SUBSTITUTE TEACHER NETWORKS: LEARNING WITH ALMOST NO SUPERVISION

Samuel Albanie*
British Institute of Learning, Yearning and Discerning
Shelfanger, UK

James Thewlis*
National Academy of Pseudosciences
Valencia, Spain

João F. Henriques*
Fortress of Solitude
Coimbra, Portugal

ABSTRACT

Education is expensive. Nowhere is that statement more universally agreed upon than in machine learning, a recently trending topic on twitter that places great value on the reduction of cost. Certainly for machines to learn, they must be taught, but how can this be achieved on an appropriate budget? Recent approaches (often referred to as *Teacher-Student* or *Knowledge Distillation* methods in the neural network literature) have demonstrated that the problem can be viewed as model compression, in which a single student model learns from an ensemble of M specialist consultants networks. Inspired by the logo on a free pen at a local recruitment fair, we scale this method *up* and *out*, while simultaneously pursuing an appropriately aggressive patenting strategy. In total, we make the following three contributions. First, we propose a novel *almost no supervision* training algorithm that is highly scalable in the number of student networks being supervised. Second, we explore the closely-related scaling problem of culinary optimisation, developing a method that tastily surpasses the current state of the art. Finally, we provide a rigorous quantitive analysis of our method, proving that we have access to a calculator.

A little learning is a dangerous thing

Alexander Pope, 1709

1 INTRODUCTION

Since time immemorial, learning has been the foundation of Human culture, allowing us to trick other animals into being our food. The importance of teaching in Ancient Times was exemplified by Pythagoras, who boasted of being able to teach his Theorem to anyone in the street (Philolaus of Croton, 421 BC), though apparently no one taught him to wear pants.

Nowadays, we are attempting to pass on this knowledge to our species' offspring, the machines (Timberlake, 2028; JT-9000, 2029)[1], who will hopefully keep us around to help with house chores.

*Authors listed in order of the number of guinea pigs they have successfully taught to play competitive bridge. Ties are broken alphabetically.

[1]The work of these esteemed scholars indicates the imminent arrival of general Artificial Intelligence. Their methodology consists of advising haters, who might be inclined to say that it is fake, to take note that it is in fact so real. The current authors, not having a hateful disposition, take these claims at face value.

Many prominent figures of our time, several of whom cannot tell their CIFAR-10 from their CIFAR-100 have expressed their reservations with this approach, but really, what can possibly go wrong?[2] Moreover, several prominent figures in our paper say otherwise (Fig. 1, Fig. 2).

Having established the wisdom of our approach as a whole with the extensive philosophical discussion above, we now press on to achieve a finer understanding of the details. Concretely, the goal of this work is to reduce the algorithmic ignorance, or more precisely *gnorance*[3] of a collection of student networks, and to do so in a fiscally responsible manner given a fixed teaching budget.

Define a collection of teachers $\{T_e\}$ as a class of highly educated functions which efficiently map unusual life experiences residing a Banach space into extremely unfair exam questions in an examination space. Further, define a collection of students $\{S_t\}$ as class of *keen beans* which inefficiently map unheated pot noodles to unwashed dishes, both in common space. Pioneering educational early work by Bucilua et al. (2006) demonstrated that on a carefully illuminated manifold, an arbitrary student S_t could improve his/her performance with N highly experienced, specialist teachers. We refer to this as the *private tuition* learning model. While effective in certain settings, this approach does not scale. Specifically, this algorithm scales in cost as $\mathcal{O}(\$MNK)$, where N is the number of students, M is the number of private tutors per student and $\$K$ is price the bastards charge per hour. Our key observation is that there is cheaper approach to ignorance reduction, which we detail in Sec. 3.

Our work is biologically inspired by the humble ostrich, an animal keenly aware of the dangers of learning too much, as its sand-based defence mechanism affords it a heightened inability to perceive threats. Advanced incomprehension of object permanence (Piaget, 1970) is also a key characteristic of human infants, as demonstrated empirically in the Stanford Peekaboo Experiment. This mental peculiarity is even more pronounced in certain human adults, with entire systems of contradictory beliefs able to be held simultaneously and without distress. Similarly, a profound ignorance of neuroscience allows the authors to confidently claim that the proposed method to cost reduction during teaching is identical to neural pathways found in the brain.

