
THE CURIOUS CASE OF DEVELOPMENTAL BERTOLOGY

ON SPARSITY, TRANSFER LEARNING, GENERALIZATION AND THE BRAIN

Xin Wang
Cerebras Systems
poincare.disk@gmail.com

July 9, 2020

ABSTRACT

In this essay, we explore a point of intersection between deep learning and neuroscience, through the lens of large language models, transfer learning and network compression. Just like perceptual and cognitive neurophysiology has inspired effective deep neural network architectures which in turn make a useful model for understanding the brain, here we explore how biological neural development might inspire efficient and robust optimization procedures which in turn serve as a useful model for the maturation and aging of the brain.

Keywords deep learning · natural language processing · BERT · network compression · sparse neural network · neuroscience · neural development

This essay is written for machine learning researchers and neuroscientists (some jargons in both fields will be used)¹. Though it is not intended to be a comprehensive review of literature, we will take a tour through a selection of classic work and new results from a range of topics, in an attempt to develop the following thesis:

Just like the fruitful interaction between representation learning and perceptual/cognitive neurophysiology, a similar synergy exists between transfer/continual learning, efficient deep learning and developmental neurobiology.

Hopefully it would inspire the reader in one way or two, or at the very least, kill some boredom during a global pandemic.

We are going to touch on the following topics through the lens of large language models:

- How do overparameterized deep neural nets generalize?
- How does transfer learning help generalization?
- How do we make deep learning computationally efficient in practice?
- In tackling these questions, how might deep learning research benefit and benefit from scientific studies of the developing and aging brain?

1 A philosophical preamble

Before we start, it is prudent to say a few words about the brain metaphor, to clarify this author's position on the issue as it often arises central at debates.

The confluence of deep learning and neuroscience arguably took place as early as the conception of artificial neural nets, because artificial neurons abstract characteristic behaviors of biological ones (McCulloch and Pitts, 1943). However, the drastically different learning mechanisms and disparities in the kinds of intelligent functions erected a formidable barrier in between the two standing tall for decades. The success of modern deep learning in recent years rekindled another trend of integration, bearing new fruits. In addition to designing AI systems inspired by the brain (e.g. Hassabis et al., 2017), deep neural nets have recently been proposed to serve as a useful model system to understand how the

¹ Originally published on Towards Data Science.

brain works (e.g. Richards et al., 2019). The benefits are mutual. Progress is being made in reconciliation of the learning mechanisms (Lillicrap et al., 2020) but, in more than one significant aspect, the intelligence gap obstinately remain (Marcus, 2018, 2020).

Now, for a deep learning researcher or practitioner looking at this mixed landscape today, is a brain analogy *helpful* or *misleading*? It is of course simple to give an answer based on faith, and there are large numbers of believers on both sides. But for now let us not pick a side by belief. Instead, let us evaluate each analogy in its unique context entirely by its practical ramifications: *scientifically*, it is helpful only if it makes experimentally verifiable/falsifiable predictions, and *for engineering*, it is useful only if it generates candidate features that can be subject to solid benchmarking. As such, for all brain analogies we are going to raise in the rest of this essay, however appropriate or farfetched they might seem, we shall look past any prior principles and strive to articulate hypotheses that can guide future scientific and engineering work in practice, either within or beyond the limits of these pages.

2 The working analogy

What do we usually think of a deep neural net when likening it to the brain?

For most, the network architecture maps to the gross anatomy of brain areas (such as in a sensory pathway) and their interconnections, i.e. the connectome, units map to neurons or cell assemblies, and connection weights to synaptic strengths. As such, neurophysiology carries out the computation of model inference.

Learning of deep neural nets typically takes place given a pre-defined network architecture, in the form of optimizing an objective function over a training dataset. (A major difficulty lies in the biological plausibility of artificial learning algorithms, a topic we do not touch in this article — here we simply accept the similarity of function despite the differences in mechanism.) Thus, the data-driven learning by optimization is similar to experience-based neural development, i.e. *nurture*, whereas network architecture, and to a large degree initialization and some hyperparameters as well, are genetically programmed as a result of evolution, i.e. *nature*.

