# A Scalable Convolutional Neural Network for Task-specified Scenarios via Knowledge Distillation

**Mengnan Shi, Fei Qin*, Qixiang Ye, Zhenjun Han, Jianbin Jiao**

University of Chinese Academy of Sciences, Beijing, China

## Motivation

We explore the redundancy in convolutional neural network, which scales with the complexity of vision tasks. We propose a task-specified knowledge distillation algorithm to derive a simplified model, which suits the resource constraint front-end systems, like smart phones, UAVs, robot arms, self-driving vehicles, etc.

●Considering front-end systems must work with limited computation resources, a reduced CNN with less computation cost is required.

| Model | Computation/FLOPs | Storage/MB | Memory/MB | **Time**/s |
|---|---|---|---|---|
| AlexNet | 7.29e+8 | 232.56 | 264.74 | **2.93** |
| CNN-S | 2.94e+9 | 392.57 | 468.90 | **10.58** |

Table1. Test-phase cost of AlexNet and CNN-S on Huawei® Mate 7[1]

●Considering most front-end visual systems are interested in only a limited range of visual targets, the removing of task-specified redundancy helps create a simpler CNN and can promote a wide range of potential applications.

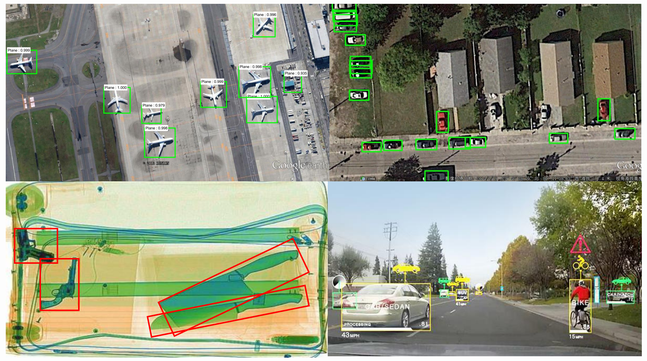

Fig1. Sample task-specified scenarios on front-end systems

## Method

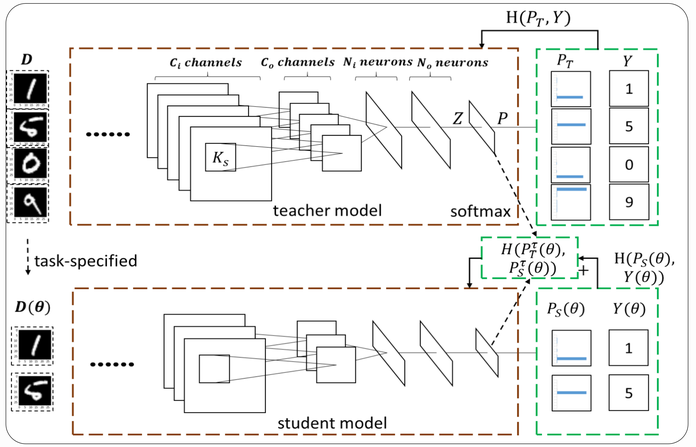


Fig2. The scheme of task-specified knowledge distillation

**Algorithm 1** Task-specified Knowledge Distillation

**Input:** $D, T, \theta$

**Output:** $S(\theta)$

1: add a softmax layer to $T$
2: use $D$ to train the teacher model $T$
3: use $T$ to capture soft targets $P_T^\tau(\theta)$ from each sample in $D(\theta)$
4: set the architecture of the student model $S(\theta)$
5: train the student model with soft targets $P_T^\tau(\theta)$ and $D(\theta)$, iteratively until the accuracy converges

$$S(\theta) = \arg\min_{S(\theta)} L_{KD}(\theta)$$

$$L_{KD}(\theta) = \frac{1}{N}\sum_N((1-\lambda)H(Y(\theta), P_S(\theta)) + \lambda H(P_T^\tau(\theta), P_S^\tau(\theta)))$$

$$H(P,Q) = -\sum_i p_i log(q_i)$$

We propose the task-specified knowledge distillation algorithm to derive a simplified model.

The new model could satisfy the constraints of both the computation cost and residue accuracy.

The detailed algorithm and loss function are shown left.

## Results

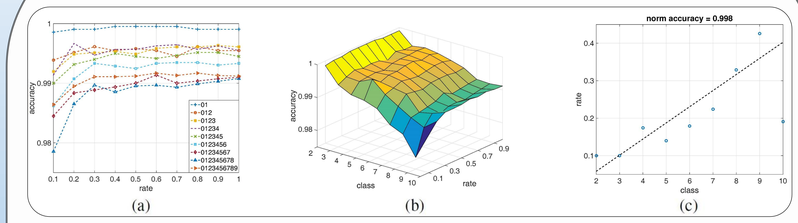


Fig3. (a) Results of MNIST; (b) re-drawed 3D version; (c) relationship between rate and class number given normalized accuracy

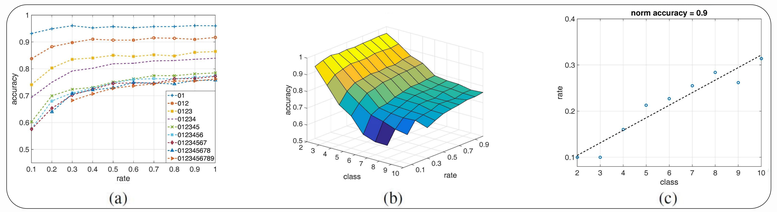


Fig4. (a) Results of CIFAR10; (b) re-drawed 3D version; (c) relationship between rate and class number given normalized accuracy

Given the same normalized accuracy, the more simpler the task is, i.e., the less the class number becomes, the more redundancy in the teacher model could be reduced for MNIST and CIFAR10.

## Contact

**Email:** shimengnan14@mails.ucas.ac.cn; fqin1982@ucas.ac.cn

**Info:** 中国科学院大学电子学院教学科研岗位
http://www.ucas.ac.cn/site/157?u=63463
电话: 86-10-69671866 邮箱: pyxia@ucas.ac.cn

**This work is published in The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017.**