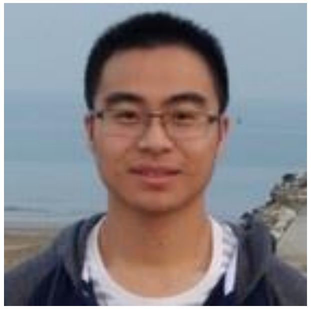
Dataset Condensation with Differentiable Siamese Augmentation

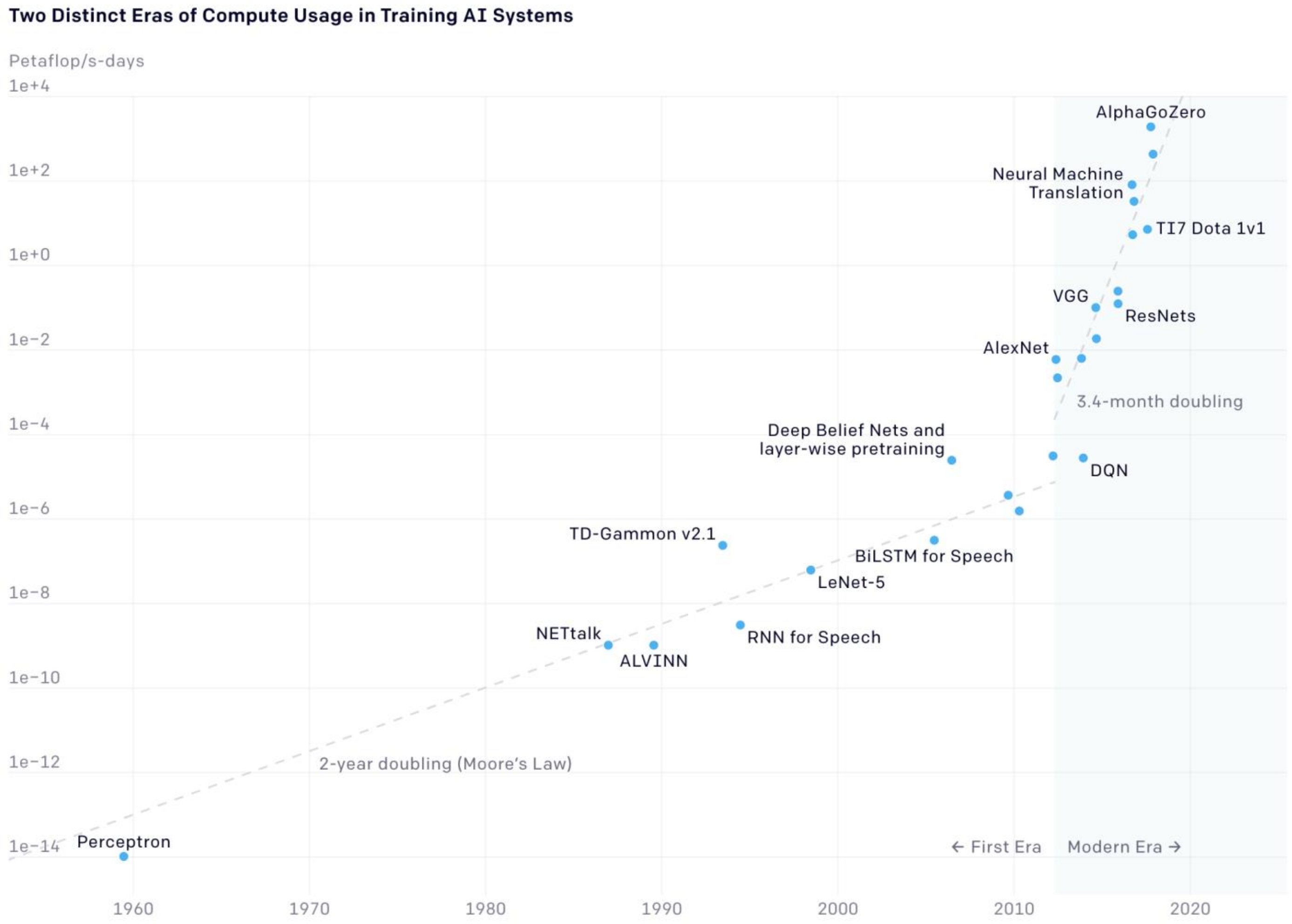
ICML 2021


Bo Zhao


Hakan Bilen

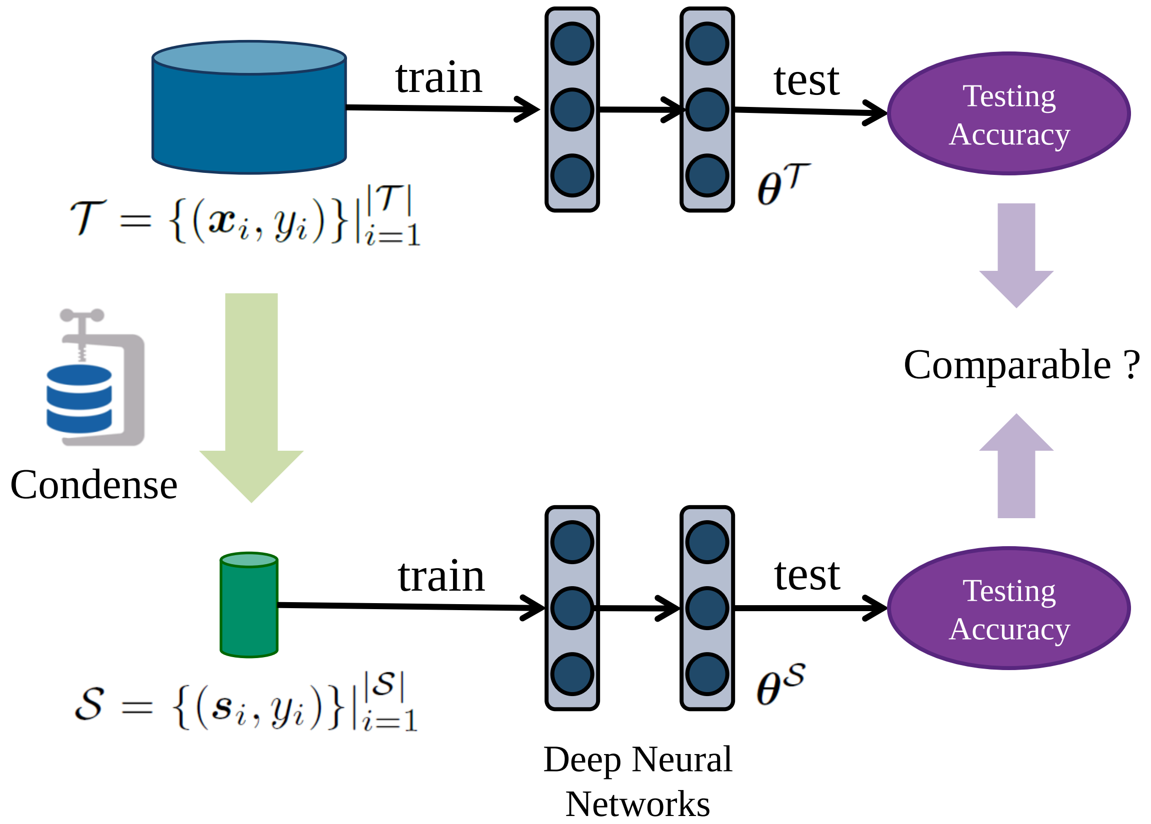