Remark. *It should be noted that modern deep net architectures, either implicitly engineered by hand or explicitly optimized through neural architecture search (NAS, for review see e.g. Wistuba et al., 2019), are also a consequence of data-driven optimization, engendering the inductive bias — the free lunch is paid for by all the unfit that failed to survive natural selection.*

Thanks to the rapid growth of data and computing power, the decade of 2010s saw a Cambrian explosion of deep neural net species, spreading rapidly across the world of machine learning.

3 BERTology

The plot thickens as the evolution of modern deep learning produces a cluster of new species in the past two years. They thrive in the continent of natural language understanding (NLU), on fertile deltas of mighty rivers carrying immense computing power, such as the Google and the Microsoft. These remarkable creatures share some key commonalities: they all feature a canonical cortical microcircuitry called the *transformer* (Vaswani et al., 2017), have rapidly increasing brain volumes setting historic records (e.g. Shoeybi et al., 2019; Microsoft Research, 2020; Brown et al., 2020) and are often scientifically named after one of the Muppets. But the most prominent common trait of these species crucial to their evolutionary success is the capability of *transfer learning*.

What does this mean? Well, these creatures have a two-stage neural development: a lengthy, self-supervised larval stage called *pre-training* followed by a fast, supervised maturation stage called *fine-tuning*. During self-supervised pre-training, huge corpora of unlabeled text are presented to the subject, who plays with itself by optimizing certain objectives very much similar to solving language quizzes given to human kids, such as completing sentences, filling in missing words, telling logical procession of sentences, and spotting grammatical errors. Then during fine-tuning, a well pre-trained subject can quickly learn to perform a particular language understanding task by supervised training.

Transfer learning’s sweeping conquest of the land of NLU was marked by the advent of bidirectional encoder representations from transformers (BERT, Devlin et al., 2018). BERT and its variants have advanced the state-of-the-art by a considerable margin. Their remarkable success piqued tremendous interest in the inner workings of these models, creating the study of “BERTology” (for review, see Rogers et al., 2020). Not unlike neurobiologists, BERTologists stick electrodes into the model brain to record activities for interpretation of the neural code (i.e. activations and attention patterns), make targeted lesions of brain areas (i.e. encoding layers and attention heads) to understand their functions, and study how experiences in early development (i.e. pre-training objectives) contribute to mature behavior (i.e. good performance in NLU tasks).

4 Network compression

Meanwhile, in the world of deep learning, multi-stage development (like transfer learning) happens in more animal kingdoms than one. Particularly, in production, one often needs to compress a trained huge neural net into a compact one for efficient deployment.

The practice of network compression derives from one of the very puzzling properties of deep neural nets: *overparameterization helps not only generalization but optimization as well*. That is to say, training a small network is often not only worse than training a large one (if one can afford to do so of course, Belkin et al., 2019), but also worse than compressing a trained large one to the same small size. In practice, compression can be realized by sparsification (pruning), distillation, etc.

Remark. *It is worth noting that the phenomenon of best sparse network arising from optimizing and then compressing a dense one (see e.g. Zhu and Gupta, 2017; Gale et al., 2019) is very much like the developing brain, in which over-produced connections are gradually pruned (Navlakha et al., 2018).*

The type of multi-stage development in model compression, however, is very different from transfer learning. The two stages of transfer learning see the same model being optimized for different objectives, whereas in model compression, the original model morphs into a different one in order to retain optimality for a same objective. If the former resembles maturation to acquire new skills, then the latter is more like graceful aging without losing already learned skills.

5 Learning weights vs. learning structures: a duality?

When a network is compressed, its *structure* often undergoes changes. It could mean either the *network architecture* (e.g. in the case of distillation) or *parameter sparseness* (e.g. in the case of pruning). These structural changes are usually imposed by heuristics or regularizers that constrain the otherwise already effective optimization.

But can *structure* rise above being merely an efficiency constraint and become an effective means for learning? An increasing number of emerging studies seem to suggest so.

One intriguing case is weight-agnostic networks (Gaier and Ha, 2019). These jellyfish-like creatures do not have to learn during their lifespan, but still are extremely well adapted to their ecological niches, because evolution did all the heavy lifting in choosing an effective brain structure for them.

