Discovering and Overcoming Limitations of Noise-engineered Data-free Knowledge Distillation

- Piyush Raikwar[1], Deepak Mishra[2]

[1]ABV-IIITM, Gwalior, India [2]IIT Jodhpur, India

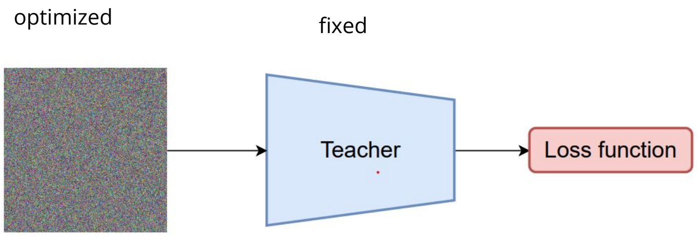
Data-free knowledge distillation

Traditionally, we assume the availability of original data.

Data-free distillation; we do not have the original data.

Prior works

- Optimize input.



- Train a generative model.

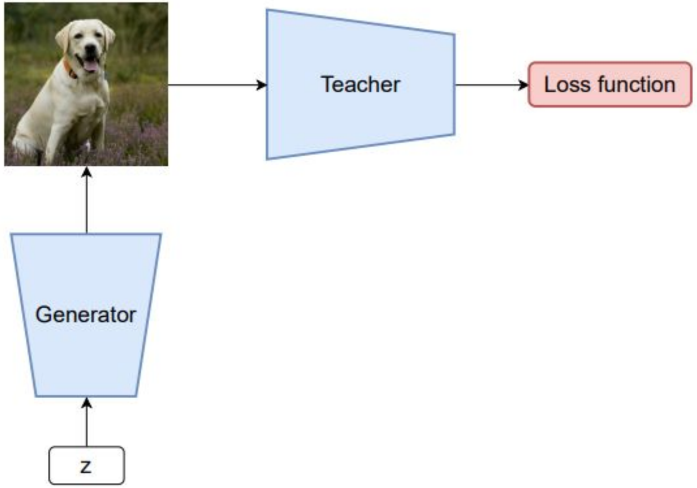


Figure 1: Prior works

The most straightforward alternative

Random noise

- Easy to generate, almost no computational burden.
- Previous attempts at distillation using Gaussian noise.
- Should not work directly, as it is basically some gibberish to teacher.

But, technically..

- Different input distribution.
- Covariate shift in hidden layer activations.

How to make it work?

Use current statistics instead of running statistics in teacher.

$$\overline{h_i} = \frac{h_i - \mu_r}{\sqrt{\sigma_r^2 + \epsilon}} \longrightarrow \overline{h_i} = \frac{h_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

