A Contrastive Knowledge Transfer Framework for Model Compression and Transfer Learning (Oral Session in ICASSP 2023)

Kaiqi Zhao, Yitao Chen, Ming Zhao

Arizona State University

http://visa.lab.asu.edu

VISA Virtualized Infrastructures, Systems, & Applications

Model Compression & Transfer Learning

- Deep learning is moving towards edge
  - DNNs are resource-demanding
  - But edge devices are resource-constrained

→ Model Compression

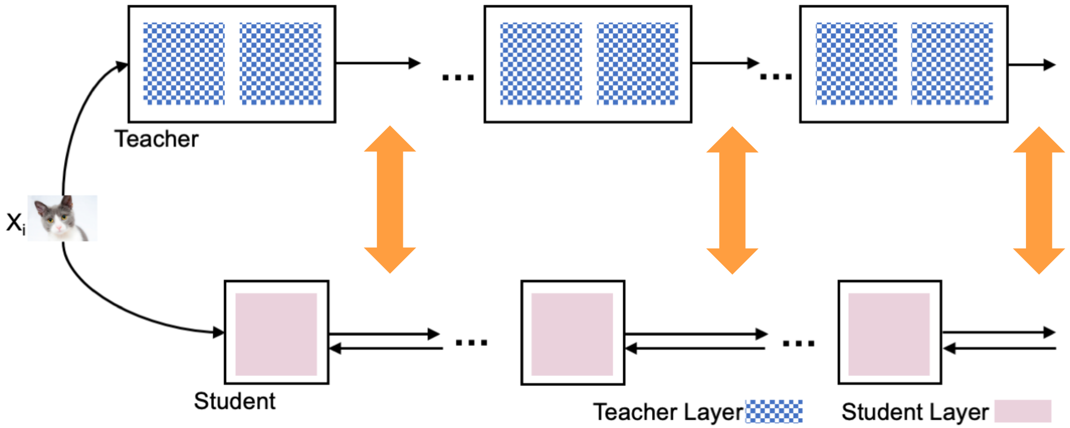
- DNN training requires sufficient labeled data
  - But many real-world scenarios do not have sufficient labeled data

→ Transfer Learning

Source Domain (Abundant labeled data) → Transfer Knowledge → Target Domain (Unavailable labels)

Knowledge Transfer

- Knowledge Transfer (KT)
  - Minimize the difference of the conditionally independent output distributions
  - Transfer soft logits (softmax outputs)
    - Knowledge Distillation (KD)
  - Transfer intermediate representations
    - Attention Transfer (AT)



