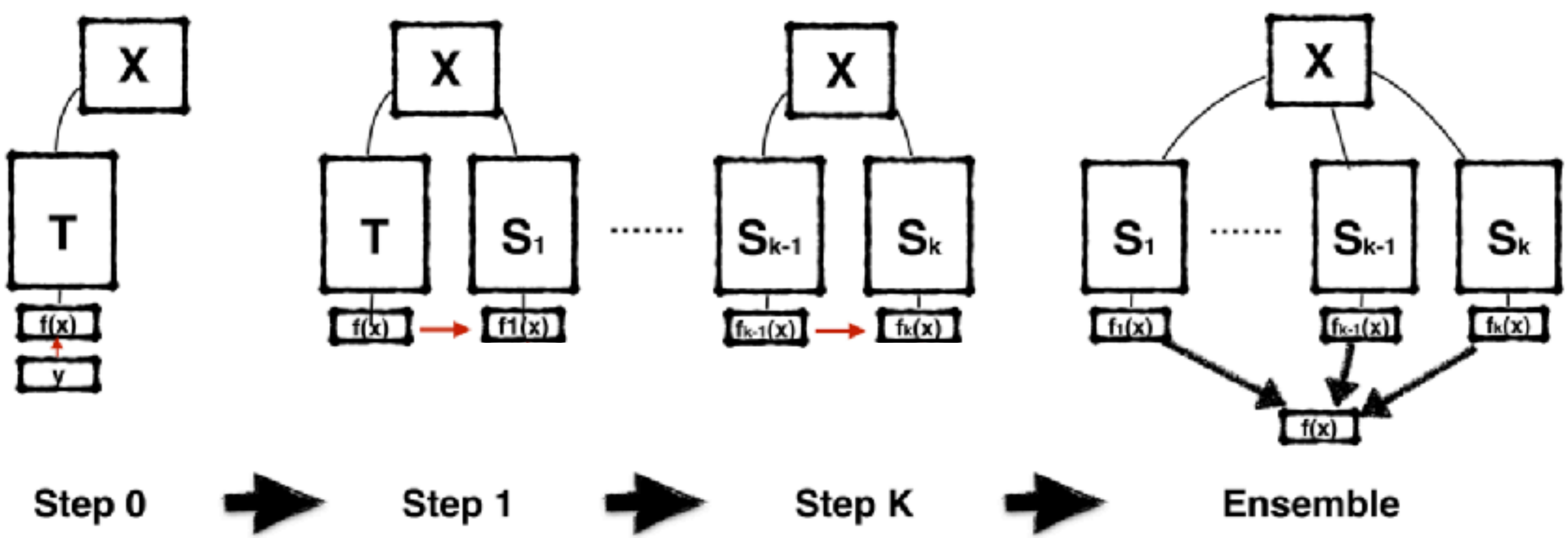
Born Again Neural Networks

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen,
Laurent Itti, and Anima Anankdumar

furlanel@usc.edu or for twitter trolling → @furlanel

Born Again Neural Networks

Knowledge Distillation between **identical** neural network architectures systematically **improves the student performance**



Born Again Neural Networks

Why Born Again ???

BORN AGAIN TREES

Leo Breiman
Statistics Department
University of California
Berkeley, CA 94720
leo@stat.berkeley.edu

Nong Shang
School of Public Health
University of California
Berkeley, CA 94720
shang@stat.berkeley.edu

ABSTRACT

Tree predictors such as CART or C4.5 are often not as accurate as neural nets or use of multiple trees. But these latter methods lead to predictors whose structure is difficult to understand, whereas trees have a universal simplicity. Because of this, it is appealing to try and find tree representations of more complex predictors. We study tree representers of multiple tree predictors. These representers are larger, more stable and more accurate than trees grown the usual way. For this reason, we call them "born again" trees.

Born Again Neural Networks

Why Born Again ???

BORN AGAIN TREES

Dark Knowledge Under the Light

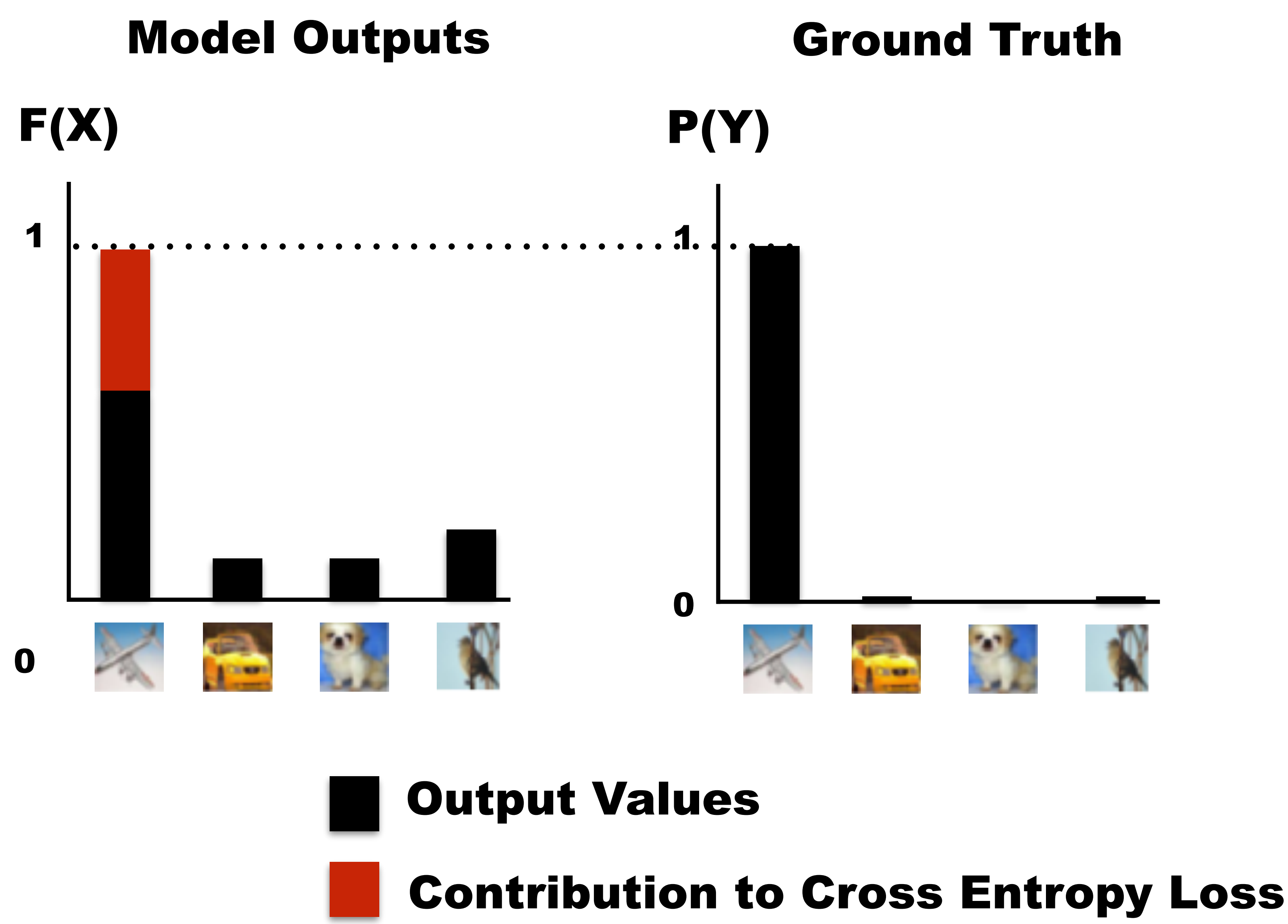
Knowledge Distillation general interpretation is that conveys some "**Dark knowledge"** hidden in the **output scores of the teacher** that reveals learned similarities between target categories