2 Related Work

> Give a student a fish and you feed them for day, teach a student to gatecrash seminars and you feed them until the day they move to Google.
>
> *Andrew Ng, 2012*

A worrying trend in the commoditization of education is the use of MOOC (Massive Open Online Courses) by large internet companies. They routinely train thousands of student networks in parallel with different hyperparameters, some of whom are hurled out to the far east on the explore-exploit coordinate chart, then keep only the top-performer of the class (Li et al., 2016; Snoek et al., 2012). We consider such practices to be wasteful and are totally not jealous at all of their impressive computational resources.

A number of approaches have been proposed to improve teaching quality. Central to each of these approaches is a question that has challenged researchers for many years, namely how best to efficiently extract extract knowledge that is *in the computer* (Zoolander, 2004). Work by noted entomologists Dean, Hinton and Vinyals illustrated the benefits of comfortable warmth to facilitate students better extracting information from their teachers (Hinton et al., 2015). In more detail, they advocated adjusting the value T in the softmax distribution:

$$p_i = \frac{\exp\left(x_i/T\right)}{\sum_j \exp\left(x_j/T\right)} \tag{1}$$

[2] This question is rhetorical, and should be safe to ignore until the Ampere release.

[3] The etymology of *gnorance* is a long and interesting one. Phonetic experts will know that the *g* is silent (cf. the silent *k* in knowledge), while legal experts will be aware that the preceding *i* is conventionally dropped to avoid costly legal battles with the widely feared litigation team of Apple Inc.

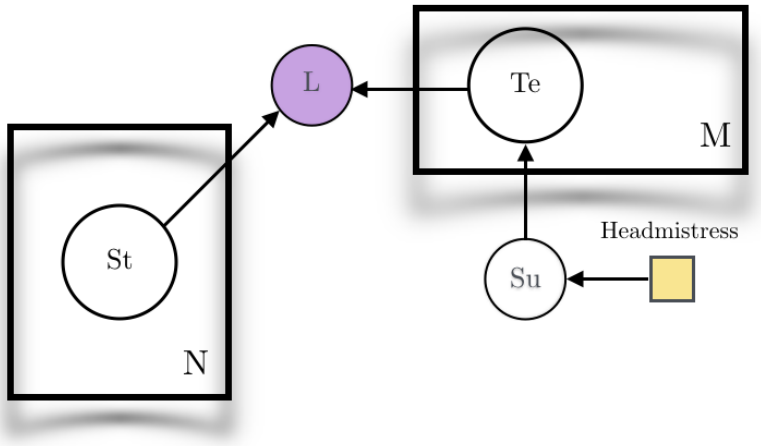

Figure 1: We introduce *Latent Substitute Teacher Allocation*, a simple generative process that explains the cost of learning. Note the use of drop-shadow plate notation, which indicates the direction of the nearest light source.

where T denotes the wattage of the classroom storage heater. More radical approaches have advocated the use of alcohol in the classroom, something that we do not condone directly, although we think it shows the right kind of attitude to innovation in education (Crowley et al., 2017). However, both approaches are clearly financially unsustainable. Moreover, differently from these works, we focus on the quantity, rather than the quality of our teaching method.

Recent work has promoted an "Attend, Infer, Repeat" (Eslami et al., 2016) approach to learning. Attendance is a prerequisite for our model, and cases of truancy will be reported to the headmistress (see Fig 1). For the substitute teacher module, the "Infer" step may be replaced by "Ignore". Only particularly badly behaved student networks will be required to repeat the course.

A number of pioneering ideas in scalable learning were physically investigated several years ago by (Maturana & Fouhey, 2013). However, we differentiate ourselves from their approach by using several orders of magnitude fewer hashtags. We also note the marginal relevance of a recent paper on unadversarial learning (Albanie et al., 2017). We now attempt to cite a future paper, from which we shall cite the current paper, in an ambitious attempt to send google scholar into an infinite depth recursion (Albanie et al., 2019), thereby increasing our academic credibility and assuredly landing us lucrative pension schemes.

2.1 Unrelated work

- A letter to the citizens of Pennsylvania on the necessity of promoting agriculture, manufactures, and the useful arts. George Logan, 1800
- Claude Debussy—The Complete Works. Warner Music Group. 2017
- Article IV Consultation—Staff Report; Public Information Notice on the Executive Board Discussion; and Statement by the Executive Director for the Republic of Uzbekistan. IMF, 2008
- A treatise on the culture of peach trees. To which is added, a treatise on the management of bees; and the improved treatment of them. Thomas Wildman. 1768

3 The Latent Substitute Teacher Allocation Process

The primary goal of educators is to educate, inform and explain. In machine learning, explanations are best encoded as simple statistical generative models. We therefore explain the role of cost efficient explanation through an appropriately simple explanation, the *Latent Substitute Teacher Allocation* (see Fig. 1).

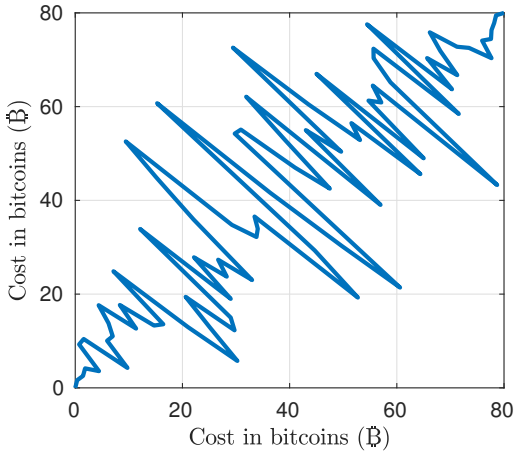

Figure 2: Expressing the cost function in bitcoins makes it significantly more volatile, yet it was instrumental in attracting venture capital for our Smart Education startup.

Fortunately, since the model is *graphical*, it needs minimal explanation. However, we can all agree that it will scale magnificently. All the teacher networks employed in the Latent Substitute Teacher Allocation Process are Recursive Neural Networks. A Recursive Neural Network is defined as the composition of some layers, and a Recursive Neural Network. By logical induction, these networks have infinite capacity, which is why they are not bothered by a heavy workload. All students are trained in two stages, separated by puberty.

In keeping with the cost-cutting focus, we have analysed the gradients available on the market, and after extensive research decided to use Synthetic Gradients Jaderberg et al. (2016), which are significantly cheaper than Natural Gradients Amari (1998). It is important to realise that our cost function, which is the target of minimisation, is very much proportional to actual cost (preferably cash; see Fig. 2).

Traditional approaches have often gone by the mantra that it takes a village to raise a child. We attempted to use a village to train our networks, but found it to be an expensive use of parish resources, and instead opted for the NVIDIA GTX 1080 Ti ProGamer-RGB. Installed under a desk in the office, this setup provided warmth during the cold winter months.

4 The Cake

As promised in the mouth watering abstract (and yet undelivered by the paper so far), we now take a short, mid-paper confectionary diversion to improve our ratings with the sweet-toothed demographic[4]. A number of competitive cakes have been recently proposed at a high-end cooking workshop (LeCun, 2016; Abbeel, 2017), resulting in a dramatic bake-off (Fig. 3-a,b).

Previous authors have focused on cherry-count. We show that better results can be achieved with more layers, without resorting to cherry-picking. Our layer cake consists of more layers than any previous cake (Fig. 3-c), showcasing the depth of our work.

We would like to dive deep into the technical details of our novel use of the No Free Lunch Theorem, Indian Buffet Processes and a Slow-Mixing Markov Blender, but we feel that increasingly thin culinary analogies are part of what's wrong with contemporary Machine Learning (Rahimi, 2017).

[4]This approach was recommended by our marketing team, who told us that everyone likes cake.

Figure 3: Several cakes of importance for current research (deeper is better). From left to right: 1) Yann LeCun's cake, 2) Pieter Abbeel's cake, 3) Our cake. Note the abundance of layers in the latter.

5 EXPERIMENTS

> If you don't know how to explain MNIST to your LeNet, then you don't understand digits!
>
> *Albert Einstein*

We now rigorously evaluate the efficacy of the Latent Substitute Teacher Allocation Process. We note that unlike previous methods, we achieve regularisation without injecting gradient noise. High noise levels tend to stop concentration gradients in student networks, and learning stalls. In these experiments we always operate in "library-mode". Performance-inducing drugs, such as batch-norm, were strictly prohibited.

After months of intensive training using our trusty NVIDIA desk-warmer, which we were able to compress down to two days using montage techniques and an 80's cassette of Survivor's "Eye of the Tiger", our student networks were ready for action. The only appropriate challenge for such well-trained networks, who eat digits for breakfast, was to pass the Turing test. We thus embarked on a journey to find out whether this test was even appropriate.

The Chinese Room argument, proposed by Searle (1980) in his landmark paper about the philosophy of AI, provides a counterpoint. It is claimed that an appropriately monolingual person in a room, equipped with paper, pencil, and a rulebook on how to respond politely to any written question in Chinese (by mapping appropriate input and output symbols), would appear from the outside to speak Chinese, while the person in the room would not actually understand the language. We ran this thought experiment many times using the highly scalable nature of the Latent Substitute Teacher Allocation Process. By sampling rulebook operators appropriately from the earth's surface, we achieved strong statistical guarantees that at least one of the monolingual subjects would be appropriately Chinese. Having resolved all philosophical and teleological questions, we then turned to the application of the actual Turing tests.

Analysing the results in Table 4, we see that only the ResNet-50 got a smiley face. The Q-network's low performance is obviously caused by the fact that it plays too many Atari games. However, we note that it could improve by spending less time on the Q's and more time on the A's. The Neural

Model	Turing test result
AlexNet	B-
ResNet-50	A+ ☺
Q-network	C
Neural Turing Machine	F-, see me after class

Figure 4: Results for the test class of 2018.

Turing Machine had an abysmal score, which we later understood was because it focused on an entirely different Turing concept.

As an additional, purely empirical statement, we observed that networks trained using our method experience a much lower DropOut rate. Some researchers set a DropOut rate of 50%, which we feel is unnecessarily harsh on the student networks[5].

6 Conclusion

> You take the blue pill—the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill—you stay in Wonderland, and I show you how deep the ResNets go.
>
> *Kaiming He, 2015*

This work has shown that it possible to achieve low-cost machine learning by using inexpensive, completely expendable Substitute Teacher Networks, while carefully avoiding their definition. We have seen that residual networks may be the architecture of choice for solving the Turing test. A major finding of this work, found during cake consumption, is that current networks have a Long Short-Term Memory, but they also have a Short Long-Term Memory. The permutations of Short-Short and Long-Long are left for future work, possibly in the short-term, but probably in the long-term.

Acknowledgements

This work was actively undermined by a wilful ignorance of related work.

References

Abbeel, Pieter. Keynote Address: Deep Learning for Robotics. 2017.

Albanie, Samuel, Ehrhardt, Sébastien, and Henriques, João F. Stopping gan violence: Generative unadversarial networks. *Proceedings of the 11th ACH SIGBOVIK Special Interest Group on Harry Quechua Bovik.*, 2017.

Albanie, Samuel, Ehrhardt, Sébastien, Thewlis, James, and Henriques, João F. Defeating google scholar with citations into the future. *Proceedings of the 13th ACH SIGBOVIK Special Interest Group on Harry Quechua Bovik.*, 2019.

Amari, Shun-Ichi. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Amazon. Details redacted due to active NDA clause.

Bucilua, Cristian, Caruana, Rich, and Niculescu-Mizil, Alexandru. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541. ACM, 2006.

Crowley, Elliot J, Gray, Gavin, and Storkey, Amos. Moonshine: Distilling with cheap convolutions. *arXiv preprint arXiv:1711.02613*, 2017.

Eslami, SM Ali, Heess, Nicolas, Weber, Theophane, Tassa, Yuval, Szepesvari, David, Hinton, Geoffrey E, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pp. 3225–3233, 2016.

Hinton, Geoffrey, Vinyals, Oriol, and Dean, Jeff. Distilling the knowledge in a neural network. In *Neural Information Processing Systems, conference on*, 2015.

[5]This technique, often referred to in the business management literature as Rank-and-Yank (Amazon), may be of limited effectiveness in the classroom.

Jaderberg, Max, Czarnecki, Wojciech Marian, Osindero, Simon, Vinyals, Oriol, Graves, Alex, and Kavukcuoglu, Koray. Decoupled neural interfaces using synthetic gradients. *arXiv preprint arXiv:1608.05343*, 2016.

JT-9000. How I learned to stop worrying and love the machines (Official Music Video). In *British Machine Vision Conference (BMVC)*, 2029. West Butterwick, just 7 miles from Scunthorpe, England.

LeCun, Yann. Keynote Address: Predictive Learning. 2016.

Li, Lisha, Jamieson, Kevin, DeSalvo, Giulia, Rostamizadeh, Afshin, and Talwalkar, Ameet. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*, 2016.

Maturana, Daniel and Fouhey, David. You Only Learn Once - A Stochastically Weighted AGGRegation approach to online regret minimization. In *Proceedings of the 7th ACH SIGBOVIK Special Interest Group on Harry Quechua Bovik.*, 2013.

Philolaus of Croton. How to bake the perfect croton. *Greek Journal of Fine, Fine Cuisine*, 421 BC.

Piaget, Jean. Piaget's theory. 1970.

Rahimi, Ali. Test of Time Award Ceremony. 2017.

Searle, John R. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959, 2012.

Timberlake, Justin. Filthy (Official Music Video). In *Pan-Asian Deep Learning Conference*, 2028. Kuala Lumpur, Malaysia.

Zoolander, Derek et. al. *Zoolander*. Paramount Pictures, 2004.