Even with a fixed architecture chosen by nature, *learning sparse structure can still be as effective as learning synaptic weights*. Recently, Ramanujan et al. (2019) managed to find sparsified versions of initialized convolutional nets which, if made wide and deep enough, generalize no worse than dense ones undergoing weight training. Theoretical investigations also suggest that sparsification of random weights can be just as effective as optimizing parameters if the model is sufficiently overparameterized (Malach et al., 2020; Ye et al., 2020).

Thus, in the grossly overparameterized regime of modern deep learning, we have in sheath a doubled-edged sword: optimization of *weights* and of *structure*. This is reminiscent of both *synaptic* and *structural plasticity* as mechanisms underlying biological learning and memory (e.g., see Gage, 2004; Johansen-Berg, 2007).

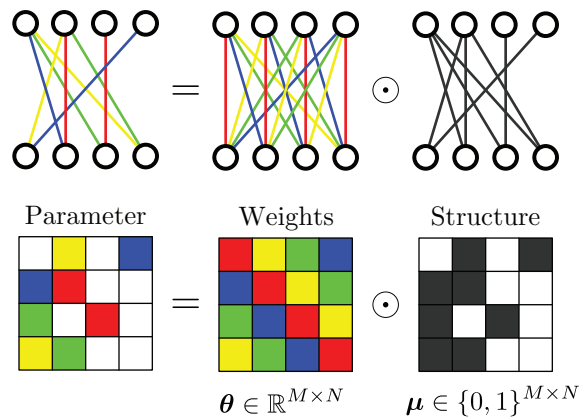


Figure 1: The parameter-mask formulation of structural sparseness of model parameters.

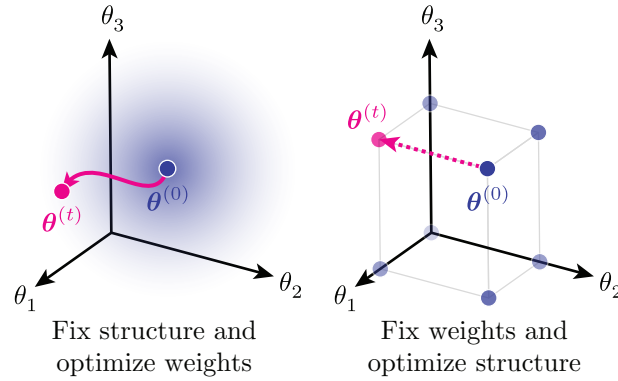


Figure 2: Learning weights versus learning structure.

Remark. A formal way of describing parameter sparseness is through the formulation of a parameter mask (Figure 1). Learning can be realized either by optimization of continuous weights within a fixed structure, or by optimization of discrete structure given a fixed set of weights (Figure 2).

6 Fine-tuning by sparsification

Now that structure, just like weights, can be optimized for learning, can this mechanism be used to make transfer learning better?

Yes, it can indeed. Recently, Radiya-Dixit and Wang (2020) made BERT pick up this new gene and evolve to something new. They showed that BERT can be effectively fine-tuned by sparsification of pre-trained weights without changing their values, as demonstrated systematically with the General Language Understanding Evaluation (GLUE) tasks (Wang et al., 2019).

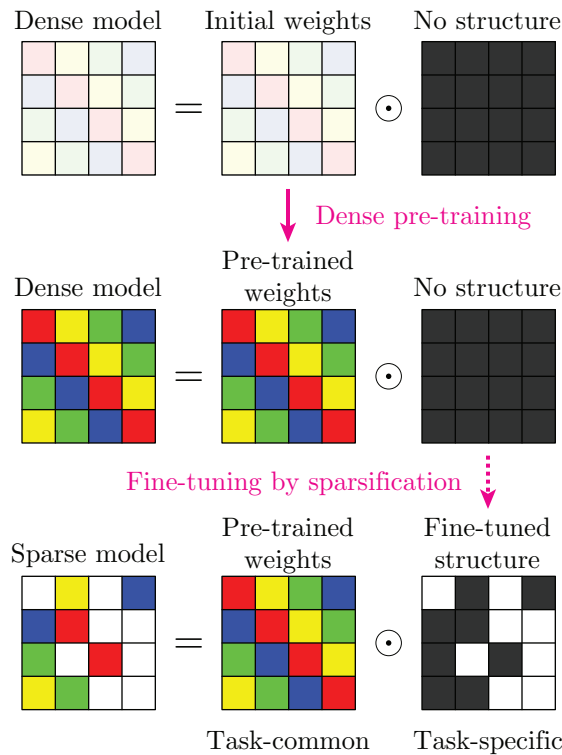


Figure 3: Fine-tuning BERT by sparsification (Radiya-Dixit and Wang, 2020).