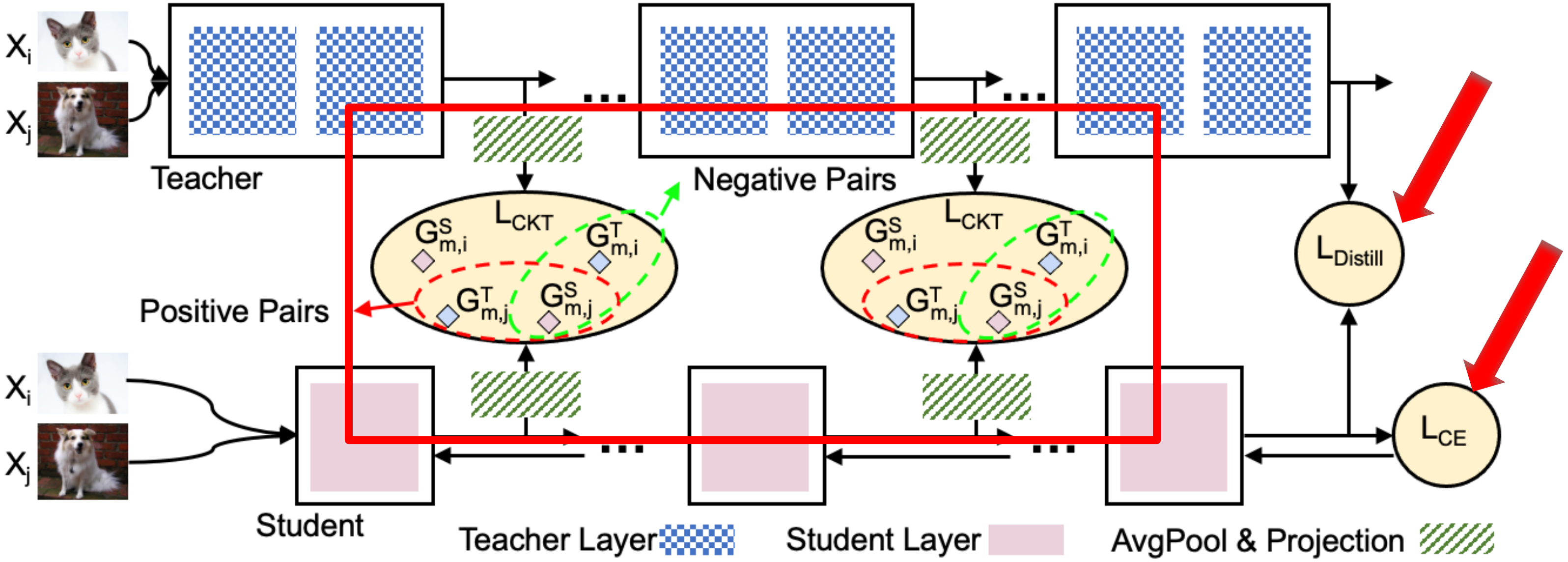
- Limitations
  - Overlook the structural knowledge from the intermediate representations
    - High-dimension
    - Crucial for guiding gradient updates
  - Lack a commonly agreed theory → Challenging to generalize
  - Fail to consistently outperform the conventional KD

Contrastive Knowledge Transfer Framework (CKTF)

- Optimization objective

$$L = \boxed{\gamma L_{CE}(Y, S_h)} + \boxed{L_{CKT}(\{T_m\}_{m=1}^{M}, \{S_m\}_{m=1}^{M}, T_h, S_h)} + \boxed{\theta L_{Distill}(T_h, S_h)}$$

  - Cross entropy loss with the ground truth labels: $L_{CE}(Y, S_h), \gamma \in [0, 1]$
  - Contrastive loss: $L_{CKT}(\{T_m\}_{m=1}^{M}, \{S_m\}_{m=1}^{M}, T_h, S_h)$
  - Distillation loss from other KT methods: $L_{Distill}(T_h, S_h), \theta \in [0, 1]$



Process Intermediate Representations

- Intermediate representations
  - Different dimensions between the teacher and student
  - Huge feature dimensions → Memory issues or Increase the training time
    - E.g., One intermediate representation of ResNet-50 on ImageNet: about 8.39 millions

- Process
  - Apply an average pooling → Reduce features

$$\bar{S}_m = AvgPool(S_m), \bar{T}_m = AvgPool(T_m)$$

  - Apply a reshape function → Reduce space from 4D to 2D

$$H_m^S = h(\bar{S}_m), H_m^T = h(\bar{T}_m)$$

  - Apply the projection network → Same dimensions
    - Linear v.s. Multi-Layer Perceptron (MLP)

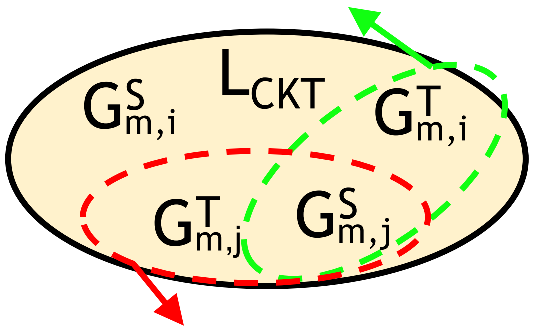
$$G_m^S = g(H_m^S), G_m^T = g(H_m^T)$$

$$S_m \in R^{B \times o_m^S \times k_m^S \times k_m^S}$$
$$T_m \in R^{B \times o_m^t \times k_m^t \times k_m^t}$$