Dark Knowledge Under the Light

Ground Truth Baseline

Cross-Entropy Loss Function with one-hot Labels:

- Only the dimension corresponding to correct category contributes to the loss function.

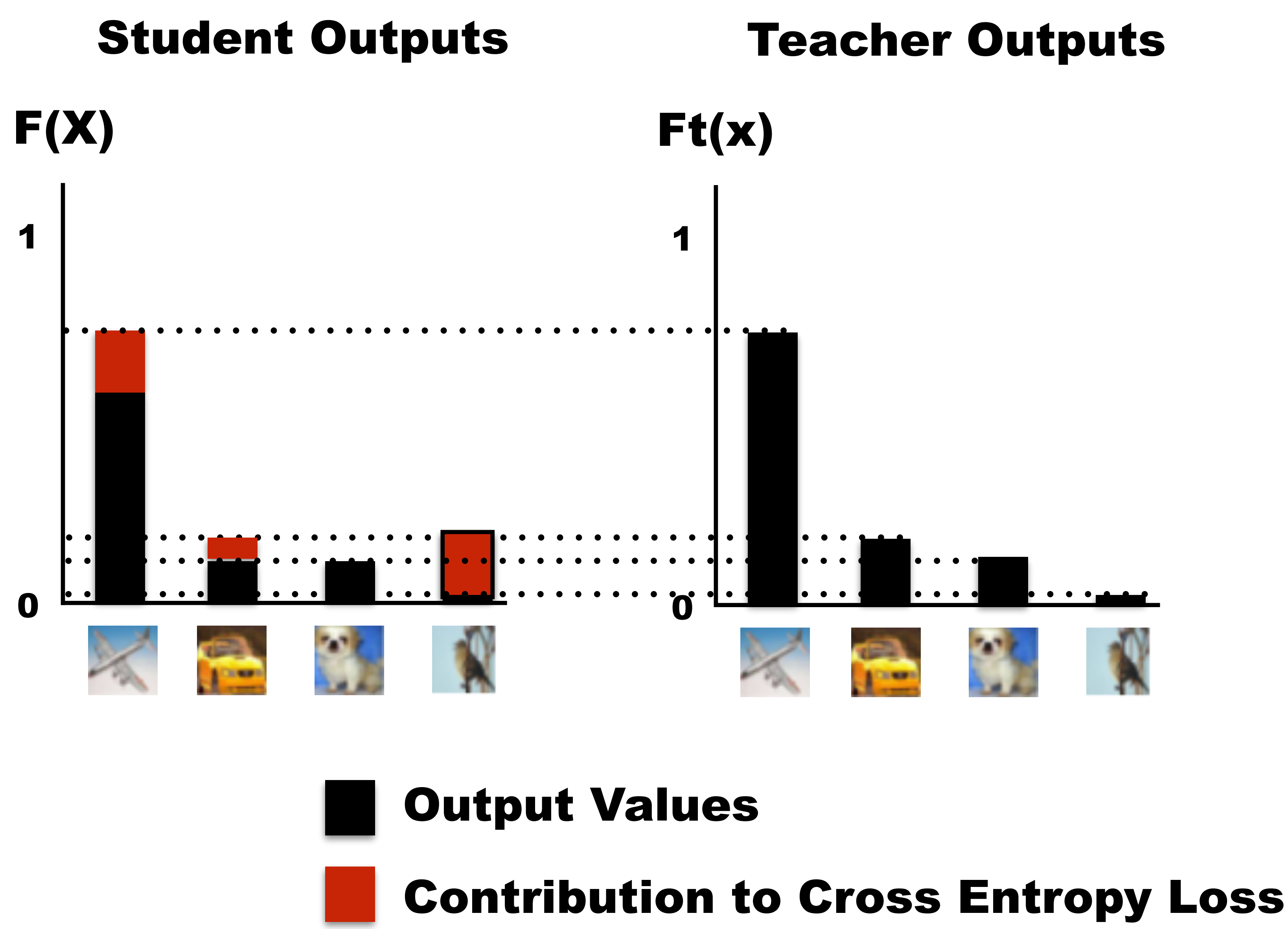


Dark Knowledge Under the Light

Knowledge Distillation

Cross-Entropy Loss Function with teacher outputs:

- The error in the output of all categories contributes to the loss function.
- If the teacher is highly accurate and certain it is virtually identical to using original labels.

