Meta-meta-learning for Neural Architecture Search through arXiv Descent

Antreas Antoniou
MetaMind
aa@mm.ai

Nick Pawlowski
Googel x^2
nick@x.x

Jack Turner
slow.ai
jack@slow.ai

James Owers
Facebrook AI Research Team
jim@fart.org

Joseph Mellor
Institute of Yellow Jumpers
joe@anditwasall.yellow

Elliot J. Crowley
ClosedAI
elliot@closed.ai

Abstract

Recent work in meta-learning has set the deep learning community alight. From minute gains on few-shot learning tasks, to discovering architectures that are slightly better than chance, to solving intelligence itself[1], meta-learning is proving a popular solution to every conceivable problem ever conceivably conceived ever.

In this paper we venture deeper into the computational insanity that is meta-learning, and potentially risk exiting the simulation of reality itself, by attempting to meta-learn at a third learning level. We showcase the resulting approach—which we call *meta-meta-learning*—for neural architecture search. Crucially, instead of *meta-learning* a neural architecture differentiably as in DARTS (Liu et al., 2018) we *meta-meta-learn* an architecture by searching through arXiv. This *arXiv descent* is GPU-free and only requires a handful of graduate students. Further, we introduce a regulariser, called *college-dropout*, which works by randomly removing a single graduate student from our system. As a consequence, procrastination levels decrease significantly, due to the increased workload and sense of responsibility each student attains. The code for our experiments is publicly available at [redacted]. Edit: we have decided not to release our code as we are concerned that it may be used for malicious purposes.

1 Introduction

Meta-learning, originally described by Donald B. Maudsley (1979) was invented by Jürgen Schmidhuber (Schmidhuber, 1997) in the great renaissance of 1997. The idea is believed to have come to him as a residual (He et al., 2016) effect of the inhalation of cosmic matter originating from a rift in space-time caused by the great old one, Shub-Niggurath (Lovecraft & Niggurath, 1923) although the details of this—and cosmic horrors more generally—are beyond the scope of this work and human comprehension.

Meta-learning, or *learning to learn*, or *post-GAN-hypetrain* is a learning paradigm involving approximately two levels of abstraction. Consider MAML (Finn et al., 2017): the objective is to learn

[1] Probably, DeepMind wouldn't tell us when we asked.

Preprint. Rejected work.

Figure 1: A mammal. This is not to be mistaken for MAML, the popular meta-learning algorithm, but is equally as difficult to train.

a good set of initial weights for a neural network (Schmidhuber, 1997), such that it can quickly adapt to a few-shot classification task on unseen data. The lower level in this case is learning from each individual task in the training data. The higher, or *Hintonian* level is learning the across-task information. This involves calculating some second-order derivatives, but fortunately autograd means we don't have to understand what is actually going on. An illustration of a mammal is given in Figure 1 for clarity.

DARTS (Liu et al., 2018)—not to be mistaken for darts (Wikipedia, 2019)— performs neural architecture search or NAS (Zoph et al., 2018; Wu et al., 2018; Zhang et al., 2018) in a similar manner. The lower level of learning is concerned with classifying 32×32 images of frogs or boats (Krizhevsky, 2009)—a task which naturally extends to a whole host of real-world applications—and the higher level is learning the architecture with which to do this.

In this paper, we explicitly add *another* level of abstraction which we sycophantically term the **Schmidhubrian level** for neural architecture search. At a level this high, one or more graduate students search through arXiv—a process which we term arXiv Descent—for meta-learning papers, that learn-to-learn neural networks that perform optimally on a given task. As this task is always one of CIFAR, Omniglot, or a variant of ImageNet, this narrows down the search somewhat. Once they have obtained a good meta-learning system they pass this architecture one level down to the *Hintonian* level. At this level, another graduate student, usually one collaborating or being supervised by the Schmidhubrian-level graduate student, will apply the selected learning-to-learn algorithm on a novel new set of tasks/CIFAR-10. There is a non-negligible probability that the student will just use a CapsulesNet (Sabour et al., 2017) for the fun of it. Finally, at the lowest level the network is trained using a whole host of carefully thought-out[2] hyperparameters.

2 Method

We begin by writing a project proposal for MSc and PhD students. Once submitted, we begin the interview procedures. At this stage, a multitude of PhD/MSc students are examined for their ability to digest highly complex literature, produce creative solutions to previously unseen problems[3] and work consistently and reliably for an average of 90 hours a week or 18 hours a day[4]. Once the interviews have completed, we mostly chose the students that we liked the most, based on anything other than quantitative/objective information.

We then teach our students how to descend arXiv. arXiv descent works as follows; first the arXiv identifier is initialised following the Xavier uniform scheme, with two digits for year (YY), two for month (MM), a period (.), and a 4 digit submission number.

[2]We decided to not harm the climate by running an extensive optimisation using an unseen amount of GPUs.

[3]This is a major requirement for meta-meta-learning.