$$G_m^S \in R^{B \times d}$$
$$G_m^T \in R^{B \times d}$$

Construct Contrastive Loss

- Representation pairs
  - Positive representation pairs $(G^S_{m,i}, G^T_{m,i})$ Push Closer
    - Outputs from the same input sample x_i
  - Negative representation pairs $(G^S_{m,i}, G^T_{m,j})$ Push Apart
    - Outputs from two different input samples x_i, x_j



- Contrastive loss on intermediate representations
  - Maximize the lower bound of the mutual information

$$L_{MCKT}\,(G^S_m, G^T_m) = -E\left[log\frac{f\left(G^S_{m,i}, G^T_{m,i}\right)}{\sum_{j=1}^{N} f\left(G^S_{m,i}, G^T_{m,j}\right)}\right]$$

$$f(G^S_{m,i}, G^T_{m,i}) = \frac{\exp(G^S_{m,i}, G^T_{m,i}/\tau)}{\exp\left(G^S_{m,i}, \frac{G^T_{m,i}}{\tau}\right) + N/N_d}$$

We are the first to construct multiple contrastive objectives on the intermediate representations of image classification models for KT

Construct Contrastive Loss (Cont.)

- Contrastive loss on penultimate representations

$$L_{PCKT}\left(S_h, T_h\right) = -E\left[log \frac{f\left(S_{h,i}, T_{h,i}\right)}{\sum_{j=1}^{N} f\left(S_{h,i}, T_{h,j}\right)}\right]$$

- Contrastive loss
  - Weighted sum of L_{MCKT} and L_{PCKT}

Contrastive loss on penultimate representations

$$L_{CKT} = \alpha_1 \sum_{m=1}^{M} \mathrm{L_{MCKT}}\left(\mathrm{G_m^S}, \mathrm{G_m^T}\right) + \alpha_2 L_{PCKT}\left(S_h, T_h\right)$$

Contrastive loss on intermediate representations

Model Compression Results

- Outperform
  - KD by 0.5% to 2.41%
  - Other KT by 0.04% to 11.59%
  - CRD by 0.04% to 0.97%
  - W/o KT 0.95% to 4.41%

