# Search to Distill: Pearls are Every where but not the Eyes

报告人：时间：2020年7月9日

# Outline

- Motivation
- Method
- Search space
- Experiments

# motivation

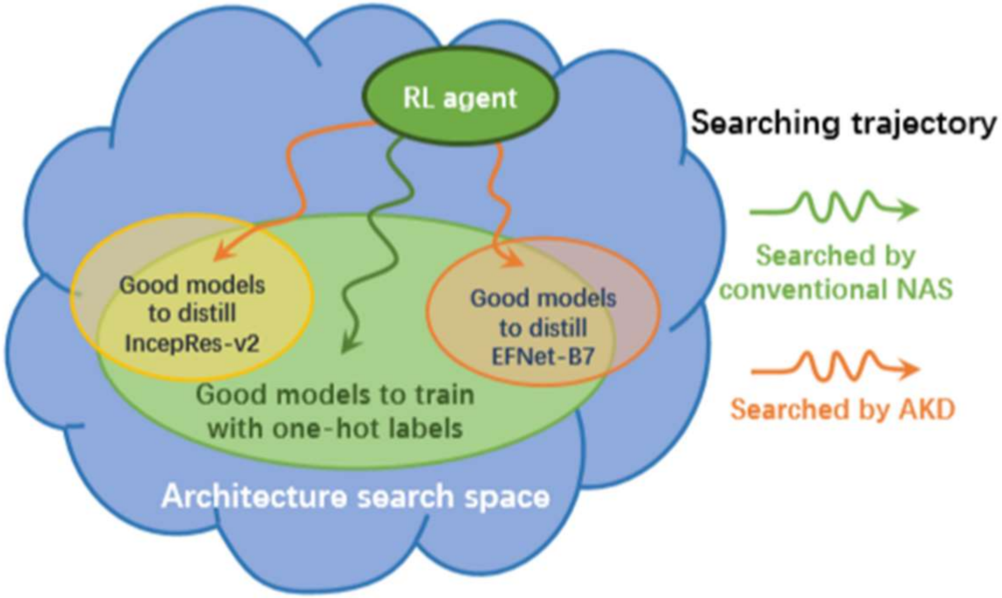



Figure 1. Searching neural architectures by the proposed AKD and conventional NAS [34] lead to different optimal architectures.

| Teachers | Student1 | Student2 | Comparison |
|---|---|---|---|
| EfficientNet-B7 [35] | 65.8% | 66.6% | student1 < student2 |
| Inception-ResNet-v2 [32] | 67.4% | 66.1% | student1 > student2 |

Table 1. ImageNet accuracy for students with different teachers.

# motivation

S2 is the pearl (best student) in the eye of T(A)

| Tag | Model name | Input size | Top-1 accuracy |
|---|---|---|---|
| T(A) | EfficientNet-B7 [35] | 600 | 84.4 |
| T(B) | PNASNet-large [18] | 331 | 82.9 74 |
| T(C) | SE-ResNet-154 [11] | 224 | 81.33 |
| T(D) | PolyNet [41] | 331 | 81.23 |
| T(E) | Inception-ResNet-v2 [32] | 299 | 80.217 |
| T(F) | ResNeXt-101 [38] | 224 | 79.431 |
| T(G) | Wide-ResNet-101 [40] | 224 | 78.84 |
| T(H) | ResNet-152 [5] | 224 | 78.31 |

Table 2. A comparison of popular off-the-shelf models, sorted by top-1 accuracy.

| | Teacher models | | | |
|---|---|---|---|---|
| GT | T(A) | T(B) | T(E) | T(F) |
| $S_3$ (65.6) | $S_2$ (66.6) | $S_3$ (66.9) | $S_1$ (67.4) | $S_5$ (67.1) |
| $S_4$ (65.6) | $S_3$ (66.5) | $S_5$ (66.4) | $S_4$ (67.0) | $S_1$ (67.1) |
| $S_5$ (65.5) | $S_4$ (66.3) | $S_4$ (66.1) | $S_5$ (66.9) | $S_4$ (66.6) |
| $S_1$ (65.5) | $S_5$ (66.0) | $S_1$ (65.7) | $S_3$ (66.5) | $S_3$ (66.3) |
| $S_2$ (65.4) | $S_1$ (65.8) | $S_2$ (65.4) | $S_2$ (66.1) | $S_2$ (66.0) |

Distilling the same teacher model to different students leads to different performance results,
and no student architecture produces the best results across all teacher networks.