Remark. Note that similar fine-tuning by sparsification has been successfully applied to computer vision, e.g. Mallya et al. (2018). Also take note of existing work sparsifying BERT during pre-training (Gordon et al., 2020).

Fine-tuning by sparsification has *favorable practical implications*. On the one hand, pre-trained parameter values remain the same in learning multiple tasks, reducing task-specific parameter storage to only a binary mask; on the other hand, sparsification compresses the model, potentially obviates many “multiply-by-zero-and-accumulate” operations with proper hardware acceleration. One stone kills two birds.

Beyond the practical benefits, however, the possibility of fine-tuning by sparsification brought about a few new opportunities towards a deeper understanding of language pre-training and its potential connections to the biological brain. Let us take a look of them in the next sections.

7 Winning tickets of a different lottery

First we study the nature of language pre-training from the perspective of optimization.

It seems that language pre-training meta-learns a good initialization for learning downstream NLU tasks. As Hao et al. (2020) recently showed, pre-trained BERT weights have good task-specific optima that are closer and flatter in loss landscape. This means pre-training makes fine-tuning easier, and the fine-tuned solutions generalize better.

Similarly, pre-training also makes discovery of fine-tuned sparse subnetworks easier (Radiya-Dixit and Wang, 2020). As such, interestingly, pre-trained language models have all the key properties of a “winning lottery ticket” as formulated by Frankle and Carbin (2018), but of exactly the complementary kind given the duality of optimizing weights vs. structure (Figures 3,4):

- The *Frankle-Carbin winning ticket* is a specific sparse structure that facilitates weight optimization. It is sensitive to weight initialization (Frankle and Carbin, 2018). It is potentially transferable across vision tasks (Morcos et al., 2019).
- A *pre-trained language model* is a specific set of weights that facilitates structural optimization. It is sensitive to structural initialization (Radiya-Dixit and Wang, 2020). It is transferable across NLU tasks (Radiya-Dixit and Wang, 2020).

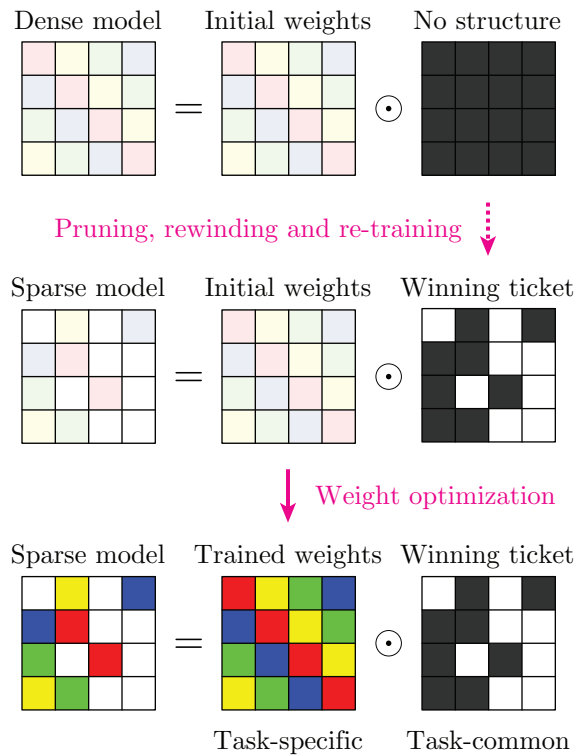


Figure 4: The *Frankle-Carbin winning ticket* (Frankle and Carbin, 2018), cf. fine-tuning by sparsification (Figure 3).