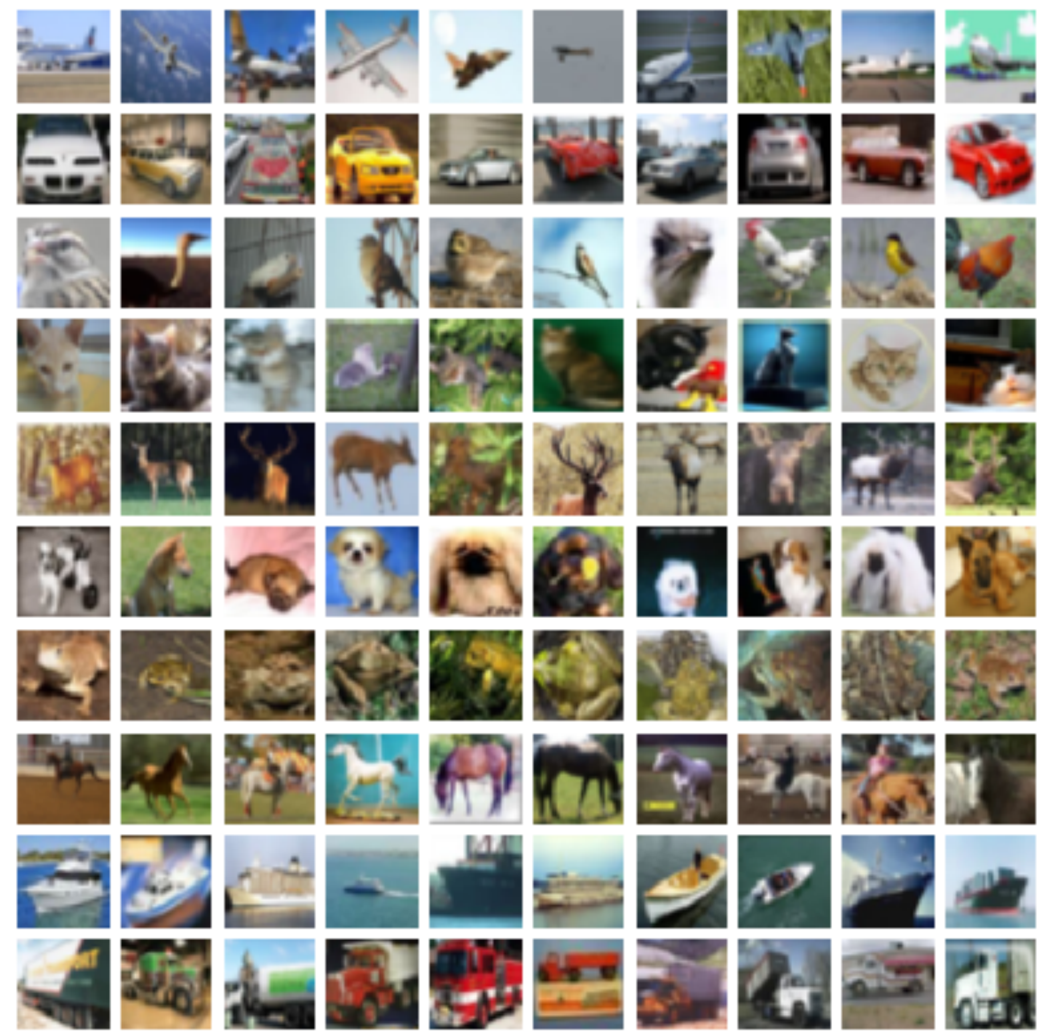


BAN - DenseNets

Cifar-100 Object Classification (100 Categories)

- **Students** have **systematically lower test error** than identical teacher.
- The **most complex** baseline model **DenseNet-80-120 with 50.4M params** reaches a **test error of 16.87**
- The **smallest BAN-DenseNet-112-33** with **6.3M** params after 3 generations reaches a **test error of 16.59,** lower than the most complex baseline.

Network	Teacher	BAN-1	BAN-2	BAN-3
DenseNet-112-33	18.25	17.61	17.22	**16.59**
DenseNet-90-60	17.69	16.62	**16.44**	16.72
DenseNet-80-80	17.16	16.26	16.30	**15.5**
DenseNet-80-120	16.87	**16.13**	16.13	/



BAN - DenseNets

Ban+L uses both labels and knowledge distillation

Inter-generational ensembles improve over the individual models

Network	Teacher	BAN+L	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNet-112-33	18.25	17.68	17.61	17.22	**16.59**	15.77	15.68
DenseNet-90-60	17.69	16.93	16.62	**16.44**	16.72	15.39	15.74
DenseNet-80-80	17.16	16.5	16.26	16.30	**15.5**	15.46	15.14
DenseNet-80-120	16.87	16.41	**16.13**	16.13	/	**15.13**	**14.9**

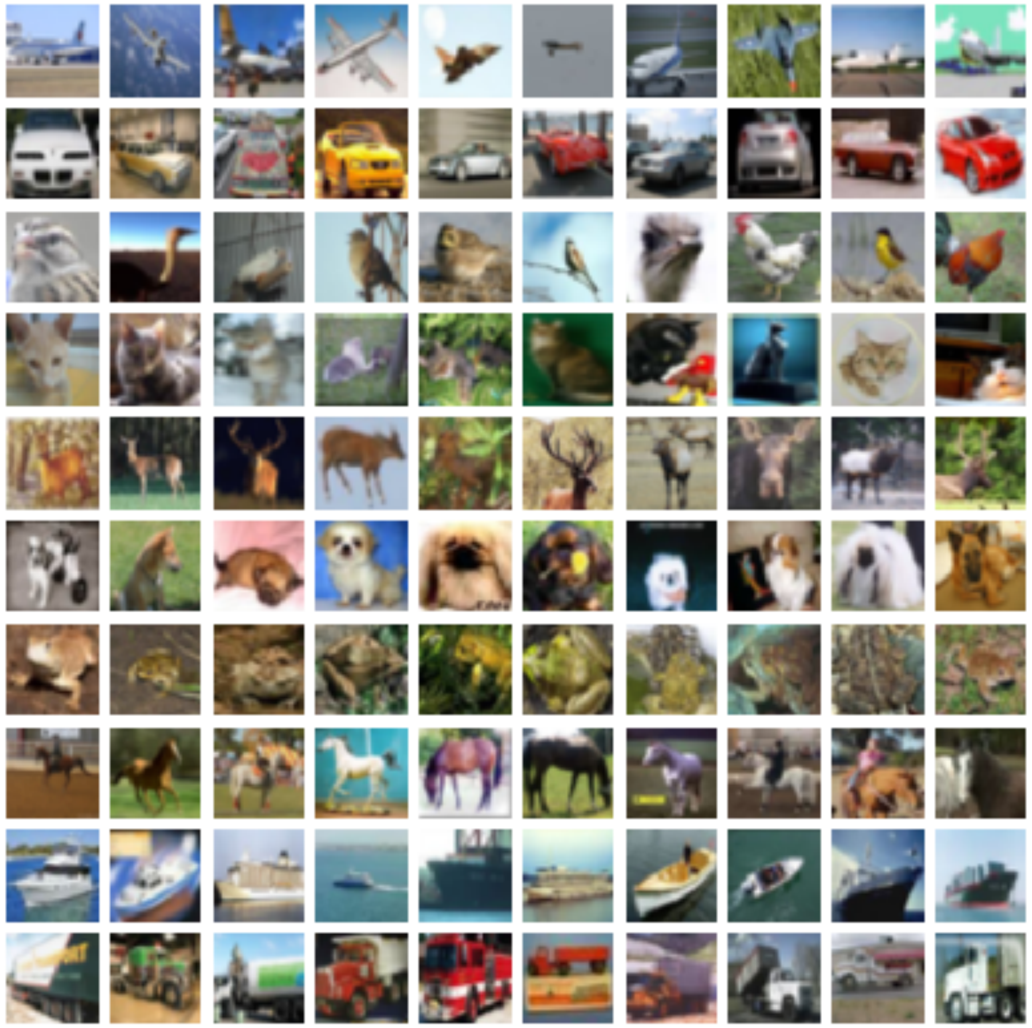
DenseNet-90-60 is used as teacher with students that share the same size of hidden states after each spatial transition but differs in depth and compression rate

Densenet-90-60	Teacher	0.5*Depth	2*Depth	3*Depth	4*Depth	0.5*Compr	0.75*Compr	1.5*compr
Error	17.69	16.95	16.43	16.64	16.64	19.83	17.3	18.89
Parameters	22.4 M	21.2 M	13.7 M	12.9 M	1 2.6 M	5.1 M	10.1 M	80.5 M

BAN -Cifar10

Cifar-10 Object Classification (10 Categories)

Network	Parameters	Teacher	BAN
Wide-ResNet-28-1	0.38 M	6.69	**6.64**
Wide-ResNet-28-2	1.48 M	5.06	**4.86**
Wide-ResNet-28-5	9.16 M	4.13	**4.03**
Wide-ResNet-28-10	36 M	**3.77**	3.86
DenseNet-112-33	6.3 M	3.84	**3.61**
DenseNet-90-60	16.1 M	3.81	**3.5**
DenseNet-80-80	22.4 M	**3.48**	3.49
DenseNet-80-120	50.4 M	**3.37**	3.54



Dark Knowledge Under the Light

Two experimental treatments to disentangle the contribution to the KD loss function of :

- Single dimension corresponding to teachers **predicted** categories
- Dimensions corresponding to the teachers **non predicted** category.

Dark Knowledge with Permuted Predictions

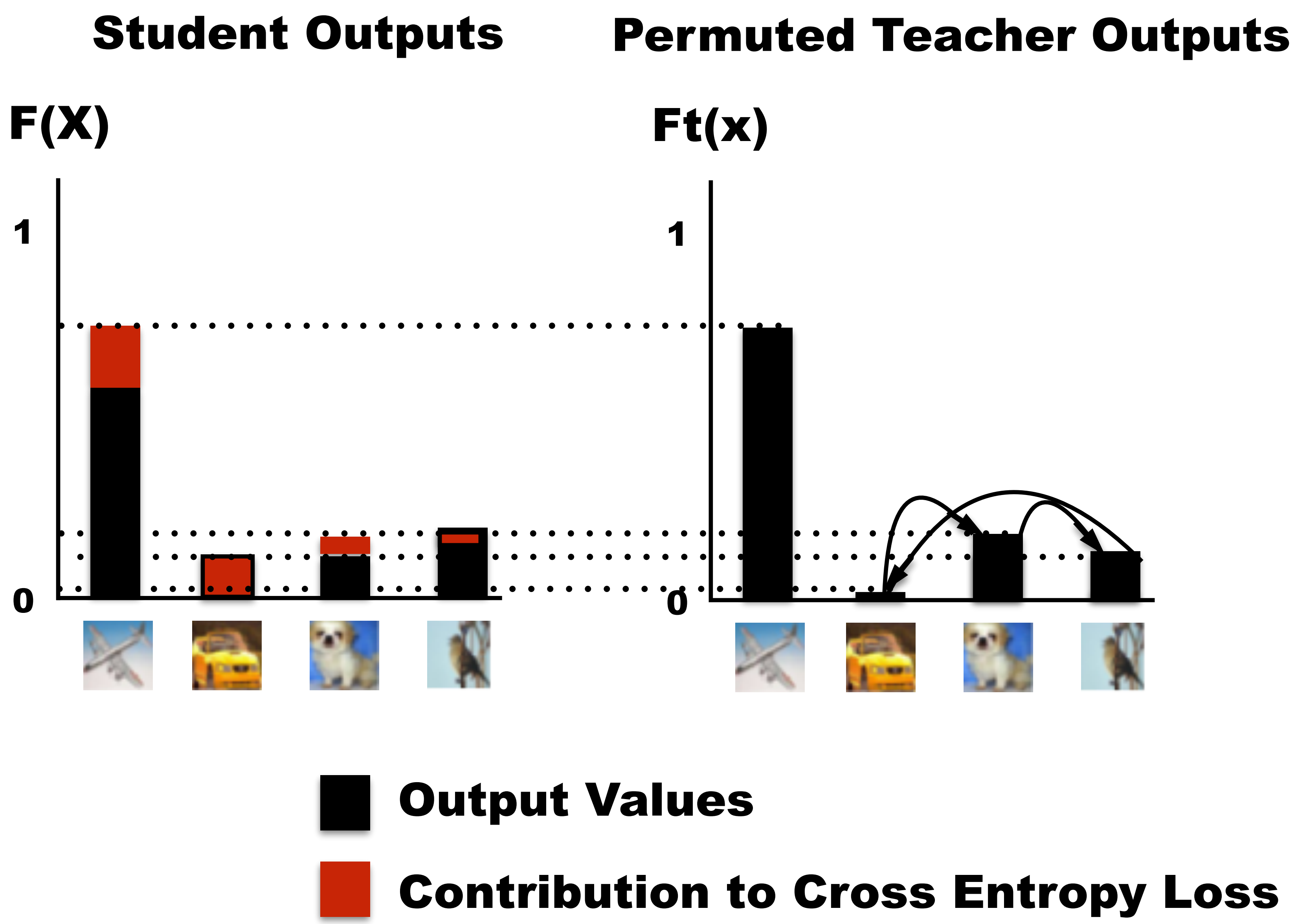
Confidence Weighted by Teacher Max

Dark Knowledge Under the Light

Dark Knowledge with Permuted Predictions

Cross-Entropy Loss Function with permuted teacher outputs for the non max categories:

- The error in the output of all categories contributes to the loss function.
- Non max categories information are **permuted**
- **Max dimension** contribution is isolated

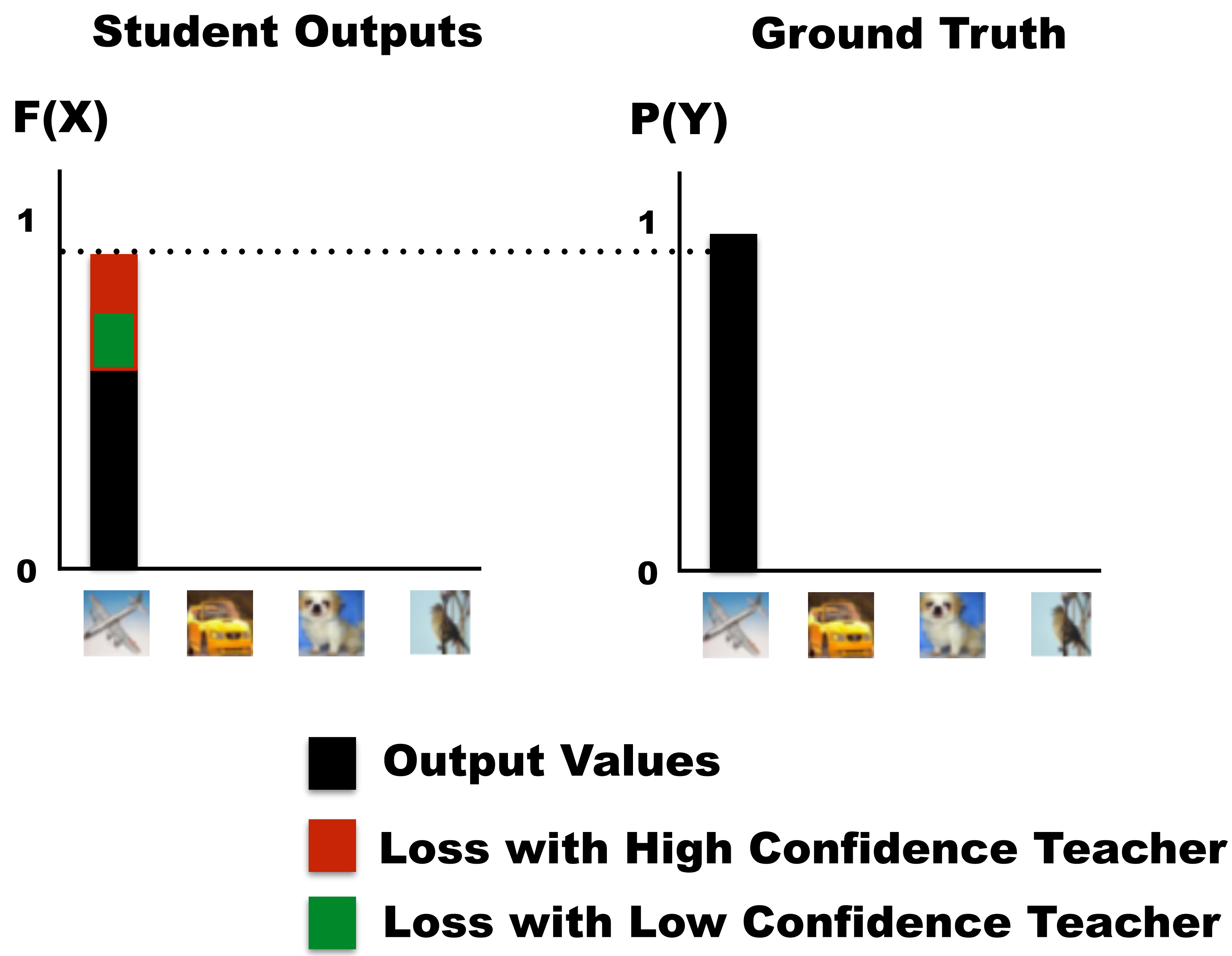


Dark Knowledge Under the Light

Confidence Weighted by Teacher Max

Cross-Entropy Loss Function with label, re-weighted by the value of the teacher max:

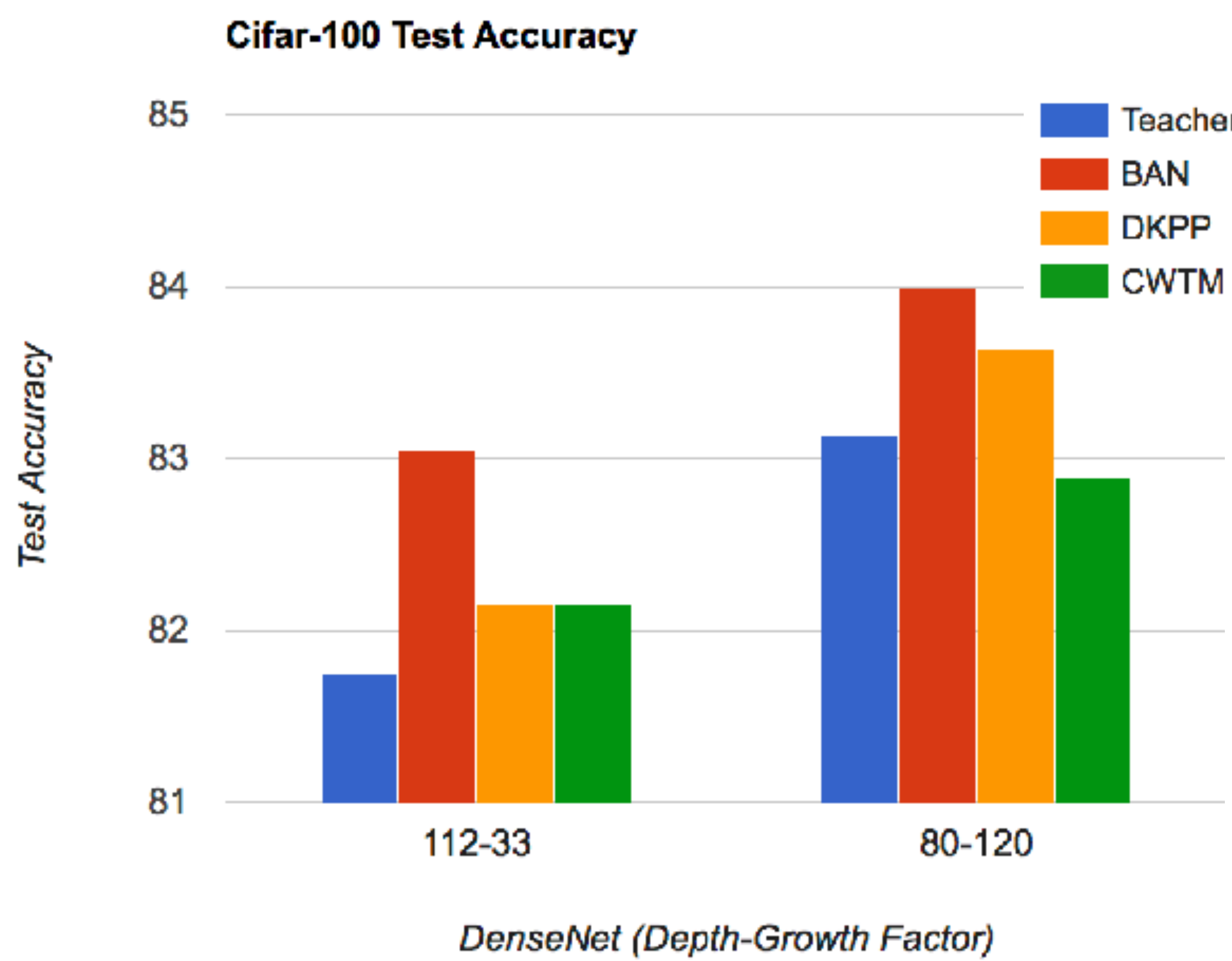
- **Only** the dimension corresponding to **correct category** contributes to the loss function.
- Loss function of **each sample** is **re-weighted by the teacher's max score.**
- Interpretation of knowledge distillation as **importance weighting of samples**, where **importance** is defined by the **teacher's confidence.**



Dark Knowledge Under the Light

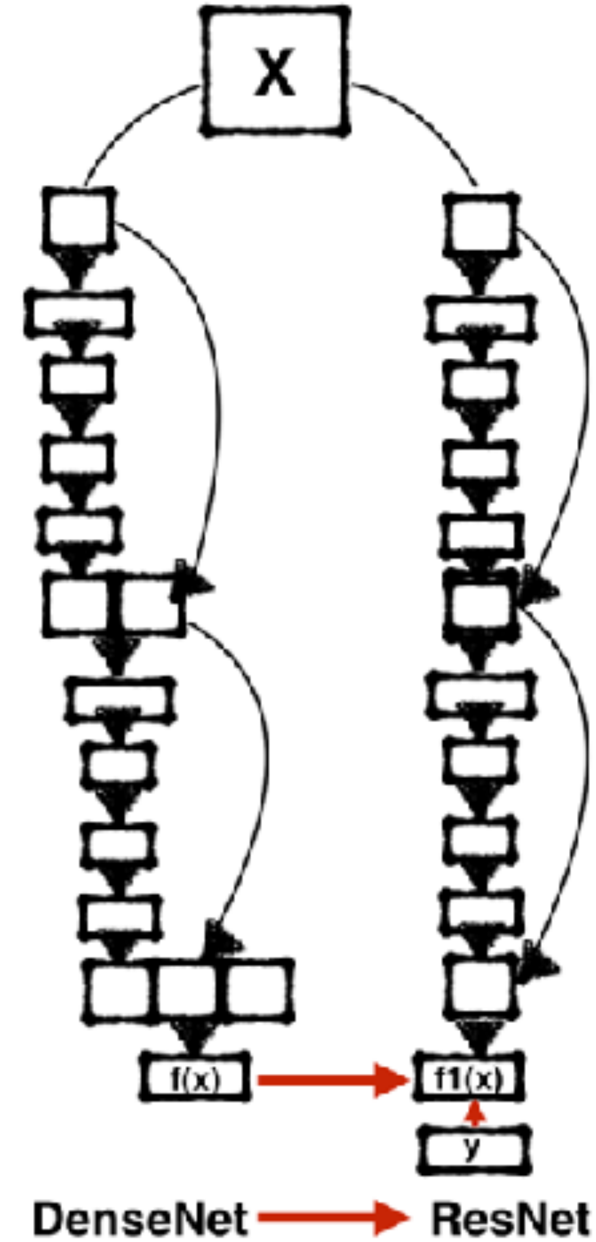
We observe that the contribution of Knowledge Distillation depends on both the correct and incorrect output categories:

- Best results on CIFAR-100 using **simple KD** with no labels.
- Permuting the incorrect output categories results in **systematic (but reduced) gains**.
- **CWTM** of samples gives **more unstable results** than **DKPP** suggesting that **higher-order information** of the complete output distribution **are important.**



BAN - ResNets

DenseNet 90-60	Parameters	Baseline	BAN
Pre-activation ResNet-1001	10.2 M	22.71	/
BAN-Pre-ResNet-14-0.5	7.3 M	20.28	18.8
BAN-Pre-ResNet-14-1	17.7 M	18.84	**17.39**
BAN-Wide-ResNet-1-1	20.9 M	20.4	19.12
BAN-Match-Wide-ResNet-2-1	43.1 M	18.83	17.42
BAN-Wide-ResNet-4-0.5	24.3 M	19.63	**17.13**
BAN-Wide-ResNet-4-1	87.3 M	18.77	17.18



BAN - LSTM

Penn Tree Bank val/test perplexities of BAN-LSTM language models

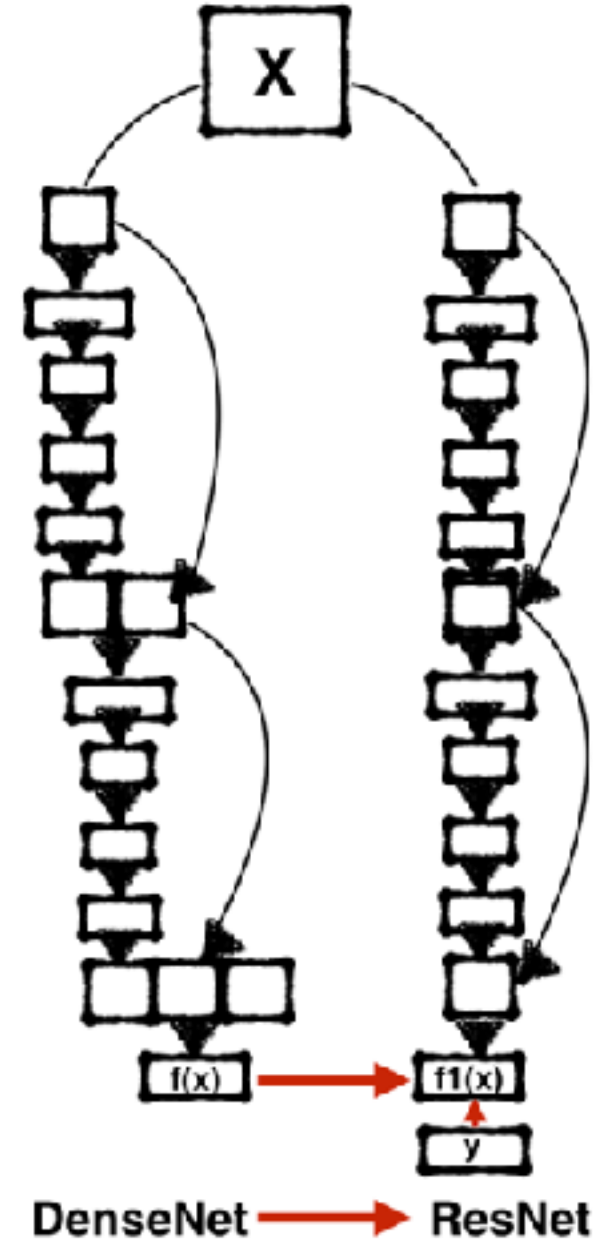
Network	Parameters	Teacher Val	BAN Val	Teacher Test	BAN Test
ConvLSTM	19M	83.69	80.27	80.05	**76.97**
LSTM	52M	75.11	71.19	71.87	**68.56**

BAN - ResNets

BAN Wide-ResNet with identical teacher

BAN Wide-ResNet Teacher Dense-90-60 Student (**17.69** baseline)

Network	Teacher	BAN	Dense-90-60
Wide-ResNet-28-1	30.05	29.43	24.93
Wide-ResNet-28-2	25.32	24.38	18.49
Wide-ResNet-28-5	20.88	20.93	17.52
Wide-ResNet-28-10	19.08	18.25	16.79



BAN - LSTM

Penn Tree Bank val/test perplexities of BAN-LSTM language models

Network	Parameters	Teacher Val	BAN Val	Teacher Test	BAN Test
ConvLSTM	19M	83.69	80.27	80.05	**76.97**
LSTM	52M	75.11	71.19	71.87	**68.56**