# motivation

| | T(A) | T(B) | T(C) | T(D) | T(E) | T(F) | T(G) | T(H) | GT |
|---|---|---|---|---|---|---|---|---|---|
| T(A) | 0.00 | 0.12 | 0.15 | 0.87 | 0.23 | 0.87 | 0.80 | 0.86 | 1.50 |
| T(B) | 0.11 | 0.00 | 0.13 | 0.61 | 0.15 | 0.62 | 0.57 | 0.62 | 1.39 |
| T(C) | 0.14 | 0.13 | 0.00 | 0.54 | 0.16 | 0.54 | 0.48 | 0.52 | 1.35 |
| T(D) | 0.32 | 0.26 | 0.24 | 0.00 | 0.20 | 0.16 | 0.16 | 0.18 | 1.08 |
| T(E) | 0.21 | 0.16 | 0.17 | 0.41 | 0.00 | 0.44 | 0.39 | 0.37 | 1.47 |
| T(F) | 0.31 | 0.26 | 0.23 | 0.15 | 0.22 | 0.00 | 0.12 | 0.18 | 0.87 |
| T(G) | 0.29 | 0.25 | 0.22 | 0.17 | 0.21 | 0.14 | 0.00 | 0.18 | 0.86 |
| T(H) | 0.44 | 0.39 | 0.36 | 0.36 | 0.29 | 0.34 | 0.30 | 0.00 | 1.72 |
| GT | 0.33 | 0.33 | 0.29 | 0.21 | 0.33 | 0.15 | 0.14 | 0.33 | 0.00 |

**(a) KL(p||q)**

| | T(A) | T(B) | T(C) | T(D) | T(E) | T(F) | T(G) | T(H) | GT |
|---|---|---|---|---|---|---|---|---|---|
| T(A) | 1.00 | 0.96 | 0.96 | 0.95 | 0.94 | 0.96 | 0.96 | 0.92 | 0.96 |
| T(B) | 0.96 | 1.00 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.92 | 0.94 |
| T(C) | 0.96 | 0.95 | 1.00 | 0.94 | 0.94 | 0.95 | 0.95 | 0.92 | 0.95 |
| T(D) | 0.95 | 0.95 | 0.94 | 1.00 | 0.94 | 0.95 | 0.95 | 0.92 | 0.94 |
| T(E) | 0.94 | 0.95 | 0.94 | 0.94 | 1.00 | 0.94 | 0.94 | 0.92 | 0.93 |
| T(F) | 0.96 | 0.95 | 0.95 | 0.95 | 0.94 | 1.00 | 0.96 | 0.92 | 0.96 |
| T(G) | 0.96 | 0.95 | 0.95 | 0.95 | 0.94 | 0.96 | 1.00 | 0.92 | 0.96 |
| T(H) | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 1.00 | 0.90 |
| GT | 0.96 | 0.94 | 0.95 | 0.94 | 0.93 | 0.96 | 0.96 | 0.90 | 1.00 |

**(b) Top-1 matching ratio**

Figure 2. Confusion matrix for models' outputs, p or q denotes softmax probability output, GT means one-hot label.

These observations inspire us to rethink the knowledge in a teacher network, which we argue depends not only on its parameters or performance but also on its architecture.