Remark. Note that the “winning ticket” property of pre-trained BERT is different from the wide-and-deep regime as in Ramanujan et al. (2019). It remains an open question whether large transformer-based language models, if made sufficiently wide and deep (bound to be astronomically large provided their already huge sizes), might be effectively fine-tuned from random initializations without pre-training.

Though learning weights of a winning lottery ticket and searching for a subnetwork within pre-trained weights lead to the same outcome — a compact, sparse network that generalizes well, the biological plausibility of the two approaches are drastically different: finding a Frankle-Carbin ticket involves repeated rewinding in time and re-training, a process only possible across multiple biological generations if earlier states could be genetically encoded and then reproduced in the next generation so as to realize rewinding. But weight pre-training followed by structural sparsification are similar to development and aging, all within a single generation. Thus, dense pre-training and sparse fine-tuning might be a useful model for neural development.

8 Robustness: same function from different structures

Another uncanny similarity between BERT and the brain is its *structural robustness*.

There seems to be an abundance of good subnetworks of pre-trained BERT at a wide range of sparsity levels (Radiya-Dixit and Wang, 2020): a typical GLUE task can be learned by eliminating from just a few percent to over half of pre-trained weights, with good sparse solutions exist everywhere in between (Figure 5, left). This is reminiscent of structural plasticity at play in the maturing and aging brain — its acquired function remains the same while the underlying structure undergoes continuous changes over time. This is very different from the brittle *point solutions* by traditional engineering.

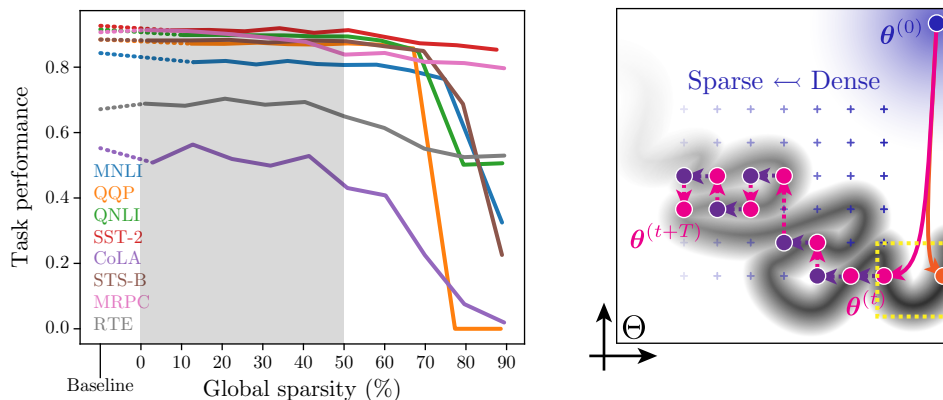


Figure 5: Structural robustness of fine-tuned language models by sparsification. (Left) There exist many good subnetworks of pre-trained BERT that span a wide range of sparsity (from a few percent to more than half Radiya-Dixit and Wang, 2020). (Right) A cartoonistic view of the loss landscape during continual sparsification. Dense training (solid magenta and orange arrows) finds low-loss solutions lying on a continuous manifold, dotted yellow box similar to Figure 1 of Draxler et al. (2018). As long as any structural perturbation by weight elimination (purple dotted arrows and circles) does not deviate far from the low-loss manifold, a quick structural fine-tuning (magenta dotted arrows and circles) can restore optimality, continually. The blue grid represents the discrete set of sparse parameters.

This phenomenon stems primarily from overparameterization of deep neural nets. In the modern regime of gross overparameterization, optima in the loss landscape are typically high-dimensional continuous non-convex manifolds (Draxler et al., 2018; Fort and Jastrzebski, 2019). This is strangely similar to biology, where identical network behavior can arise from vastly different underlying parameter configurations, forming a non-convex set in the parameter space, e.g. see Marder (2011).

Here comes the interesting part. Just like the life-long homeostatic adjustment in biology, a similar mechanism might support continual learning in overparameterized deep nets (illustrated in Figure 5, right): early-stage learning of dense connections finds a good solution manifold, along which an abundance of good sparse solutions exist; as the network ages, continual and gradual sparsification of the network can be quickly fine-tuned by structural plasticity (like the brain that maintains life-long plasticity). Having a large, but sparse and plastic brain has functional advantages (e.g. Ahmad and Scheinkman, 2019).

