# Finding the Right Teacher for a Difficult Student

Daniel Whettam

The University of Edinburgh / The University of Bristol

March 22, 2020

# Overview


1. Network Distillation
2. Neural Architecture Search (NAS)
3. Combining Distillation with NAS
4. Fisher Distillation
5. Conclusions

# Network Distillation

Two ways of thinking about it:

- Using large networks to improve the performance of a small networks
- Compressing a large network into a smaller one
- Key idea: When we can't simply scale up our network (depth, width etc.) how can we improve network performance?

# Network Distillation

Basic Idea:

- If we have large model (teacher) that works well, we can use it to guide the training procedure of a smaller model (student)

We can do this through a few ways:

- Simply make predictions with our teacher and use those new labels to train the student
- Introduce another term into the loss function when training the student network:
  - to penalise the student if predictions are different to the predictions of our teacher (knowledge distillation)
  - or to penalise the student if the activations at specified layers are different to those in the teacher (attention transfer)

# Network Distillation - Knowledge Distillation

- Encourages the students predictions to be similar to the teachers predictions (Hinton et al., 2015)

$$\mathcal{L}_{KD} = (1-\alpha)\mathcal{L}_{CE}(\mathbf{y}, \sigma(\mathbf{s})) + \alpha T^2 \mathcal{L}_{CE}\left(\sigma\left(\frac{\mathbf{t}}{T}\right), \sigma\left(\frac{\mathbf{s}}{T}\right)\right) \tag{1}$$

- First term is the standard cross entropy for the student network
- Second term is the cross entropy loss between the teacher and student, with a normalising term.

# Network Distillation - Attention Transfer

- Encourages the activations at each layer of the student to be similar to the activations of the teachers (Zagoruyko and Komodakis, 2016)

$$\mathcal{L}_{AT} = \mathcal{L}_{CE}(\mathbf{y}, \sigma(\mathbf{s})) + \beta \sum_{i=1}^{N_L} \left\| \frac{\mathbf{f}\left(A_i^t\right)}{\left\|\mathbf{f}\left(A_i^t\right)\right\|_2} - \frac{\mathbf{f}\left(A_i^s\right)}{\left\|\mathbf{f}\left(A_i^s\right)\right\|_2} \right\|_2 \tag{2}$$

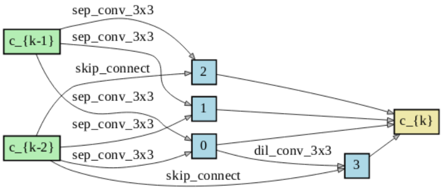
- First term is again the standard cross entropy loss
- Second term works like a regulariser. Penalises loss if the activation maps vary a lot between teacher and student
- Outperforms knowledge distillation
- We can also combine knowledge distillation and attention transfer

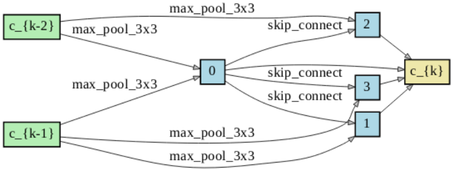
# Neural Architecture Search

1. Can we learn an architecture specific for our task?
2. Traditionally done via genetic algorithms or reinforcement learning
    - Very computationally expensive
    - e.g. SOTA RL approach costs 2000 GPU hours (Zoph et al., 2018)
3. Can also be done through gradient descent (Liu et al., 2018)

# Differentiable Architecture Search (DARTS)

- Relaxes the search space so it can be optimised via gradient descent
- Done by considering multiple possible operations for each node in a DAG. Optimised to find the most probable operations for the task.



# Combining Distillation with NAS

- My work asks the question of how can we do network distillation when we are working with models generated through NAS?
- What makes a good teacher for a NAS model?
- Can we use something off-the-shelf (e.g. ResNet), or do we need some way of growing the student into a teacher?

# Combining Distillation with NAS

- I focused on CNN's with CIFAR-10, however the ideas should remain the same for feed-forward nets.
- Initially I looked at how a NAS student performs when using various off-the-shelf networks as a teacher with attention transfer

| Teacher | Student | Top-1 Error | Top-5 Error |
|---|---|---|---|
| DenseNet | DenseNet | $6.3000 \pm 0.35$ | $0.1833 \pm 0.02$ |
| DenseNet | WRN | $6.2467 \pm 0.12$ | $0.2000 \pm 0.04$ |
| DenseNet | DARTS | $8.4600 \pm 1.30$ | $0.24 \pm 0.04$ |
| WRN | DenseNet | $5.7467 \pm 0.33$ | $0.1767 \pm 0.02$ |
| WRN | WRN | $6.1567 \pm 0.17$ | $0.2200 \pm 0.02$ |
| WRN | DARTS | $8.5433 \pm 0.14$ | $0.2233 \pm 0.02$ |
| DARTS | DenseNet | $6.7833 \pm 0.02$ | $0.2233 \pm 0.01$ |
| DARTS | WRN | $6.6767 \pm 0.42$ | $0.2067 \pm 0.01$ |
| DARTS | DARTS | $7.7900 \pm 0.30$ | $0.2100 \pm 0.01$ |

# Combining Distillation with NAS

- Teachers that are similar to the student seem to work well
- DARTS students perform very badly regardless of teacher, although a DARTS teacher is best
- Clearly something more sophisticated is needed to teach a DARTS model.
- I then looked at how to go about generating a teacher for a given student...

# Fisher Distillation

- Fisher information is a measure of how much information a known variable, $X$, contains about an unknown parameter, $\theta$. (Lehmann and Casella, 2006)
- We can use this to estimate which channel of the CNN has the biggest impact on the loss function:

$$\Delta_c = \frac{1}{2N} \sum_{n=1}^{N} g_{nc}^2 \tag{3}$$

- $g_{nc}^2$ is the gradient with respect to the $n^{th}$ data point for the channel of interest, $c$.
- Can then grow the student by growing the number of channels in the cell of the most important channel

# Fisher Distillation

I then created three different teacher networks for attention transfer with a 10 layer DARTS student (DARTS_V1_10)

- DARTS_V1_25
  - The same model as the DARTS student, but with 25 layers
- DARTS_V2_10 - u
  - The number of channels in each cell has been scaled uniformly until the model is approximately twice the size of the student
- DARTS_V1_10 - f
  - The number of channels in each cell has been scaled using fisher information until the model is approximately twice the size of the student

# Fisher Distillation

| Teacher | Student | Top-1 Error | Top-5 Error |
|---|---|---|---|
| N/A | DARTS_V1-10 | $8.4700 \pm 1.11$ | $0.2200 \pm 0.04$ |
| DARTS_V1-25 | DARTS_V1-10 | $7.7900 \pm 0.30$ | $0.2100 \pm 0.01$ |
| DARTS_V2-10-u | DARTS_V1-10 | $6.4233 \pm 0.17$ | $0.1833 \pm 0.06$ |
| DARTS_V2-10-f | DARTS_V1-10 | $6.4033 \pm 0.08$ | $0.1333 \pm 0.03$ |

# Fisher Distillation

- Fisher info works well, although not much better than uniform
- A depth-wise scaling improves on the baseline, but is quite bad.
- More generally, similar networks seem to make a good teacher/student pairing. This can be visualised by looking at the activations:

# Fisher Distillation

No teacher, DenseNet student

DenseNet teacher, DenseNet student

WRN teacher, DenseNet student

DARTS teacher, DenseNet student

# Conclusions

- Distillation can be effective, but only when a good teacher/student pairing is found
- Similar architectures give better pairings
- Non-standard architectures e.g. NAS require a bespoke teacher.
- Creating this teacher via a channel wise scaling seems to work well
- It may be possible to maximise performance by using Fisher information to guide this scaling, however not conclusive.

# Applications to Apple/Siri

- It is probable that performance can be improved through distillation, particularly attention transfer
- It is also possible that neural architecture search may give better results
- If so, combining the two may be a worthwhile consideration. This work proposes a way to do that

# References I

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Liu, H., Simonyan, K., and Yang, Y. (2018). Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Zagoruyko, S. and Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710.