DataSet	CIFAR-100							Tiny-ImageNet			
Model											
Teacher	WRN-40-2	WRN-40-2	ResNet-56	ResNet-110	ResNet-110	ResNet-32*4	VGG-13	VGG-19	VGG-16	ResNet-34	ResNet-50
Student	WRN-16-2	WRN-40-1	ResNet-20	ResNet-20	ResNet-32	ResNet-8*4	VGG-8	VGG-8	VGG-11	ResNet-10	ResNet-10
Compression Ratio	3.21	3.96	3.10	6.24	3.67	6.03	2.39	5.01	1.59	4.28	4.78
Baselines											
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64	61.62	61.35	65.38	65.34
Student (w/o KT)	73.26	73.54	69.06	69.06	71.14	72.5	70.36	54.61	58.60	58.01	58.01
Method											
KD [2]	74.92	73.54	70.66	70.67	73.08	73.33	72.98	55.55	62.51	58.92	58.63
FitNet [3]	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)	55.24 (↓)	59.08 (↓)	58.22 (↓)	57.76 (↓)
AT [4]	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)	53.55 (↓)	61.40 (↓)	59.16 (↑)	58.92 (↑)
SP [5]	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)	55.09 (↓)	61.61 (↓)	55.91 (↓)	57.17 (↓)
CC [6]	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.71 (↓)	54.87 (↓)	58.34 (↓)	57.18 (↓)	57.36 (↓)
VID [7]	74.11 (↓)	73.3 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)	54.94 (↓)	60.07 (↓)	58.53 (↓)	57.65 (↓)
RKD [8]	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)	54.13 (↓)	59.96 (↓)	57.35 (↓)	57.05 (↓)
PKT [9]	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)	55.35 (↓)	60.46 (↓)	58.41 (↓)	58.66 (↑)
AB [10]	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)	50.31 (↓)	55.65 (↓)	57.22 (↓)	58.05 (↓)
FT [11]	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)	53.65 (↓)	58.84 (↓)	56.22 (↓)	56.48 (↓)
FSP [12]	72.91 (↓)	N/A	69.95 (↓)	70.11 (↓)	71.89 (↓)	72.62 (↓)	70.23 (↓)	N/A	N/A	N/A	N/A
NST [13]	73.68 (↓)	72.24 (↓)	69.60 (↓)	69.53 (↓)	71.96 (↓)	73.30 (↓)	71.53 (↓)	51.08 (↓)	58.47 (↓)	59.23 (↑)	47.83 (↓)
CRD [14]	75.48 (↑)	74.14 (↑)	71.16 (↑)	71.46 (↑)	73.48 (↑)	75.51 (↑)	73.94 (↑)	56.99 (↑)	62.04 (↓)	60.02 (↑)	59.31 (↑)
CKTF	**75.85** (↑)	**74.49** (↑)	**71.20** (↑)	**71.80** (↑)	**73.84** (↑)	**75.74** (↑)	**74.31** (↑)	**57.57** (↑)	**63.01** (↑)	**60.39** (↑)	**59.42** (↑)
CRD+KD [14]	75.64 (↑)	74.38 (↑)	71.63 (↑)	71.56 (↑)	73.75 (↑)	75.46 (↑)	74.29 (↑)	58.09 (↑)	63.66 (↑)	61.99 (↑)	61.26 (↑)
CKTF+KD	**75.89** (↑)	**74.94** (↑)	**71.86** (↑)	**71.66** (↑)	**74.07** (↑)	**75.97** (↑)	**74.55** (↑)	**58.76** (↑)	**63.97** (↑)	**62.31** (↑)	**61.51** (↑)

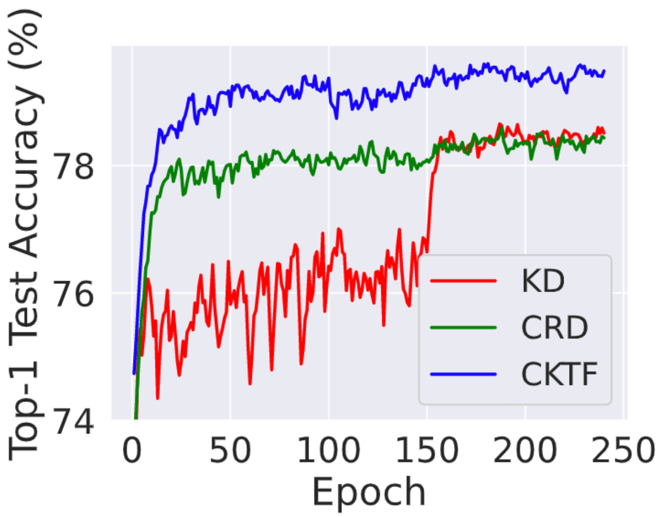
Model Compression Results (Cont.)

- Incorporate KT methods
  - Improve existing KT works by 0.89% to 3.02%
  - Provide a generalized agreement behind knowledge transfer

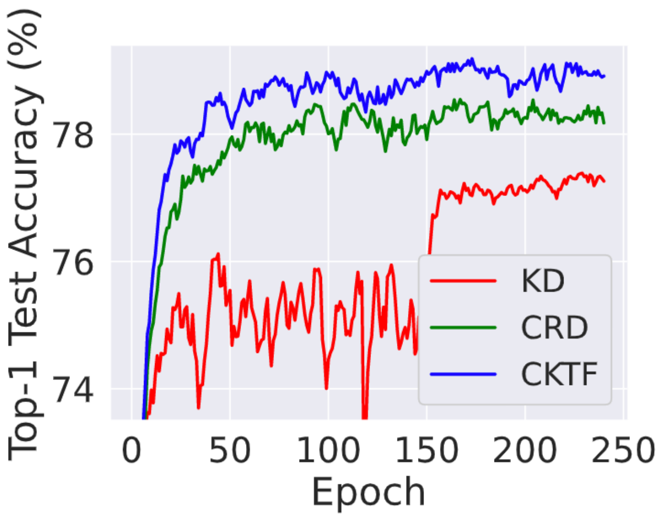
	CKTF +FitNet	CKTF +AT	CKTF +SP	CKTF +CC	CKTF +VID	CKTF +RKD	CKTF +PKT	CKTF +AB	CKTF +FT	CKTF +NST
T: ResNet-32×4 S: ResNet-8×4 (CIFAR-100)	73.18 (1.68 ↑)	74.92 (1.48 ↑)	75.30 (2.36 ↑)	75.86 (2.89 ↑)	75.43 (2.34 ↑)	74.92 (3.02 ↑)	75.82 (2.18 ↑)	75.38 (2.21 ↑)	75.39 (2.53 ↑)	75.08 (1.78 ↑)
T: VGG-19 S: VGG-8 (Tiny-ImageNet)	56.19 (0.95 ↑)	55.33 (1.78 ↑)	56.22 (1.13 ↑)	55.99 (1.12 ↑)	56.34 (1.4 ↑)	55.96 (1.83 ↑)	56.82 (1.47 ↑)	52.63 (2.32 ↑)	56.39 (2.74 ↑)	51.97 (0.89↑)

Transfer Learning Results

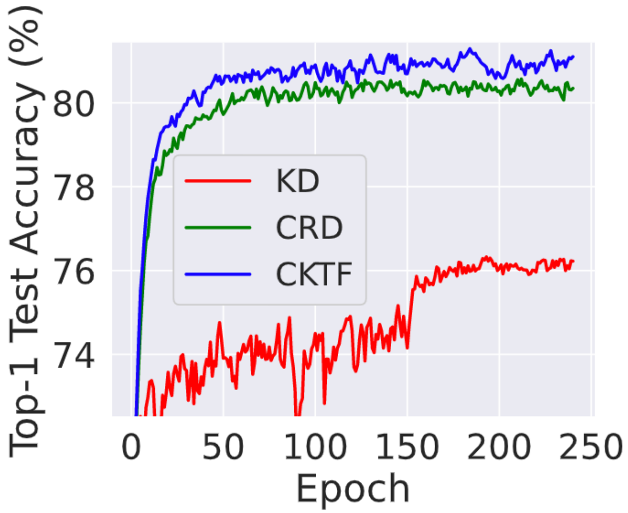
- Tiny-ImageNet (Labeled) → STL-10 (Unlabeled)
- Comparison with KD and CRD
  - Converge speed: Faster
  - Final Top-1 accuracy: Outperform by 0.4% to 4.75%



T: VGG-19 / S: VGG-19



T: VGG-19 / S: VGG-8



T: ResNet-18 / S: ResNet-18

Conclusions and Future Work

- Conclusions
  - Enable the transfer of high-dimension structural knowledge by optimizing multiple contrastive objectives across the intermediate representations
  - Provide a generalized agreement to existing KT methods and increase their accuracy significantly by deriving them as specific cases of CKTF
  - Outperform the existing KT works by 0.04% to 11.59% in model compression and by 0.4% to 4.75% in transfer learning
- Future work
  - Investigate the effectiveness of CKTF in ensemble knowledge transfer
  - Study the effectiveness of CKTF in large-scale language model compression

Acknowledgement

- National Science Foundation
  - Awards CNS-1955593, OAC-2126291

- VISA Lab @ ASU

VISA Research Laboratory for Virtualized Infrastructures, Systems, and Applications