# Outline

- Motivation
- Method
- Search space
- Experiments

# Method

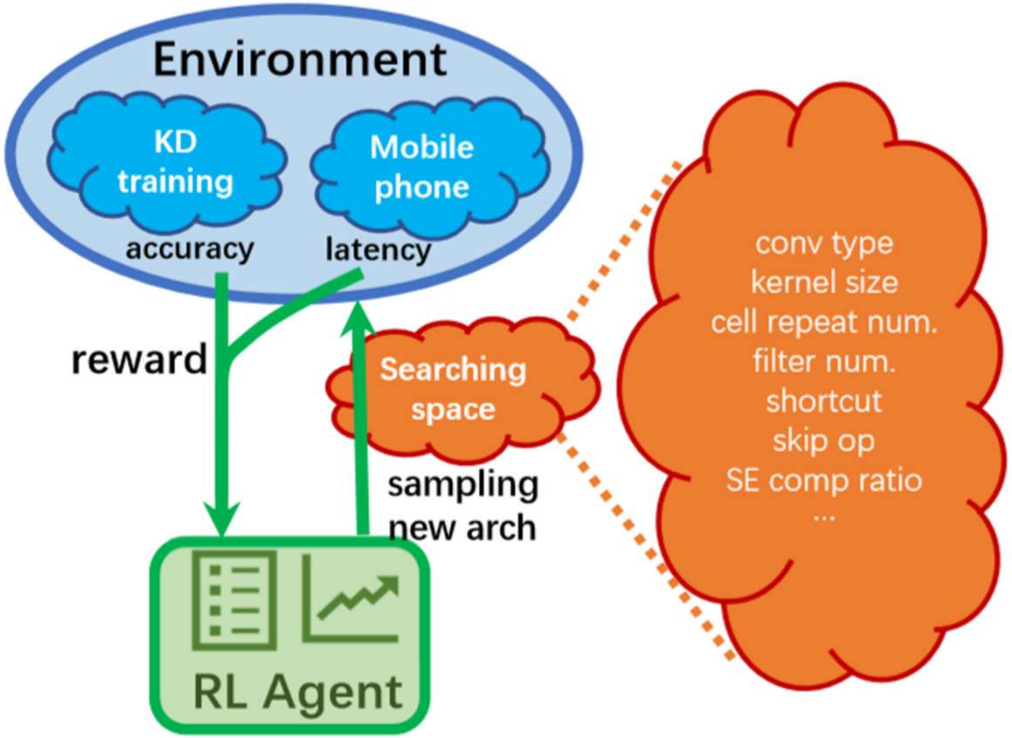



Figure 3. Pipeline of searching process in AKD. there are three core components: environment, RL agent and search space.

Search space:
the number of layers,
the convolution and skip operation type,
conv kernel size,
squeeze-and-excite ratio,
input/output filter size,

# Method

We treat each sampled model as a student, and distill teacher's knowledge by training on the mini-train for 5 epochs

Then evaluating on the mimi-val to obtain its accuracy.