Table 1: Fine-tuning by sparsification of quantized pre-trained parameters. Shown are F1 scores of fine-tuned BERT and related models for MRPC, mean \pm S.D. of 3 independent runs. Thanks to Hugging Face’s `transformer`, experiments like this are a breeze.

Model	Fine-tune weights	Fine-tune structure	Fine-tune structure on weights quantized to			
			8-bit	4-bit	2-bit	1-bit
bert-base	.8890 \pm .0061	.9035 \pm .0032	.9059 \pm .0044	.8935 \pm .0086	.8122 \pm .0000	.5338 \pm .3512
bert-large	.8968 \pm .0148	.8996 \pm .0020	.8995 \pm .0101	.8977 \pm .0082	.8122 \pm .0000	.3792 \pm .1595
xlnet-base	.8950 \pm .0159	.9061 \pm .0048	.9009 \pm .0034	.8929 \pm .0021	.8129 \pm .0010	.8122 \pm .0000
xlnet-large	.9132 \pm .0051	.9048 \pm .0038	.9015 \pm .0312	.8918 \pm .0152	.8122 \pm .0000	.8122 \pm .0000
roberta-base	.9131 \pm .0057	.9031 \pm .0025	.9046 \pm .0095	.8459 \pm .0275	.8122 \pm .0000	.8122 \pm .0000
roberta-large	.9158 \pm .0028	.9186 \pm .0066	.9124 \pm .0069	.8638 \pm .0061	.8122 \pm .0000	.7699 \pm .0733
albert-base	.9008 \pm .0056	.8984 \pm .0064	.8962 \pm .0083	.8945 \pm .0054	.8086 \pm .0059	.6906 \pm .0377
albert-large	.9108 \pm .0021	.9080 \pm .0029	.9124 \pm .0094	.8251 \pm .0223	.8121 \pm .0001	.6682 \pm .0305
albert-xlarge	.9043 \pm .0025	.9096 \pm .0062	.9162 \pm .0064	.9124 \pm .0060	.8122 \pm .0000	.3656 \pm .2633

From the neurobiological perspective, if one accepts *the optimizational hypothesis* (Richards et al., 2019), then the life-long plasticity must carry out some functional optimization continually during lifespan. Following this logic, neural developmental disorders that arise from this process going awry should essentially be *optimizational diseases*, with etiological characterizations such as bad initialization, unstable optimizer dynamics, etc.

Whether the aforementioned hypothesis holds true for deep neural nets in general, and adequate for them to serve as a good model for neural development and pathophysiology, are open questions for future research.

9 How much did BERT learn?

Finally, let us apply some neuroscientific thinking to BERTology.

We ask the question: how much information is stored in *pre-trained* BERT parameters relevant for solving an NLU task? It is not an easy question to answer because sequential changes in parameter values during pre-training and during fine-tuning confound each other.

This limitation is no longer there in the case of BERT fine-tuned by sparsification, where pre-training only learns weight values and fine-tuning only learns structure. To a biologist, it is always good news if two stages of development involve completely different physiological processes, in which case one of them can be used to study the other.

Now let us do exactly this. Let us perturb the pre-trained weight values and study the downstream consequences. For this experiment, we do not make physiological perturbations (such as lesioning attention heads), but a pharmacological one instead: systemic application of a substance that affects every single synapse in the entire brain. This drug is *quantization*. Table 1 summarizes some preliminary dose-responses: though BERT and related species have developed large brains, it seems knowledge learned during language pre-training might be described by just a few bits per synapse.

In practice, this means that, since pre-trained weights do not change values during fine-tuning by sparsification, one might only need to store a low-precision integer version of all BERT parameters without any adverse consequences — a significant compression. The upshot: *all you need is a quantized integer version of pre-trained parameters shared across all tasks, with a binary mask fine-tuned for each task.*

Remark. Note that existing work on quantization of BERT weights quantizes fine-tuned weights (e.g. *Q-BERT*, Shen et al., 2019) instead of pre-trained weights.