Correspondence: bo.zhao@ed.ac.uk

Training Deep Models is Increasingly More Expensive



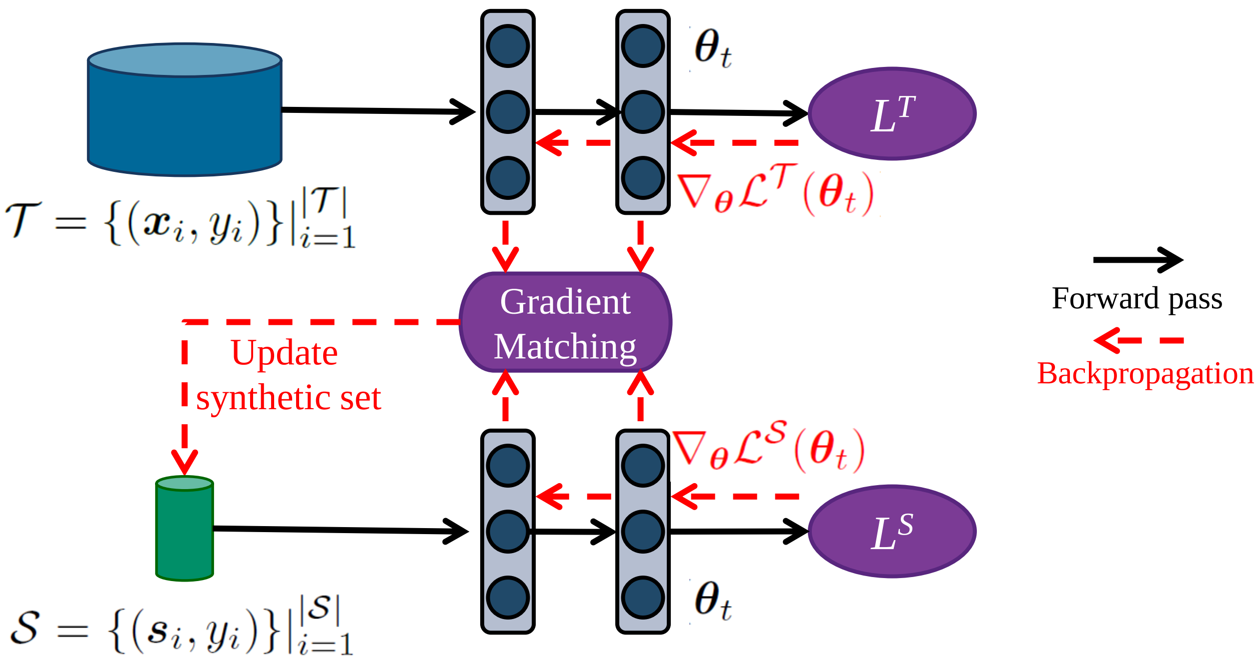
Source: "AI and Compute", OpenAI

Goal – Condensing Training Data



Our goal is to condense a large training set T into a small synthetic set S such that the model trained on the small synthetic set can obtain comparable testing performance to that trained on the large training set.