We do not use a shared weight among different sampled models

Each experiment takes about 5 days on 200 TPUv2 Donut devices

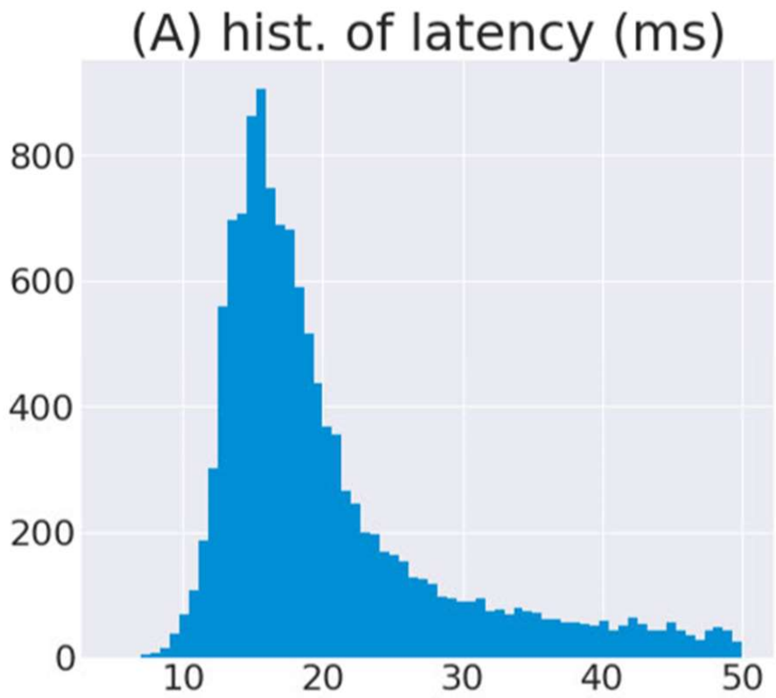
# Outline

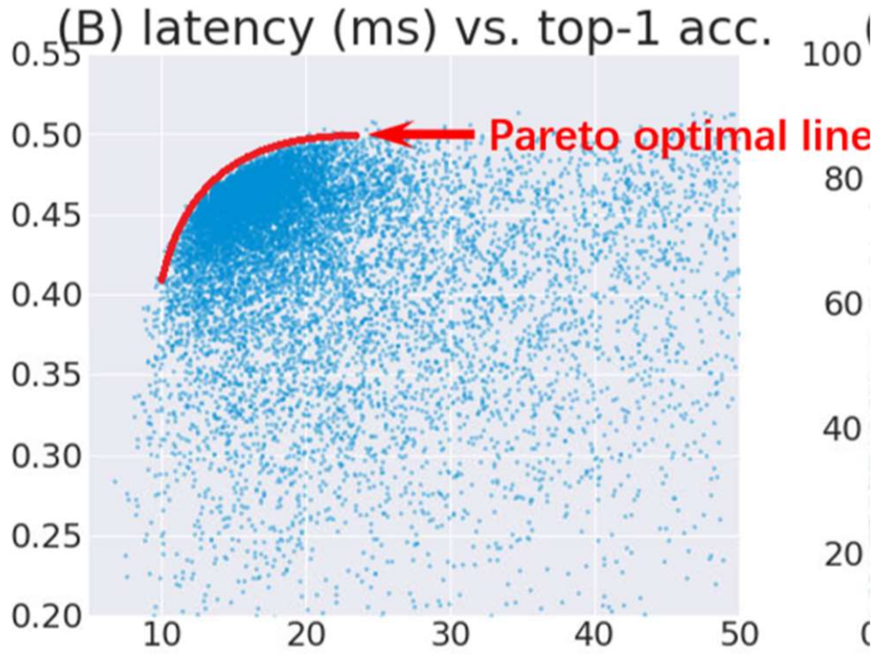
- Motivation
- Method
- Experiments

# Experiment

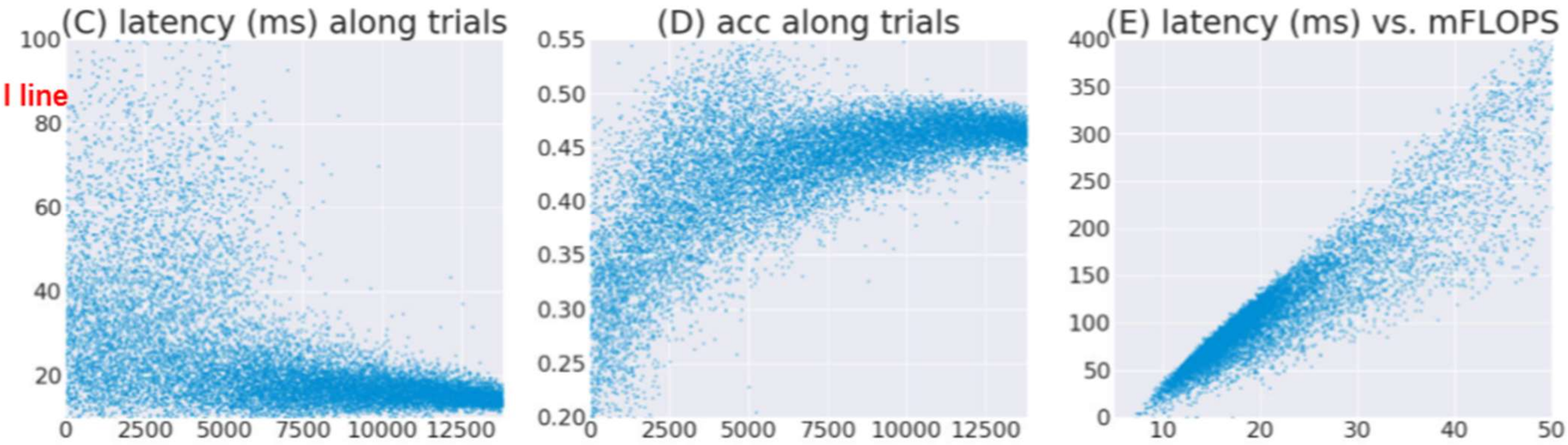
Understanding the searching process

Latency target:15ms



# Experiment



(C) latency (ms) along trials
l line
(D) acc along trials
(E) latency (ms) vs. mFLOPS