10 Epilogue

Deep neural nets and the brain have obvious differences: at the lowest level, in learning algorithms, and at the highest level, in general intelligence. Nevertheless, profound similarities at intermediate levels have proven beneficial for the advancement of both deep learning and neuroscience.

For instance, perceptual and cognitive *neurophysiology* has already inspired *effective deep network architectures* which in turn make a useful model for understanding the brain. In this essay, we proposed another point of intersection: biological *neural development* might inspire *efficient and robust optimization procedures* which in turn serve as a useful model for maturation and aging of the brain.

Remark. *It should be noted that neural development in the context of traditional connectionism was proposed in the 1990s (e.g. see Elman et al., 1996).*

Specifically, we have reviewed some recent results on weight learning and structural learning as complementary means to optimization, and how they, in combination, realize efficient transfer learning in large language models.

As structural learning becomes increasingly important in deep learning, we shall see corresponding hardware accelerators emerge (e.g. Nvidia’s Ampère architecture supporting sparse weights, NVIDIA Blog, 2020). This is likely to bring about a new wave of architectural diversification of specialized hardware — acceleration of structural learning requires smart data movement adapted to specific computations, a new frontier for exploration.

Bibliography

- Ahmad, S. and Scheinkman, L. (2019). How Can We Be So Dense? The Benefits of Using Highly Sparse Representations.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. (2018). Essentially No Barriers in Neural Network Energy Landscape.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.
- Fort, S. and Jastrzebski, S. (2019). Large Scale Structure of Neural Network Loss Landscapes.
- Frankle, J. and Carbin, M. (2018). The Lottery Ticket Hypothesis: Finding Small, Trainable Neural Networks.
- Gage, F. H. (2004). Structural plasticity of the adult brain.
- Gaier, A. and Ha, D. (2019). Weight Agnostic Neural Networks.
- Gale, T., Elsen, E., and Hooker, S. (2019). The State of Sparsity in Deep Neural Networks.
- Gordon, M. A., Duh, K., and Andrews, N. (2020). Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning.
- Hao, Y., Dong, L., Wei, F., and Xu, K. (2020). Visualizing and understanding the effectiveness of BERT. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 4143–4152. Association for Computational Linguistics (ACL).
- Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence.
- Johansen-Berg, H. (2007). Structural Plasticity: Rewiring the Brain.
- Lillicrap, T. P., Santoro, A., Marris, L., and Akerman, C. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, pages 1–12.
- Malach, E., Yehudai, G., Shalev-Shwartz, S., and Shamir, O. (2020). Proving the Lottery Ticket Hypothesis: Pruning is All You Need.
- Mallya, A., Davis, D., and Lazebnik, S. (2018). Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11208 LNCS, pages 72–88.
- Marcus, G. (2018). Deep Learning: A Critical Appraisal.
- Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence.
- Marder, E. (2011). Variability, compensation, and modulation in neurons and circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 108(SUPPL. 3):15542–15548.

- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.
- Microsoft Research (2020). Turing-NLG: A 17-billion-parameter language model by Microsoft.
- Morcos, A. S., Yu, H., Paganini, M., and Tian, Y. (2019). One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers.
- Navlakha, S., Bar-Joseph, Z., and Barth, A. L. (2018). Network Design and the Brain. *Trends in Cognitive Sciences*, 22(1):64–78.
- NVIDIA Blog (2020). What Is Sparsity in AI Inference?
- Radiya-Dixit, E. and Wang, X. (2020). How fine can fine-tuning be? Learning efficient language models.
- Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. (2019). What’s Hidden in a Randomly Weighted Neural Network?
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience.
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works.
- Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. (2019). Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT.
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism.
- Vaswani, A., Uszkoreit, J., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. (Nips).
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.
- Wistuba, M., Rawat, A., and Pedapati, T. (2019). A Survey on Neural Architecture Search.
- Ye, M., Gong, C., Nie, L., Zhou, D., Klivans, A., and Liu, Q. (2020). Good Subnetworks Provably Exist: Pruning via Greedy Forward Selection.
- Zhu, M. and Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression.