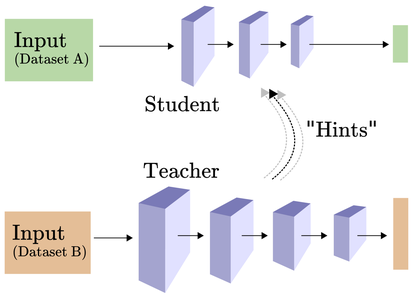
Knowledge Transfer with Jacobian Matching

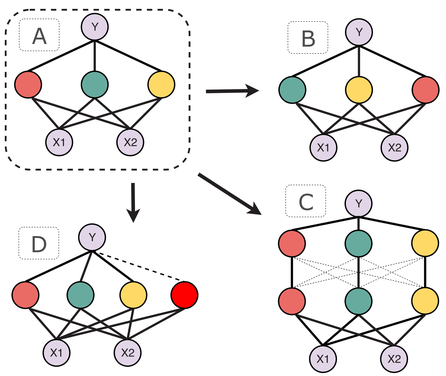
Suraj Srinivas & François Fleuret
{suraj.srinivas, francois.fleuret}@idiap.ch

Introduction



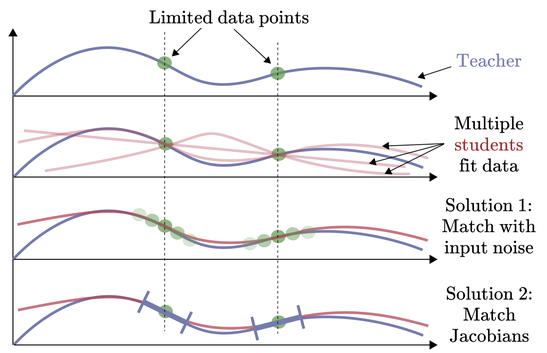
- **Problem:** Improve performance of a student network by using 'hints' from a pre-trained teacher
- **If** dataset A == dataset B :
  task = distillation
  else :
  task = transfer learning

What 'hints' to use?



- 'Hints' must be invariant to parameterization
- **Previous methods** use neural net outputs y
- **Our method** uses $\nabla_x y = \left[\frac{\partial y}{\partial x_0} \quad \frac{\partial y}{\partial x_1} \quad \cdots \quad \frac{\partial y}{\partial x_D}\right]$
- Previous works [1, 2] also considered $\nabla_x y$, but did not motivate the choice of loss function or relation to previous approaches.

Jacobian Matching



Jacobian Matching is equivalent to **matching output activations with noise** added to the inputs.

Form of Loss function

$\mathcal{T}(x) \rightarrow$ Teacher network, $\mathcal{S}(x) \rightarrow$ Student network
$\boldsymbol{\xi} \rightarrow$ Gaussian noise with covariance $\sigma^2 \mathrm{I}$

$$\mathbb{E}_{\boldsymbol{\xi}}\left[(\mathcal{T}(\mathbf{x}+\boldsymbol{\xi}) - \mathcal{S}(\mathbf{x}+\boldsymbol{\xi}))^2\right] \sim (\mathcal{T}(\mathbf{x}) - \mathcal{S}(\mathbf{x}))^2 + \sigma^2 \|\nabla_x \mathcal{T}(\mathbf{x}) - \nabla_x \mathcal{S}(\mathbf{x})\|_2^2$$

$$\mathbb{E}_{\boldsymbol{\xi}}\left[-\mathcal{T}_s(\mathbf{x}+\boldsymbol{\xi}) \log\left(\mathcal{S}_s(\mathbf{x}+\boldsymbol{\xi})\right)\right] \sim -\mathcal{T}_s(\mathbf{x}) \log(\mathcal{S}_s(\mathbf{x})) - \sigma^2 \frac{\nabla_x \mathcal{T}_s(\mathbf{x})^T \nabla_x \mathcal{S}_s(\mathbf{x})}{\mathcal{S}_s(\mathbf{x})}$$

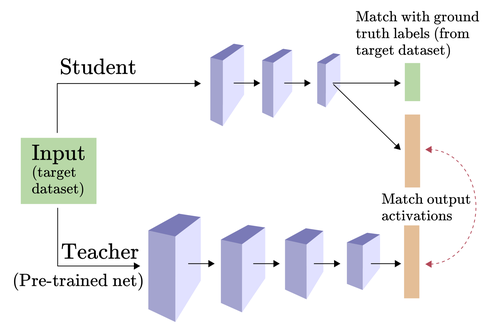
Jacobian Norm Regularization

$$\mathbb{E}_{\boldsymbol{\xi}}\left[(y(\mathbf{x}) - \mathcal{S}(\mathbf{x}+\boldsymbol{\xi}))^2\right] \sim (y(\mathbf{x}) - \mathcal{S}(\mathbf{x}))^2 + \sigma^2 \|\nabla_x \mathcal{S}(\mathbf{x})\|_2^2$$

- We encourage the neural net output to be insensitive to small changes to the input.
- For linear models, Jacobian = model weights

Transfer Learning

- We use a distillation-like method called *Learning without Forgetting* (LwF) [3]
- This enables us to transfer Imagenet trained representations without using the Imagenet dataset



Theoretical Results

- When does LwF work? If the teacher network produces meaningless outputs, LwF would not help.

Q: When does the teacher produce useful outputs?
A: Three conditions must be satisfied :
- the **teacher's loss** on the source dataset (Imagenet) is **small**
- the **teacher's Lipschitz** constant is **small**
- **"distance"** between source and target datasets is **small**

- Does improving distillation improve transfer learning?

Q: Does Jacobian matching improve over LwF?
A: Yes.

Crucial to the proof is the equivalence between Jacobian matching and matching outputs with input noise.

Distillation Results

- **Teacher** : VGG-9, obtains 64.78% accuracy
- **Student** : VGG-4
- **Dataset** : CIFAR-100

# of Data points per class	10	50	500 (full)
Regular Cross-Entropy (CE) training	20.03	37.6	54.28
Match Activations + CE	33.92	46.47	**56.65**
Match {Activations + Jacobians} + CE	**39.55**	**49.49**	54.57
Match Activations only	33.6	45.73	56.59
Match {Activations + Jacobians}	38.16	47.79	51.33

Transfer Learning Results

- **Teacher** : Imagenet pre-trained ResNet-32
- **Student** : VGG-9
- **Target dataset** : MIT Indoor Scenes (67 classes)

# of Data points per class →	25	50	Full
No Transfer Learning	35.19	46.38	59.33
Fine-tuning Oracle [a]	57.65	64.18	71.42
LwF [3]	45.08	55.22	65.22
LwF + Jacobians	45.26	56.49	66.04
LwF + attention [2]	46.01	57.80	**67.24**
LwF + attention + Jacobians	**47.31**	**58.35**	**67.31**

[a] Requires teacher and student to have the same architecture

References

[1] Czarnecki *et.al.* *Sobolev training for neural networks*. NIPS, 2017
[2] Sergey Zagoruyko and Nikos Komodakis. *Paying more attention to attention*. ICLR, 2017
[3] Zhizhong Li and Derek Hoiem. *Learning without forgetting*. PAMI, 2017

Acknowledgements

FNSNF