# Experiment

**how different the AKDNet and NASNet**

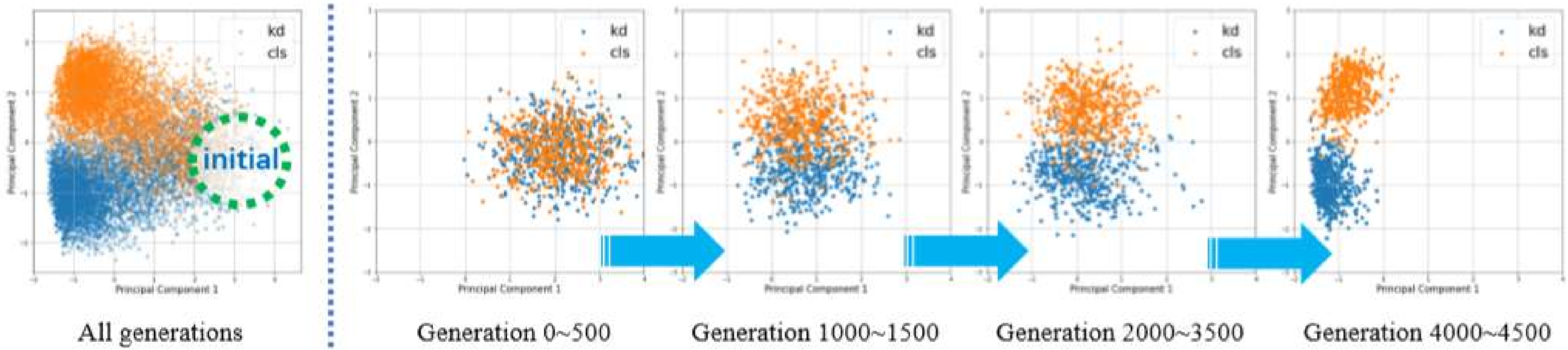
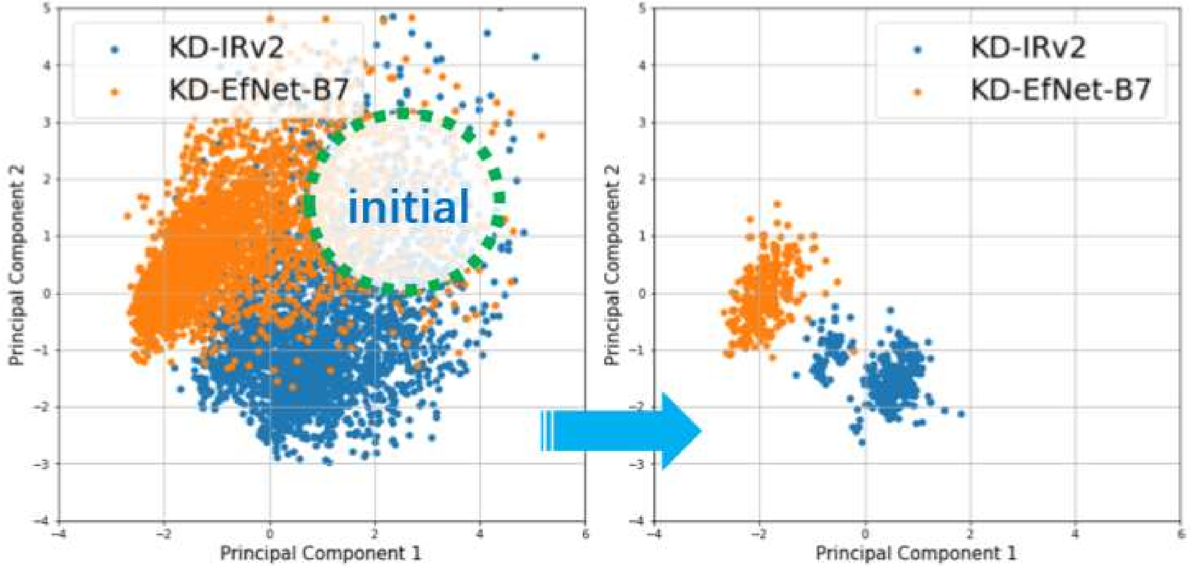



Figure 5. The architecture evolves during searching. Each dot represents an architecture. Different colors indicate different NAS policies – orange for conventional NAS and blue for AKD based NAS. PCA is used for visualization. Best view in color.

# Experiment

**Existence of structural knowledge**

If two identical RL agents perform AKD on two different teacher architectures, will they converge to different area in search space?




All generations Last 200 generations

Figure 6. All generations searched by AKD on two different teachers. Their final generations converge into different areas. This proves the structural knowledge does exist in the teacher model.

# Experiment

If two different RL agents perform AKD on the same teacher, will they converge to similar area?

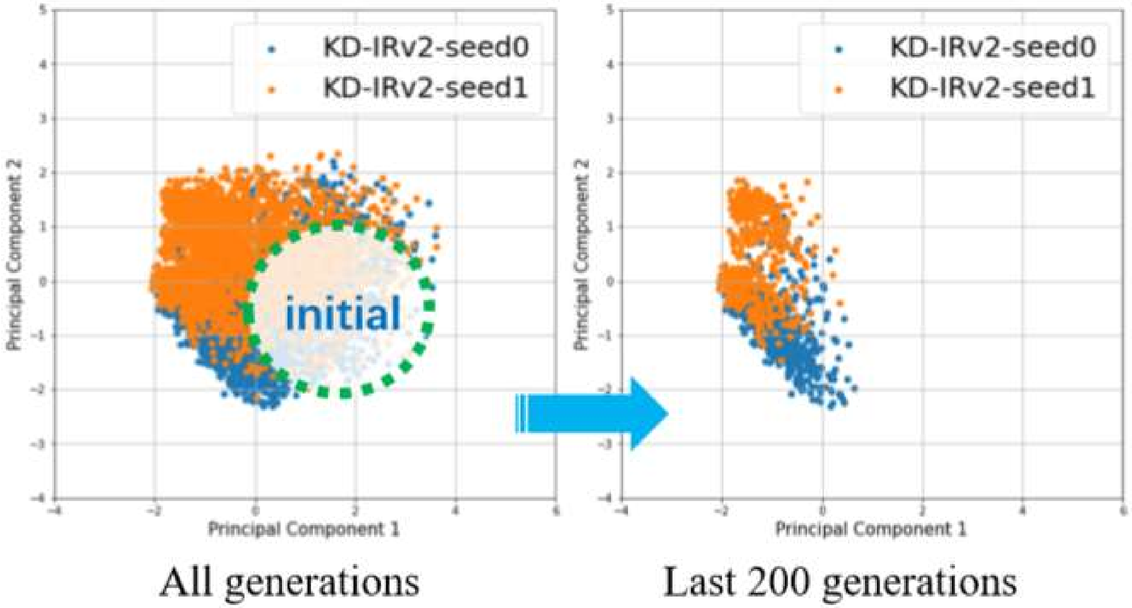
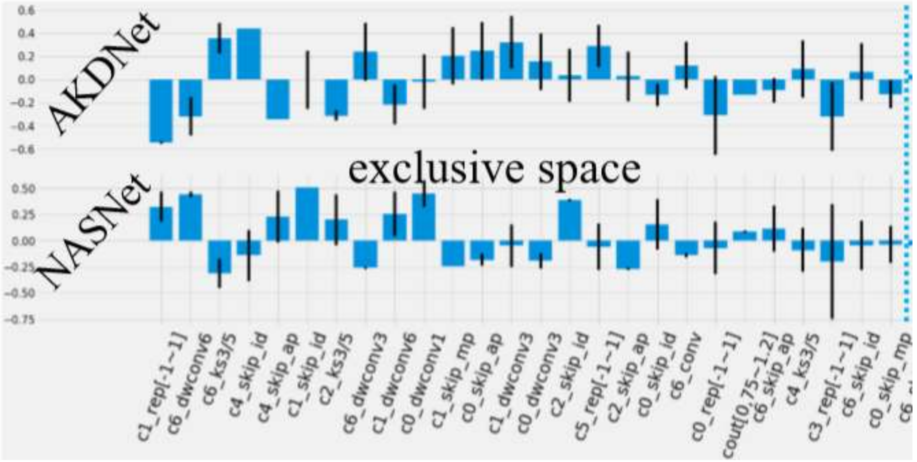
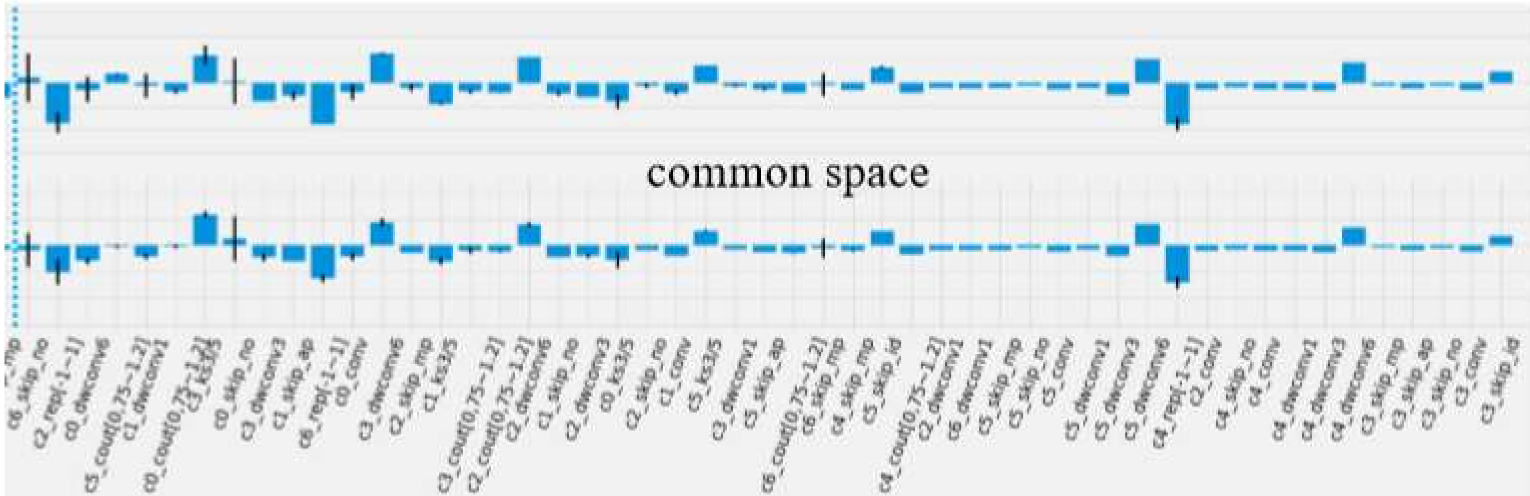



Figure 7. All generations searched by AKD on the same teacher model but different RL agents. Their final generations converge into the same area. Note that there are a large amount of blue dots overlapped by the orange dots. Best view in color.

# Experiment

**Opertaion distribution**



# Experiment

AKDNet becomes KD-friendly along searching

$$[\mathrm{KD(AKDNet)} - \mathrm{CLS(AKDNet)}] - [\mathrm{KD(NASNet)} - \mathrm{CLS(NASNet)}], \tag{1}$$

| Generation | initial | $\sim$1k | $\sim$3k | $\sim$10k |
|---|---|---|---|---|
| Winning ratio | 10 / (20-2) | 12 / 20 | 16 / 20 | **18 / 20** |
| Average gain | -0.07 | 0.10 | 0.46 | **1.05** |

Table 4. Relative performance gain between KD(AKDNet) and KD(NASNet) during different searching stages.

# Experiment

Ablation study on ImageNet

| Latency | searching by | training by | top-1 | top-5 |
|---|---|---|---|---|
| $15\pm1$ ms | hard label | hard label | 59.73 | 81.39 |
| | hard label | distillation | 63.9 | 84.26 |
| | distillation | hard label | 61.4 | 83.1 |
| | **distillation** | **distillation** | **66.5** | **87.5** |
| $25\pm1$ ms | hard label | hard label | 67.0 | 87.4 |
| | hard label | distillation | 68.1 | 88.0 |
| | distillation | hard label | 67.2 | 87.5 |
| | **distillation** | **distillation** | **69.6** | **89.1** |
| $75\pm1$ ms | hard label | hard label | 73.0 | 92.1 |
| | hard label | distillation | 74.7 | 92.54 |
| | distillation | hard label | 73.6 | 92.2 |
| | **distillation** | **distillation** | **75.5** | **93.1** |

# Experiment

whether AKD overfits the original KD policy

| Training method | Latency | Advanced KD method | | |
|---|---|---|---|---|
| | | TA-KD | RCO-KD | CC-KD |
| MNet-v2 w/o KD | 33ms | 65.4 | | |
| MNet-v2 w/ KD | | 67.6 ↑**2.2** | 68.2 ↑**2.8** | 67.7 ↑**2.3** |
| AKDNet-M w/o KD | 32.8ms | 68.9 | | |
| AKDNet-M w/ KD | | 72.0 ↑**3.1** | 72.4 ↑**3.5** | 72.2 ↑**3.3** |

Table 6. AKDNet transfers to other advanced KD method. 'MNet-v2' denotes the MobileNet-v2 0.5×. The 'M' in AKDNet-M denotes the 32.8ms version of ADKNet. Even with a much higher baseline, AKDNet consistently brings considerable gain ($\sim$ 1%) under each KD method compared to MobileNet-v2.

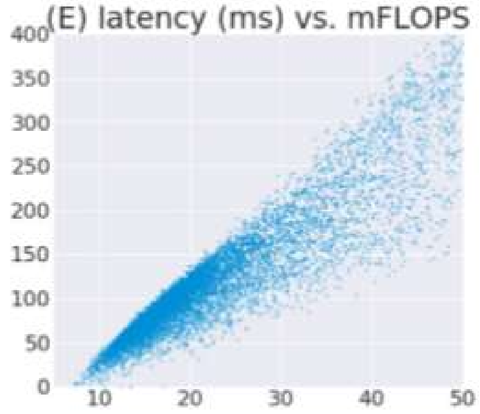
# Experiment

Compare with SOTA architectures

| Latency | architecture | with KD? | top-1 | top-5 |
|---|---|---|---|---|
| 15∼20 ms | AKDNet | | 61.4 | 83.1 |
| | AKDNet | ✓ | ↑2.6 | ↑3.24 |
| | AKDNet | RCO-KD | ↑3.1 | ↑3.8 |
| | MNet-v2-a | | 59.2 | 79.8 |
| | MNet-v2-a | ✓ | ↑1.4 | ↑2.1 |
| | MNASNet-a | | 62.2 | 83.5 |
| | MNASNet-a | ✓ | ↑1.49 | ↑2.3 |
| | MNet-v3-a | | 64.1 | 85.0 |
| | MNet-v3-a | ✓ | ↑1.3 | ↑ 2.2 |
| 25∼27 ms | AKDNet | | 67.2 | 87.5 |
| | AKDNet | ✓ | ↑2.4 | ↑1.6 |
| | AKDNet | RCO-KD | ↑2.8 | ↑1.5 |
| | MNASNet-b | | 66.0 | 86.1 |
| | MNASNet-b | ✓ | ↑1.1 | ↑0.6 |

# Experiment

Latency vs. FLOPS




$$3.4 \times (\text{latency} - 7) \leq \text{mFLOPS} \leq 10.47 \times (\text{latency} - 7)$$

| | searching by | training by | top-1 | top-5 |
|---|---|---|---|---|
| NASNet | hard label | hard label | 69.92 | 89.1 |
| | hard label | distillation | 71.2 | 90.4 |
| AKD | distillation | hard label | 70.0 | 89.4 |
| | **distillation** | **distillation** | **72.1** | **91.7** |
| | **distillation** | **RCO-KD** | **73.0** | **92.2** |

# Experiment

**Towards million level face retrieval**

| Training method | KD method | Distractor num. 1e3 | 1e4 | 1e5 | 1e6 |
|---|---|---|---|---|---|
| Teacher | - | 99.56 | 99.3 | 99.0 | 98.2 |
| MNet-v2 | - | 91.49 | 84.45 | 75.6 | 65.9 |
| MNet-v2 | CC-KD | 97.93 | 95.76 | 91.99 | 86.29 |
| MNet-v2 | RCO-KD | 98.29 | 95.01 | 90.97 | 85.9 |
| **AKDNet-M** | - | 93.8 | 86.4 | 78.2 | 68.6 |
| **AKDNet-M** | CC-KD | 98.26 | 97.48 | 93.85 | 88.41 |
| **AKDNet-M** | RCO-KD | 98.42 | 97.56 | 94.1 | 90.2 |

MS-Celeb1M

Table 9. Transfer the AKDNet on MegaFace. The teacher model in all KD settings is Inception-ResNet-v2.

# Experiment

| Training Method | Architecture | Training method | Distractor num. | |
|---|---|---|---|---|
| | | | 1e5 | 1e6 |
| Ensemble Teacher | HRNet-w48 [36] + R100 [2] + EPolyFace [20] + IncRes-v2 [33] + SE154 [12] | - | 99.6 | 99.3 |
| AKDNet-M | AKDNet | hard label | 78.2 | 68.6 |
| AKDNet-M | AKDNet | original-KD | 94.9 | 90.1 |
| AKDNet-M | AKDNet | RCO-KD | **95.7** | **90.9** |

**The End!**