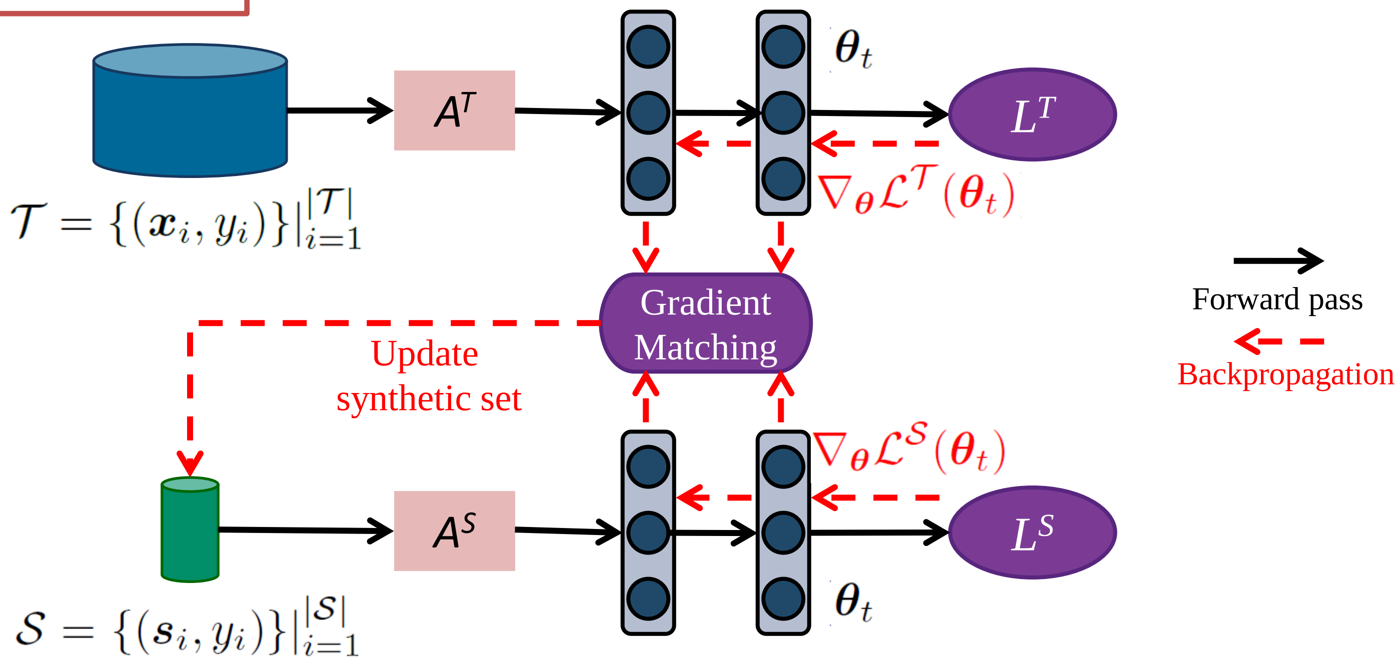
Related Work – Dataset Condensation with Gradient Matching[1]



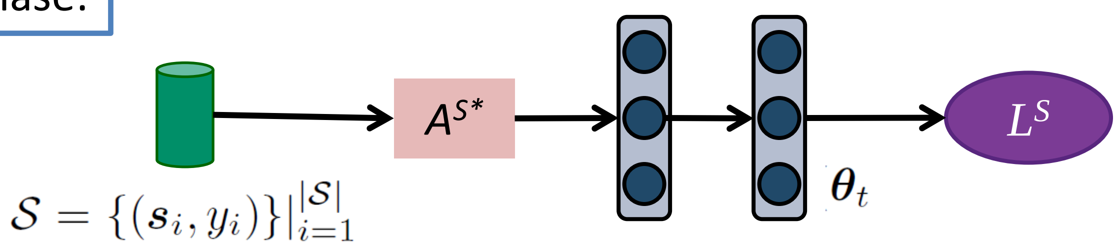
[1] Dataset Condensation with Gradient Matching. Zhao et. al. ICLR 2021.

Problem: Dataset Condensation with Data Augmentation

Condensation Phase:



Testing Phase:



A^T, A^S are traditional random augmentation.