Related Literature

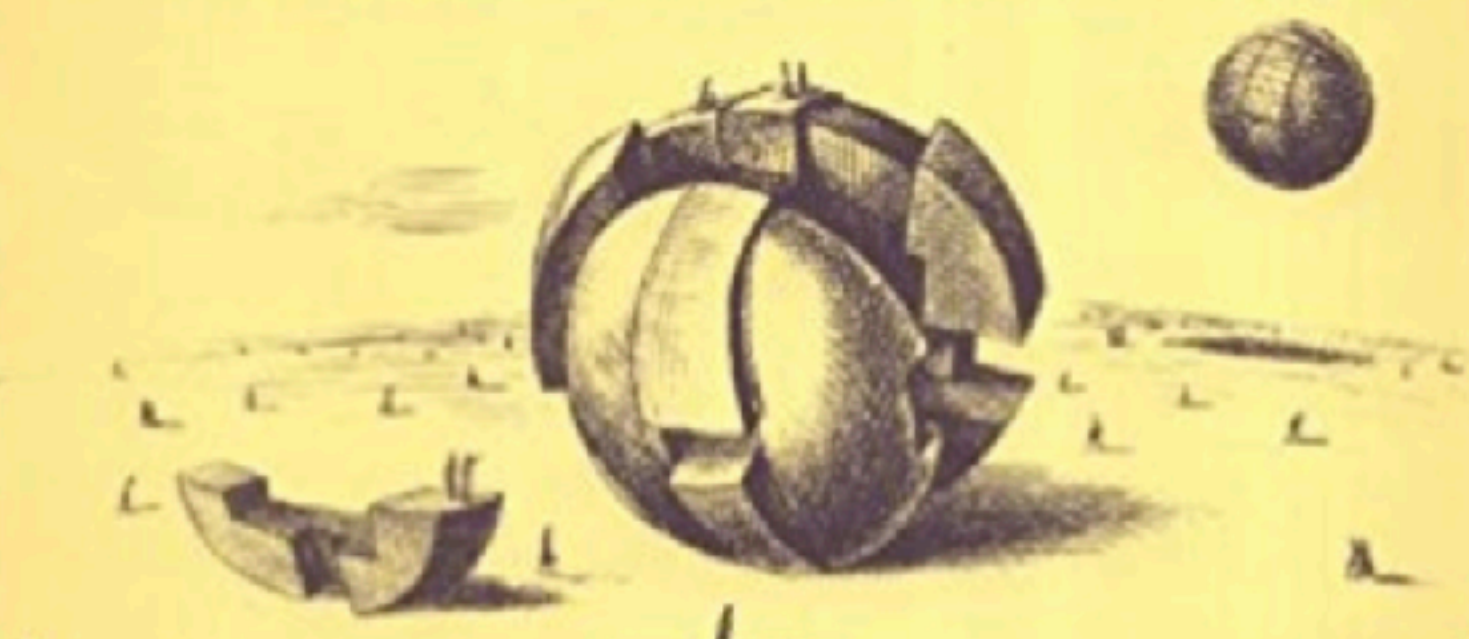
- Breiman, Leo, and Nong Shang. "Born again trees." University of California, Berkeley, Berkeley, CA, Technical Report (1996).
- Buciluǎ, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.
- Vapnik, Vladimir, and Rauf Izmailov. "Learning using privileged information: similarity control and knowledge transfer." Journal of machine learning research 16.2023-2049 (2015): 2.
- Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
- Geras, Krzysztof J., et al. "Blending lstms into cnns." arXiv preprint arXiv:1511.06433 (2015).
- Zagoruyko, Sergey, and Nikos Komodakis. "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer." arXiv preprint arXiv:1612.03928 (2016).
- Rusu, Andrei A., et al. "Policy distillation." arXiv preprint arXiv:1511.06295 (2015).
- Yim, Junho, et al. "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 2. 2017.
- Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models" Advances in neural information processing systems. 2017.
- Schmitt, Simon, et al. "Kickstarting Deep Reinforcement Learning." arXiv preprint arXiv:1803.03835 (2018).

THE SOCIETY OF MIND

"270 brilliantly original essays on... how the mind works."
—Isaac Asimov, Information Week

MARVIN MINSKY

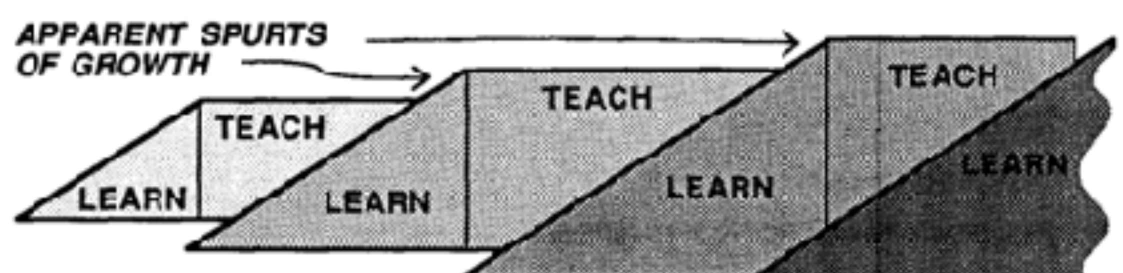
COFOUNDER OF THE ARTIFICIAL INTELLIGENCE LABORATORY, MIT



Minsky thought it first :p

17.1 SEQUENCES OF TEACHING-SELVES

Up to this point we've portrayed the mind as made of scattered fragments of machinery. But we adults rarely see ourselves that way; we have more sense of unity. In the next few sections we'll speculate that this coherency is acquired over many "stages of development." Each new stage first works under the guidance of previous stages, to acquire some knowledge, values, and goals. Then it proceeds to change its role and becomes a teacher to subsequent stages.



Overlapping Stages of Development

Extra credits to the conversations with:
Pratik Chaudhari, Kamyar Azizzadenesheli, Seb Arnold, Rich Caruana, Sammy Bengio & all the participants of NIPS 2017 Metalearning workshop

This work was supported by the National Science Foundation (CCF-1317433 and CNS-1545089), the Office of Naval Research (N00014-13-1-0563), C-BRIC (one of six centers in JUMP, a Semiconductor Re- search Corporation (SRC) program sponsored by DARPA), Intel Corporation and Amazon.com, inc. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.