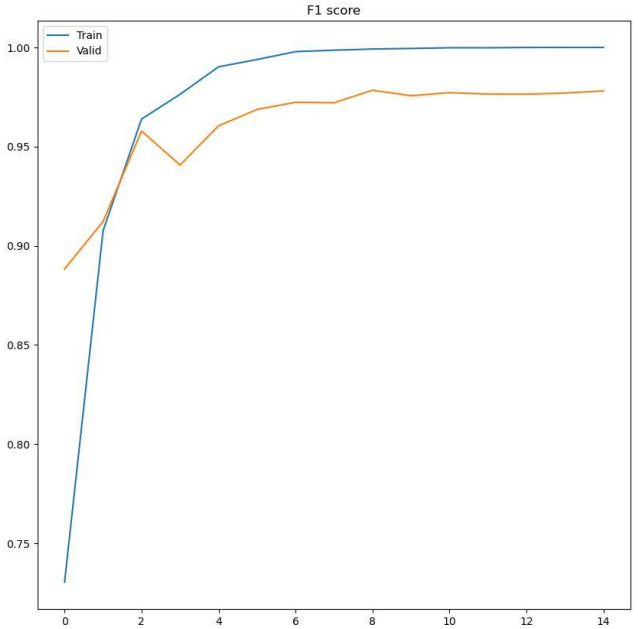
Training Details - Model Selection

- Validation F1-score is calculated after each epoch, and the model with the highest validation F1-score is selected as the final model for making predictions

Learning Curve

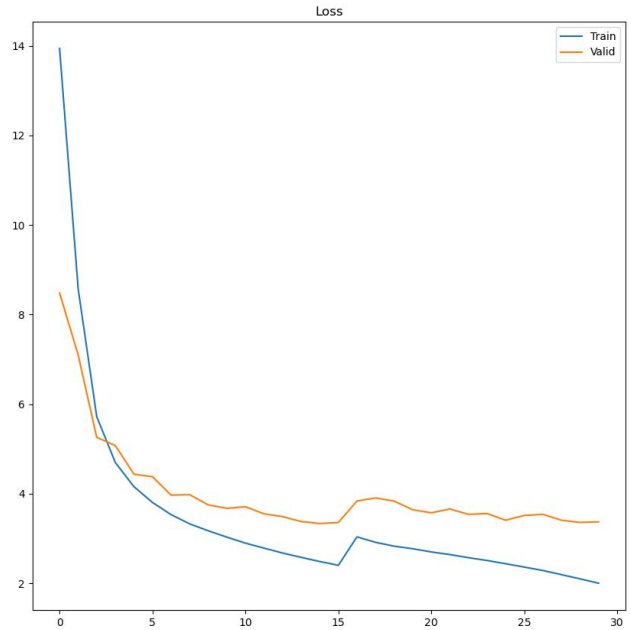
- Teacher model

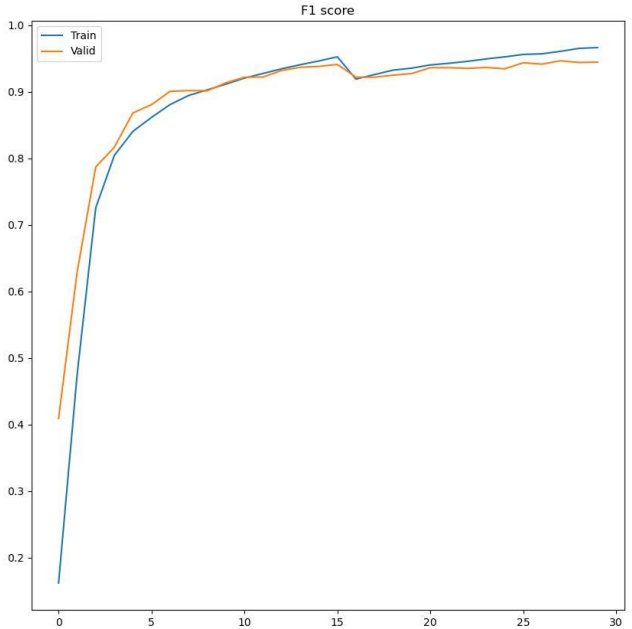




Learning Curve

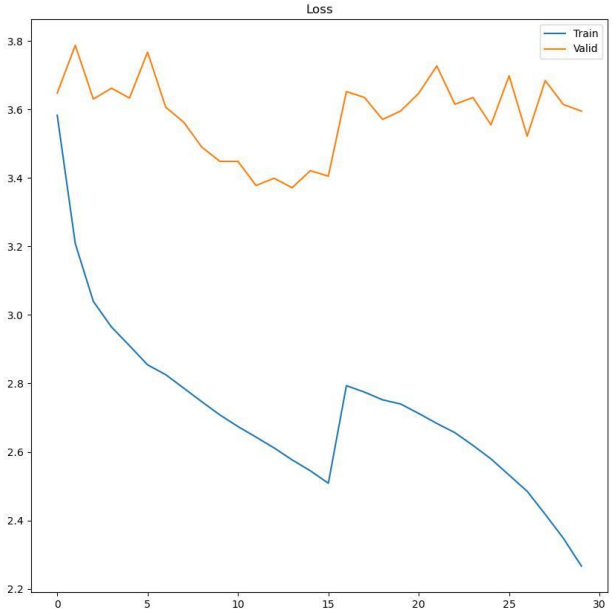
- Student model

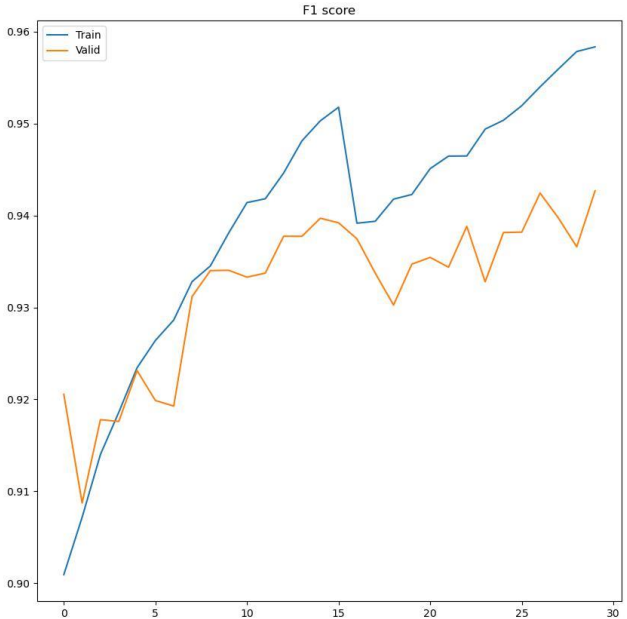




Learning Curve

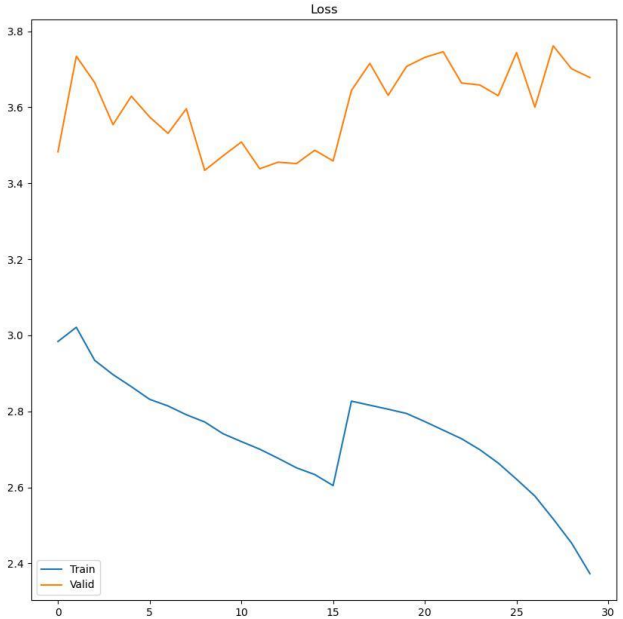
- Student model after 50% pruning

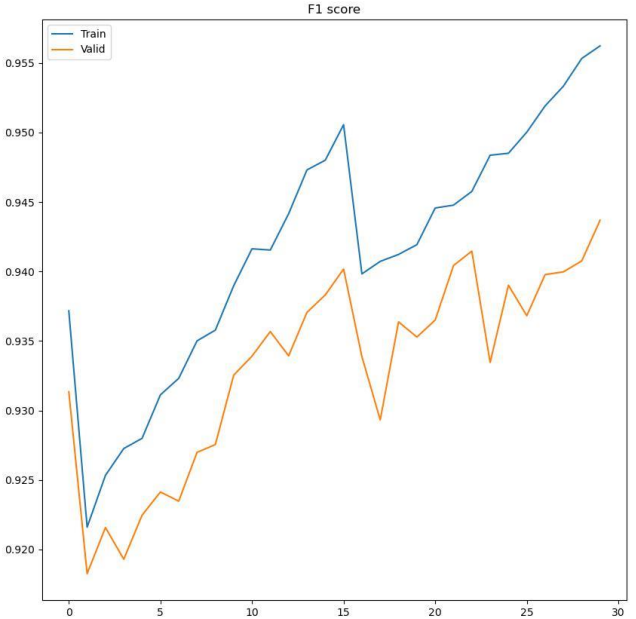




Learning Curve

- Student model after 60% pruning





Outline

1. Baseline

- Knowledge Distillation
- Model Pruning

2. Our Method

- Training Pipeline
- Model Architecture
- Decreasing Temperature
- Training Details
- Learning Curve

3. Experiment Result

- **Different Methods**
- **Student Model Architecture**

Ablation Study

- Different Methods

	Test F1-score	Parameters size
Teacher	0.993	8,756,461
Student	0.916	189,677
Knowledge distillation	0.969	189,677
+ Decreasing temperature	0.972	189,677
+ Pruning 50%	0.97	98,117
+ Pruning 60%	**0.969**	**79,805**

Ablation Study

- Student Model Architecture

	Test F1-score	Parameters size
ShuffleNet v2-0.5x (conv5 out=1024)	0.986	879,917
ShuffleNet v2-0.5x (conv5 out=512)	0.982	511,789
ShuffleNet v2-0.5x (conv5 out=128)	0.977	235,693
ShuffleNet v2-0.5x (conv5 out=64)	**0.969**	**189,677**
ShuffleNet v2-0.5x (conv5 out=32)	0.948	166,669