Problem: Dataset Condensation with Data Augmentation

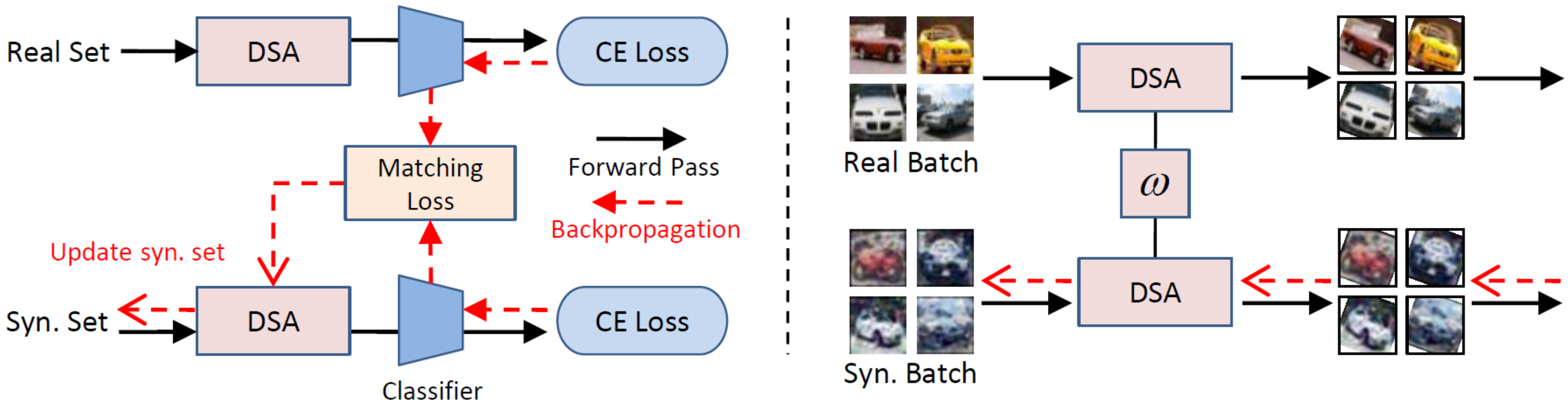
Train Synthetic Images		Train Models	Test Performance
Real	Synthetic	Synthetic	
-	-	-	45.5 ± 0.6
-	-	A^{S*}	46.9 ± 0.6
A^T	-	A^{S*}	42.8 ± 0.7
-	A^S	A^{S*}	44.6 ± 0.7
A^T	A^S	A^{S*}	44.5 ± 0.5

Problem:
- Naive augmentation schemes lead to either performance drops or negligible gains.

Reason:
- The learned synthetic images have different characteristics from natural images.
- Simply applying random augmentation to real/synthetic images leads to averaged effects which are difficult to disentangle.

Dataset Condensation with Differentiable Siamese Augmentation

Motivation: We aim to learn a synthetic training set that can be effectively used with data augmentation to train deep neural networks.

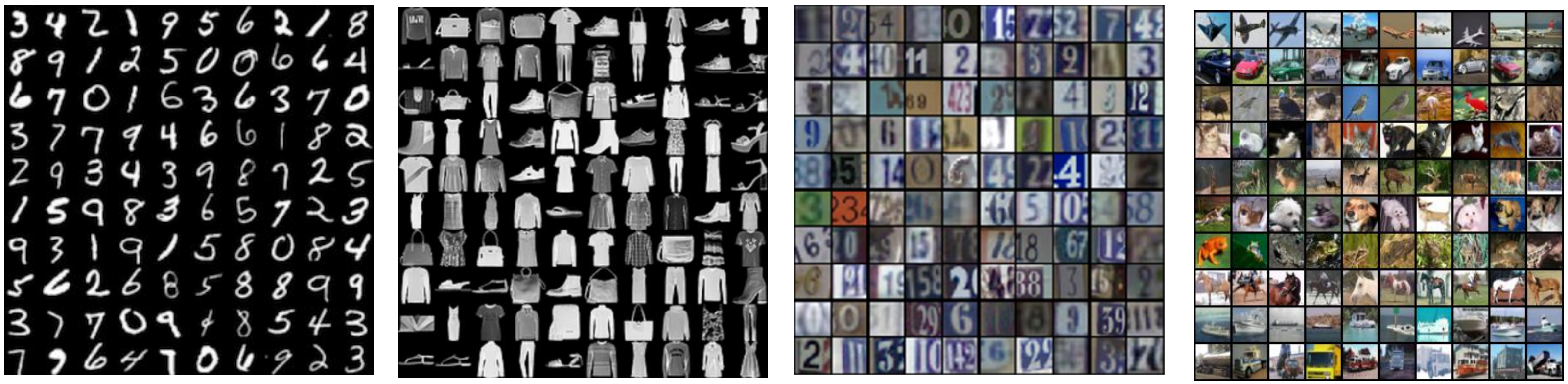


Benefits:

- Exploit more information from real/synthetic images.
- Sharing transformation enables learning prior knowledge (e.g. objects are usually horizontally on the ground).

