Association for the Advancement of Artificial Intelligence

Understanding the Role of the Projector in Knowledge Distillation

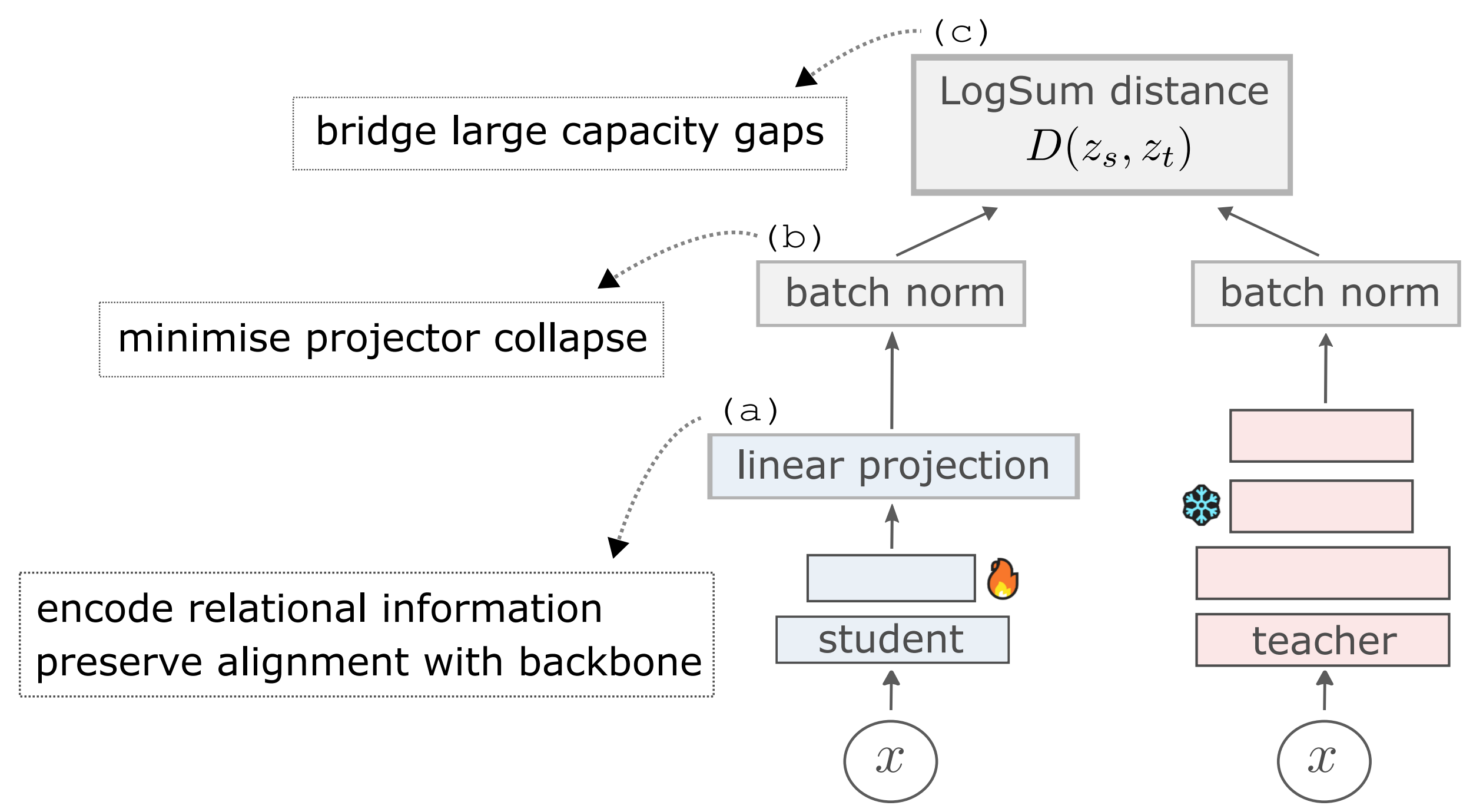
Roy Miles, Krystian Mikolajczyk

Imperial College London

Imperial College London

Overview

We explore the role of three distinct components used for knowledge distillation. In doing so we propose a very simple feature distillation pipeline that achieves state-of-the-art performance across a wide-range of vision tasks.


(c)
bridge large capacity gaps
LogSum distance
$D(z_s, z_t)$
(b)
batch norm
batch norm
minimise projector collapse
(a)
linear projection
encode relational information
preserve alignment with backbone
student
teacher
x
x

Background and Limitations

- Deep neural networks have achieved remarkable success in various applications, ranging from computer vision to natural language processing.
- However, the high computational cost and memory requirements of deep models have limited their deployment in resource-constrained environments.
- Knowledge distillation is a popular technique to address this problem through transferring the knowledge of a large teacher model to that of a smaller student model.
- This technique involves training the student to imitate the output of the teacher, either by directly minimizing the difference between intermediate features or by minimizing the Kullback-Leibler (KL) divergence between their soft predictions.
- Although knowledge distillation has shown to be very effective, there are still some limitations related to the computational and memory overheads in constructing and evaluating the losses, as well as an insufficient theoretical explanation for the underlying core principles.

Contributions

Our contributions can be summarized as follows

- Understanding the coupling and training dynamics of three distinct knowledge distillation principles.
- Projection layers implicitly encode relational information → No need to construct any expensive correlation matrices or memory banks.
- We propose a simple recipe for knowledge distillation using a linear projection, batch normalisation, and a $LogSum$ function.

The Projector Encodes Relational Information

We consider a simple L2 loss between the projected student and teacher features.

$$D(\mathbf{Z}_s, \mathbf{Z}_t\,; \mathbf{W}_p) = \frac{1}{2}\|\mathbf{Z}_s\mathbf{W}_p - \mathbf{Z}_t\|_2^2 \tag{1}$$

Taking the derivative with respect to $\mathbf{W}_p$, we can derive the update rule $\dot{\mathbf{W}}_p$

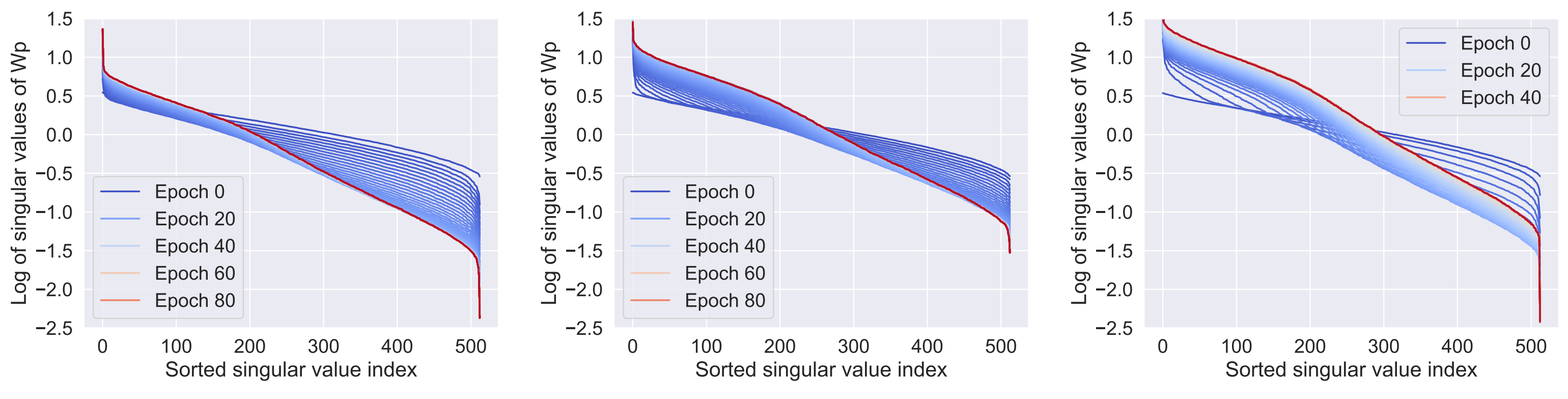
$$\dot{\mathbf{W}}_p = -\frac{\partial D(\mathbf{W}_p)}{\partial \mathbf{W}_p} = -\mathbf{Z}_s^T\mathbf{Z}_s\mathbf{W}_p + \mathbf{Z}_s^T\mathbf{Z}_t, \tag{2}$$

which can be further simplified

$$\boxed{\dot{\mathbf{W}}_p = \mathbf{C}_{st} - \mathbf{C}_s\mathbf{W}_p} \tag{3}$$

where $\mathbf{C}_s = \mathbf{Z}_s^T\mathbf{Z}_s \in \mathbb{R}^{d_s \times d_s}$ and $\mathbf{C}_{st} = \mathbf{Z}_s^T\mathbf{Z}_t \in \mathbb{R}^{d_s \times d_t}$ denote self and cross correlation matrices respectively. This result shows that the training dynamics dictate the relational information being encoded in the weights, while also removing the need to construct any hand-crafted relational objects.

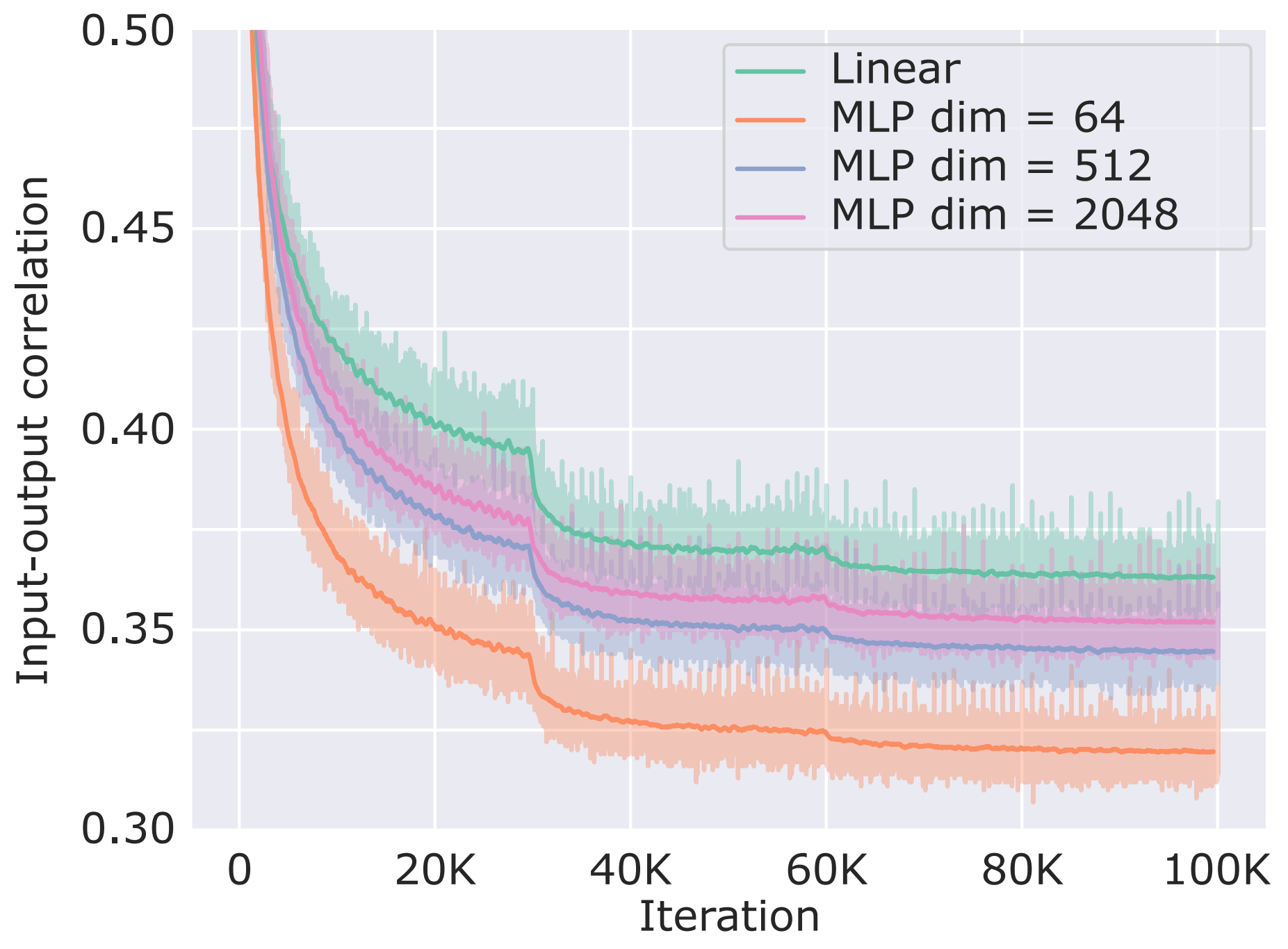
Importance of Feature Normalisation



(a) no normalisation. (b) L2 normalisation. (c) batch normalisation.

Evolution of singular values of the projection weights $\mathbf{W}_p$ under three different representation normalisation schemes. The three curves shows the evolution of singular values for the projector weights when the representations undergo no normalisation, L2 normalisation, and batch norm respectively.

Using Non-Linear Projections



Correlation between input-output features using different projector architectures. All projectors considered will gradually decorrelate the input-output features. Although this decorrelation can be discarding irrelevant information, it can also degrade the efficacy of distillation.

Data-Efficient Training of Transformer Models

Network	acc@1	Teacher	#params
RegNetY-160	82.6	none	84M
BiT-M R152x2	84.5	none	236M
DeiT-Ti	72.2	none	5M
CivT-Ti	74.9	ensemble	6M
DeiT-Ti⚗	74.5	regnety-160	6M
DearKD	74.8	regnety-160	6M
USKD	75.0	regnety-160	6M
Our Method	**77.2**	regnety-160	6M
ResNet-50	76.5	none	25M
FunMatch	80.3	bit-m r152x2	25M
DeiT-S	79.8	none	22M
CivT-S	82.0	ensemble	22M
DeiT-S⚗	81.2	regnety-160	22M
DearKD	81.5	regnety-160	22M
USKD	80.8	regnety-160	22M
Our Method	**82.1**	regnety-160	22M

Table 1. Data-efficient training of transformers and CNNs on the ImageNet-1K dataset. Unless specified, all student models are trained for 300 epochs.

We observe a significant improvement over both DeiT and CivT when the capacity gap is large. We also achieve competitive performance to other distillation methods that are trained for 1000 epochs - **over three times longer!**.

Multiple factors, such as the soft maximum function and the batch normalisation, will be contributing to this observed result. However, the explanation is more concisely described by the fact that our distillation loss transfers more translational equivariance to the student.

Image Classification

	Teacher	Student	AT	KD	CC	CRD	ReviewKD	Ours
acc@1	26.69	30.25	29.30	29.34	30.04	28.62	28.39	**28.13**
acc@5	8.58	10.93	10.00	10.12	10.83	9.51	9.42	**9.29**

Table 2. Top-1 and Top-5 error rates (%) on ImageNet. ResNet18 as student, ResNet34 as teacher.

Teacher Student	ResNet32×4 ResNet8×4	ResNet50 MobileNet-V2
Teacher	79.42	79.34
Student	72.50	64.60
KD	73.33	64.60
CRD	75.51	69.11
ReviewKD	75.63	69.89
DKD	76.32	70.35
Ours	76.55	71.53
Ours + DKD	**76.95**	**71.75**

Table 3. Our simple feature distillation pipeline is complimentary to other logit distillation approaches, resulting in further performance improvements.

We perform experiments on the two standard knowledge distillation benchmarks: CIFAR100 and ImageNet1K. The results show improved performance, while being significantly cheaper and easier to adopt in practice.

Poster Session

r.miles18@imperial.ac.uk