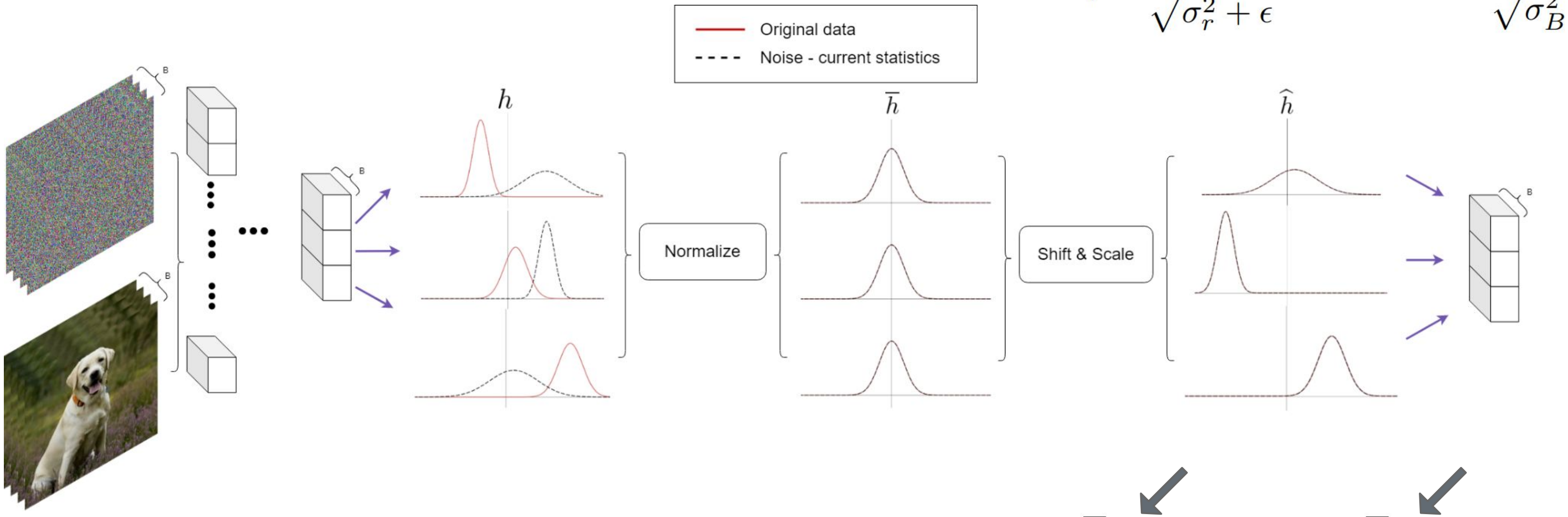


Figure 2: Inside BatchNorm layer

$$P(\overline{\boldsymbol{h}}|G, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B) = P(\overline{\boldsymbol{h}}|B, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B)$$
$$\approx \mathcal{N}(0,1)$$

A toy example

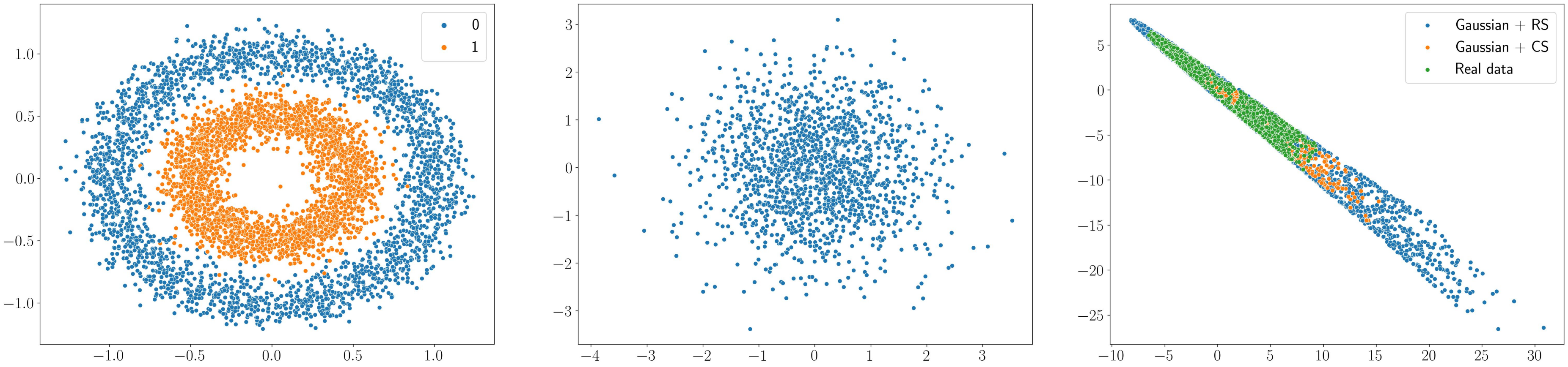


Figure 3: (Left) Circles data on which the MLP is trained. (Middle) Gaussian noise used as input to the trained MLP. (Right) Scatter plot for embeddings in different cases.

Student’s perspective

Make student accustomed to original data.

1. Use current statistics in student while evaluation.
2. Adjust student’s running statistics by just feed forwarding some original data.

Algorithm 1 Training - KD

Requires: pretrained teacher $T(.)$
Initialize: student $S(.;\theta)$ with parameters θ
for B **in** $1, 2, ..., \boldsymbol{\mathcal{B}_1}$ **do**
  $G \sim \mathcal{N}(0,1)$
  $y_T \leftarrow T(G|\mu_B, \sigma_B)$
  $y_S \leftarrow S(G|\theta, \mu_B, \sigma_B)$
  $\theta \leftarrow \theta - \eta \frac{\partial L_{KD}}{\partial \theta}$
end for

Algorithm 2 Evaluation

Requires: pretrained student $S(.;\theta)$
for B **in** $1, 2, ..., \boldsymbol{\mathcal{B}_2}$ **do**
  $X \sim D$
  $y_S \leftarrow S(X|\theta, \mu_B, \sigma_B)$
  $y_{label} \leftarrow argmax(y_S)$
end for

Experiments

ResNet-34 teacher (93.29%) trained on CIFAR10.