[4]As it is industry standard in the field; see `https://twitter.com/twinaki/status/908085572283092996`

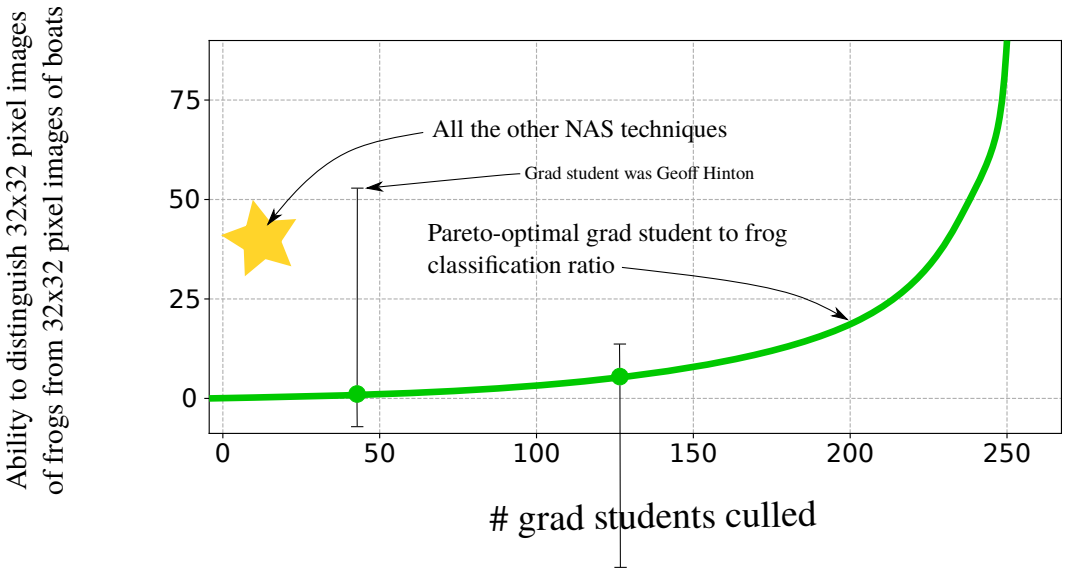

Figure 2: Experimental results. We only had two datapoints so we took the liberty of fitting this green curve to them. The star shows all the other NAS techniques, because they're all the same as random.

Graduate students then iterate the architecture by accessing the paper with the given identifier. If the paper is vaguely related to image classification or computer vision, they adapt the given setup with a probability of $p(\text{adaptarchitecture} \mid \text{CVpaper}) = \pi_{adapt}$ or alternatively decrease the 4-digit submission number by 1. Decreasing the submission number leads to the students discovering earlier work. Earlier work is often better work, as flag-planting methodology using half-baked experiments is highly desirable.

If the paper is not related to images, the student increases the month and year digits following the rules of the Gregorian calendar in the hope of finding a paper with pretty pictures. By increasing the date of the paper that is examined, we increase the probability of hitting a paper published within the period of GAN-hype, which led to the generation of many pretty images without any real application [5]. Nevertheless, such papers work on images and therefore hold useful architectures.

We implement early stopping (Caruana et al., 2001) by finely cherry-picking results to best suit our hypothesis. In cases where students are not converging fast enough, we also introduce several arbitrary hyperparameters to the optimisation process to both bewilder them and reduce internal covariate shift. Graduate students are dropped out at random, or when they become unable to afford the completely insane fees for their programme.

3 Experimental Results

We found AmoebaNet (Real et al., 2018), which is quite good. Our search process can be observed in Figure 2.

4 Rethinking Meta-Meta-Learning

Meta-meta learning has recently been proposed. Because the field of deep learning research is so saturated, this means that in a few months someone can write a paper disputing this method. This is more fashionable, and easier to do than thinking up something original.

[5] As far as the authors are concerned, DeepFakes do not constitute a real-world application.

5 Related Work

This work is entirely novel. This is why this "Related Work" section has been placed at the end as an afterthought. The only related works are previous works of the authors. We therefore acknowledge the act of unnecessary self-citation of barely relevant papers (Crowley & Pawlowski, 2015),

Although meta-meta-meta-learning has been proposed through the scientific medium of Twitter [6],we have found it impossible to implement in Keras (Chollet et al., 2015), and therefore cannot compare it to this work.

6 Conclusion

It should be obvious by now, that the decreasing size of the sections indicate that the authors are running out of steam. Furthermore, the submission deadline for this paper is effectively today, which further necessitates that we produce a complete paper. Hence, we shall conclude: Our technique is really good, and future work shall consist of whatever we think up next.

References

Caruana, R., Lawrence, S., and Giles, C. L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing systems*, 2001.

Chollet, F. et al. Keras, 2015.

Crowley, E. J. and Pawlowski, N. Neural network ensembles behave like a colony of bees. In *Retreats in Neural Information Processing Systems*, 2015.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

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

Lovecraft, H. and Niggurath, S. The colour out of space-time. *arXiv preprint arXiv:2311.01234*, 1923.

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. In *Advances in neural information processing systems*, 2017.

Schmidhuber, J. Musings of Jürgen Schmidhuber. In *International Conference on Jürgen Schmidhuber*, 1997.

Wikipedia. Darts. `https://en.wikipedia.org/wiki/Darts`, 2019.

Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., and Keutzer, K. FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. *arXiv preprint arXiv:1812.03443*, 2018.

Zhang, X., Huang, Z., and Wang, N. You only search once: Single shot neural architecture search via direct sparse optimization. *arXiv preprint arXiv:1811.01567*, 2018.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] https://twitter.com/fchollet/status/997250312359460869