Experiments – Datasets & Settings

Datasets:



Experimental Setting:

Stage 1: learn the condensed images (denoted as C)

Stage 2: train networks from scratch on the condensed images, then evaluate them on real testing data (denoted as T)

We test our method with MLP, ConvNet (default), LeNet, AlexNet,VGG-11 and ResNet-18.

We investigate different settings: 1, 10 and 50 image/class learning.

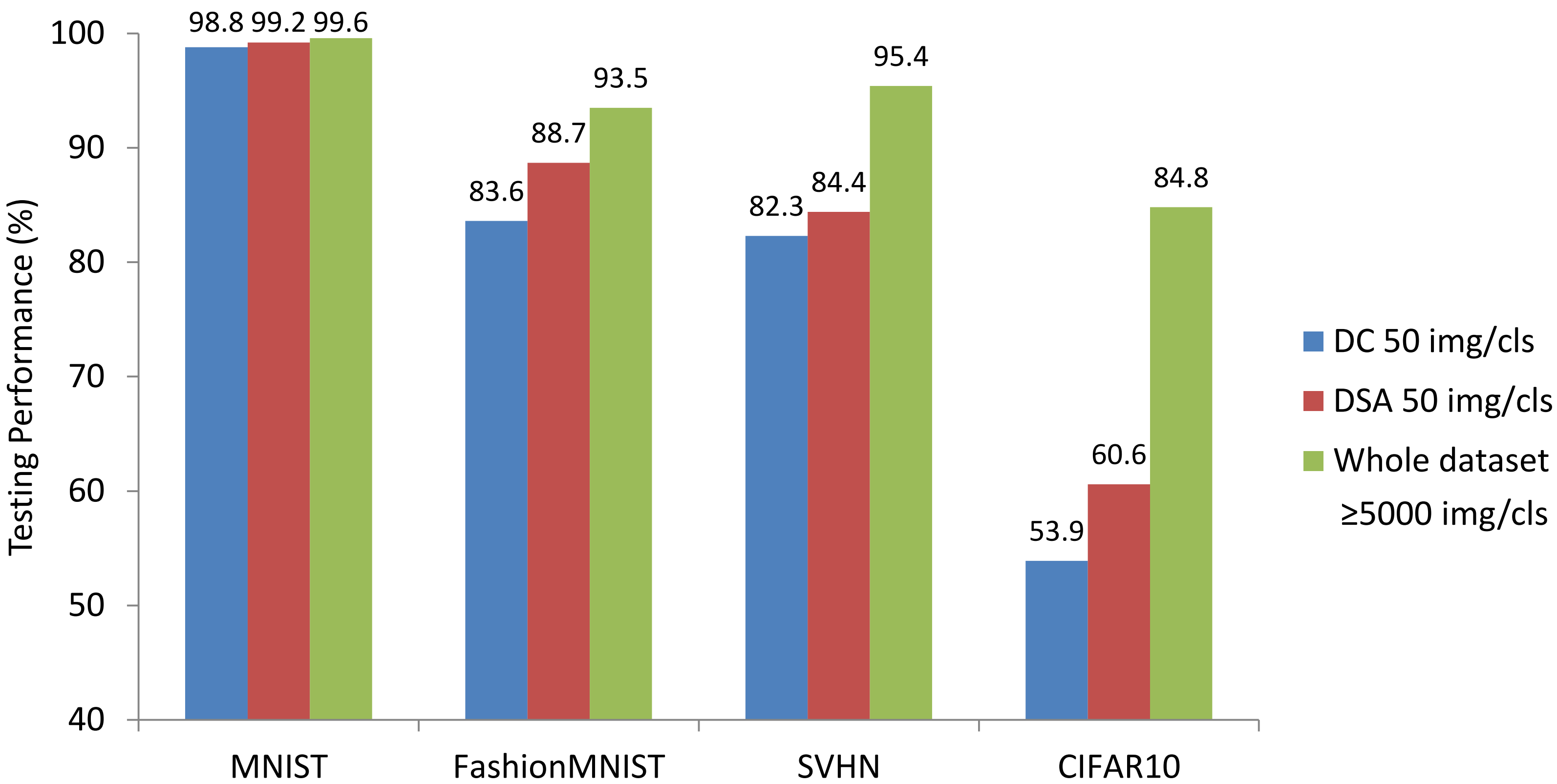
Experiments – Effectiveness of DSA

We study the effect of design choices in the proposed DSA in terms of test performance on CIFAR10 for 10 images/class learning.