Student	ResNet34	ResNet18	MobileNetV2
Supervised	93.29	93.22	91.61
Original data + RS (Oracle)	92.74 ± 0.21	92.44 ± 0.05	90.57 ± 0.22
Original data + CS	92.77 ± 0.22	92.20 ± 0.1	91.44 ± 0.13
Gaussian noise + RS	13.18 ± 0.21	13.49 ± 0.08	12.43 ± 0.3
Gaussian noise + CS (Ours)	87.11 ± 0.23	85.98 ± 0.12	82.47 ± 0.26

Table 1: CIFAR10 results

Dataset	SVHN	CIFAR100	Food101
Teacher	ResNet18	WideResNet-28-10	ResNet101
Student	MobileNetV2	WideResNet-16-8	ResNet18
Teacher supervised	94.48	80.6	73.4
Original data + RS (Oracle)	95.75	74.1	67.6
Gaussian noise + RS	45.03	1.2	0.9
Gaussian noise + CS (Ours)	92.93	65.7	54.16

Table 2: Results on other datasets

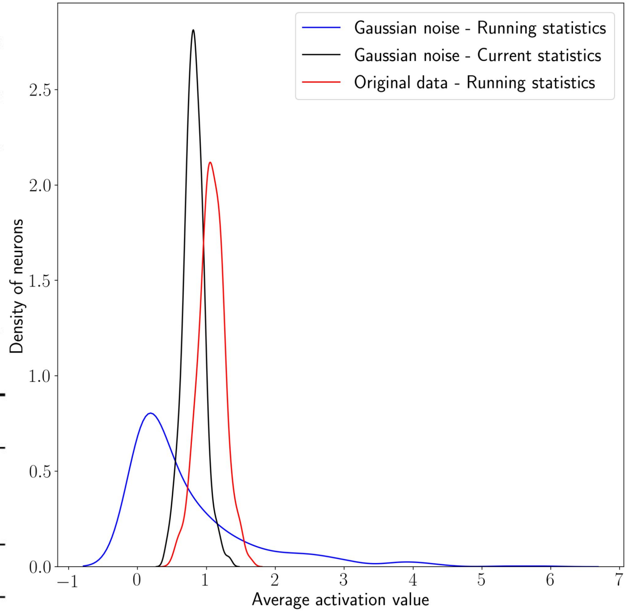


Figure 4: ResNet-34 CIFAR10 'avgpool' activation distribution

Other observations

1. Larger the batch size during training the better.
2. Larger the batch size during inference the better.

But, they have to be just enough, e.g., 256 batch size is sufficient.

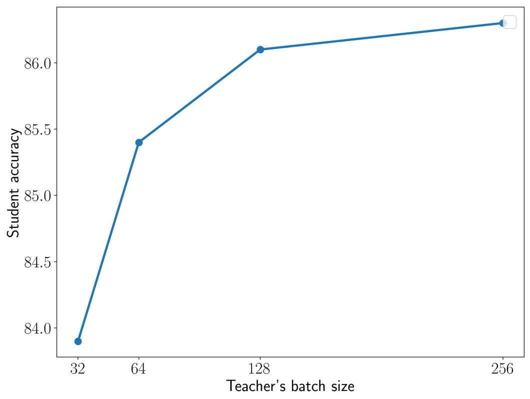


Figure 5: Batch size during training

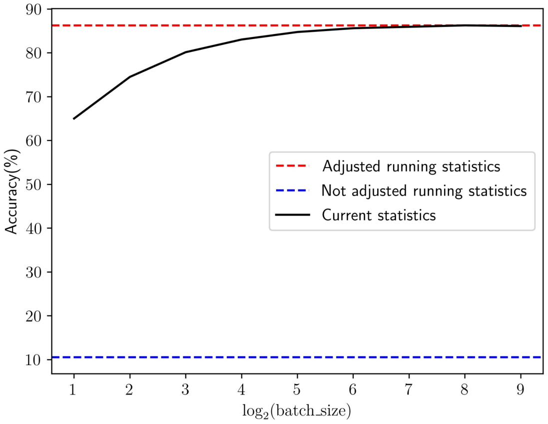


Figure 6: Handling student

3. Handling partial BN layers helps partially.

$\mathcal{P}$	Student accuracy
100 (Running statistics)	13.49
90	18.26
75	57.73
50	79.54
25	82.24
0 (Current statistics)	89.4

Table 3: Percent BN layers using running statistics

4. More the data for adjusting the running statistics of student the better.

Conclusion

- We show how covariate shift interferes with data-free distillation.
- We propose an approach to mitigate it to a significant extent and show that KD is possible using just Gaussian noise.
- We might not necessarily need realistic data, at least for KD. Thus we lay the foundations for noise-engineered data-free distillation.

Future work

- Noise of lower resolutions.
- Various other noises, such as fractals.
- Applying the proposed method to other domains like transfer learning and domain adaptation.
- Use proposed method to complement other data-free distillation approaches.

Thank you for the attention!

Paper

Code

Contact at: piyush.raikwar@cern.ch