Train Synthetic Images		Train Models	Test Performance
Real	Synthetic	Synthetic	
-	-	-	45.5 $\pm$ 0.6
-	-	A^{S*}	46.9 $\pm$ 0.6
A^T	-	A^{S*}	42.8 $\pm$ 0.7
-	A^S	A^{S*}	44.6 $\pm$ 0.7
A^T	A^S	A^{S*}	44.5 $\pm$ 0.5
A^ω (Shared)	A^ω (Shared)	A^{S*}	49.1 $\pm$ 0.6

- Our DSA learns better synthetic images.
- Naive augmentation schemes lead to either performance drops or negligible gains.

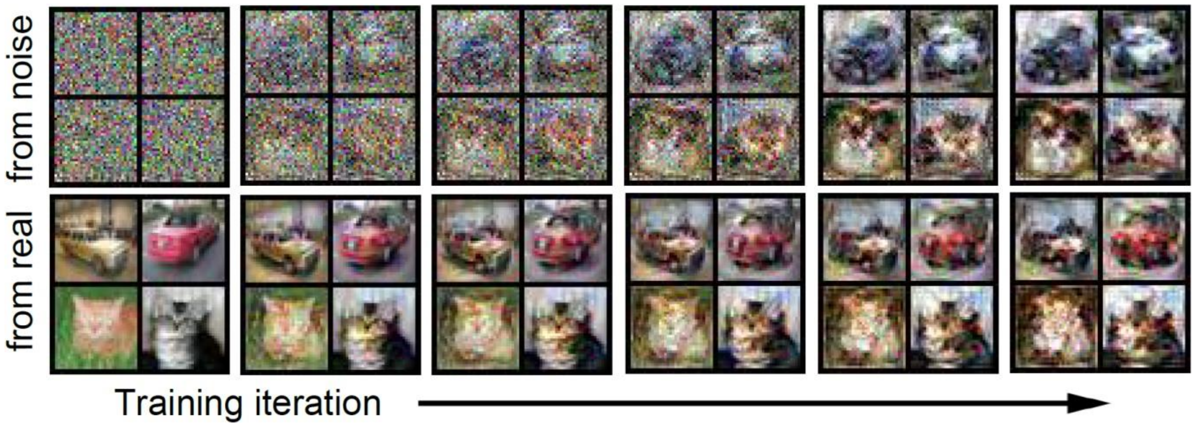
Experiments – Comparison to SOTA



- Outperform the state-of-the-art by a large margin (7% on CIFAR10).
- Obtain 99.2% testing accuracy on MNIST dataset with 50 synthetic images/class.

[1] Dataset Condensation with Gradient Matching. Zhao et al. ICLR 2021. (DC)

Experiments – Visualization



- Our method works well with both two kinds of initialization.
- The synthetic images inherit some contents from the initialization.

Conclusion

Conclusion:

- enable learning synthetic training set that can be effectively used with data augmentation.
- achieve better performance (~7% improvement on CIFAR10/100) than SOTA.
- show promising results in continual learning and neural architecture search.

Future work:

- explore the use of condensed images in challenging datasets like ImageNet.

Project page : https://github.com/VICO-UoE/DatasetCondensation
Correspondence: bo.zhao@ed.ac.uk

Project page

Thank you for listening!

Project page : https://github.com/VICO-UoE/DatasetCondensation
Correspondence: bo.zhao@ed.ac.uk

Project page