Research Statement *Liyuan Liu*

Achieving better resource productivity is crucial to accommodate the ever-growing resource consumption of neural networks, and such productivity is in great demand to advance cutting-edge models or apply them to real-world applications. I strive to build *more productive algorithms* via 1) reducing the reliance on expert annotations by utilizing other available resources; 2) reducing the computational burden by simplifying the training workflow.

OVERVIEW

The resource consumption of cutting-edge neural networks is higher than ever before. While growing computation power and data volume keep breaking the glass ceiling for neural models, they also bring great challenges. For example, although specialized hardware devices (e.g., TPU) make it possible to conduct large-scale training (e.g., BERT), the sometimes multimillion-dollar cost hinders most researchers and organizations from developing or applying cutting-edge models. Also, while the construction of the ImageNet dataset spawns the boom of deep learning, the reliance on expert annotations makes it hard to deploy cutting-edge models on low-resource domains.

Despite the effectiveness of cutting-edge models, it has been quite challenging for most researchers and organizations to meet their training cost. Since deep learning has an unquenchable thirst for resources, it is in great demand to develop more productive algorithms. Besides, given the resource-dependent nature, productive algorithms have a unique potential to advance deep learning.

I endeavor to build productive algorithms that can effectively reduce the resource consumption of deep learning. Since model training incurs two types of cost (annotation cost and computational cost), my research is naturally formed as a "pincer movement" and has two fronts. First, I leverage other available resources to supervise model training. Instead of confining to expert annotations, my research broadens supervision sources and empowers various applications. For example, when line-by-line annotations are not available, we successfully construct biomedical information extraction models with knowledge bases. Second, I seek to reduce unnecessary computations by simplifying and automating the training workflow. By analyzing existing trial-and-error approaches, I develop perspectives and theories to guide algorithm design, allowing my algorithms to be automatically adapted to different scenarios with minimal tuning.

Impacts. Besides novel ideas, my research brings practically applicable algorithms. As of Nov. 2020, my projects have been cited 1,000+ times, received 3,500+ stars (i.e., upvotes) and 600+ forks on Github.com, been awarded a Grand Prize in the Topcoder Named Entity Recognition Challenge, placed first in various benchmarks (e.g., WMT'14 En-Fr), been downloaded 42,000+ times, and been successfully deployed in a wide range of applications (from entity recognition systems for COVID-19 literature to large-scale online recommendation).

BROADEN SUPERVISION SOURCES

Typical deep learning models depend heavily on human annotations and fail to utilize other available resources. However, in many scenarios, annotations alone are insufficient to drive model training, e.g., for tasks like biomedical information extraction or legal document translation, they usually do not have sufficient annotations due to the high cost to recruit domain experts, while there exist other available resources like knowledge bases.

I strive to give algorithms the flexibility to leverage all available resources in model training, instead of confining to gold-standard annotations. Specifically, I try to 1) automatically convert other available resources to pseudo-labels, and 2) employ language models to conduct training in a self-supervised manner.

Resource-Label Conversion. Many attempts have been made to automatically convert knowledge bases and heuristic rules into pseudo-labels for model training. At the expense of quality, these labels obtain great scalability and low cost. Consequently, label noise has been long regarded as the sole challenge brought by these pseudo-labels. In my study, I challenge such common wisdom with systematic analyses—my study recognizes that label noise is not the only challenge and further reveals an important, but long-overlooked, issue [16][1]. Specifically, resulted from the bias of handcrafted heuristics, the distribution of pseudo-labels is shifted from the true label distribution. I further introduce a simple yet effective method to neutralize such shifts, which leads to consistent performance improvements across various scenarios.

Moreover, I design task-specific auxiliary tasks to complement pseudo-labels [20]. Meanwhile, I incorporate resources other than knowledge bases to generate pseudo-labels (e.g., domain-specific patterns) and propose a novel model to refine the label quality [24]. Also, my work benefits classic supervised learning by confronting label noise. It leads to consistent improvements by identifying and down-sampling low-quality instances [15].

Employ Language Models. Benefitted from the nearly unlimited corpora, neural language models can learn to generate high-quality sentences from scratch. Recognizing its ability to automatically capture language structures, my work was the first to leverage character-level language models [23], whose direct follow-up methods serve as state-of-the-art systems on several benchmarks.

Since language modeling requires no annotations, the scale of language modeling style pre-training becomes larger and larger. Although these efforts reduce the reliance on human annotation, they lead to significant computation cost. Recognizing such computation-intensive nature, my work was the first to accelerate pre-trained language models in down-stream tasks [22]. With structured pruning, my method can accelerate model inference to about 7-time faster without much performance loss.

SIMPLIFY THE TRAINING WORKFLOW

A common criticism of deep learning is the lack of theoretical guidance and the requirement of excessive tuning. Intuitively, without understanding the underlying mechanism, tunings in various aspects are required. Typical practices employ a trial-and-error approach to configure model architecture, optimization, and initialization, which instantly inflates the training cost and complicates the workflow. Meanwhile, the effectiveness of these techniques offers a unique opportunity to advance deep learning—existing configurations are empirical summarizations of what works and what does not work. Therefore, analyzing those configurations not only can reduce unnecessary computation, but also can develop perspectives and theories to understand deep learning better.

Two types of tuning are required for training, i.e., optimization tuning and architecture tuning. It is important to understand their underlying mechanisms before simplifying them. To this end, my work effectively integrates theoretical analyses with controlled experiments, by which my study gains a unique edge to improve both productivity and effectiveness.

Optimization Analyses. Learning rate warmup has achieved remarkable success in stabilizing training. Unlike the common wisdom that starts training from a large learning rate, warmup starts from a small learning rate. With systematic analyses, I recognize a long-overlooked issue for optimization, i.e., the adaptive learning rate has an undesirably large variance in the early stage of model training, due to the limited training samples [13]. Inspired by my analyses, I propose a new optimizer to rectify the adaptive learning rate, which effectively stabilizes model training and reduces the need to tune warmup settings.

[1] Citations reference the publications listed in my curriculum vitae.

Similarly, I conduct systematic analyses on robustness overfitting, i.e., if training is conducted for too long, adversarial training leads to a robustness drop. I recognize the root cause of this problem is the optimization setup for the perturbation generator, which gradually downgrades the generator and causes the overfitting [1]. Guided by those analyses, I propose an adaptive adversarial training algorithm to neutralize the robustness drop. It performs better than PGD-10 while reducing the training cost by 75%.

Architecture Analyses. Pursuing factors that can make architectures more stable, I conduct systematic comparisons on two Transformer variants [8]. My analyses show that gradient vanishing is not the direct reason complicating Transformer training. Moreover, I identify an amplification effect that greatly influences training stability and proposes a novel initialization schema to fuse the merits of two different architectures. Without introducing any hyper-parameters, it stabilizes the training of very deep Transformer models and reaches the new state of the art on the long-standing benchmark WMT14'En-Fr.

Besides, we develop an adaptive algorithm to gradually increase the model depth as training advances, accelerating the training by about one time [9]. We further build connections between residual networks and numerical analysis, which sheds insights on why deeper models usually perform better. Also, it allows our method to be adapted to other tasks with little tuning.

FUTURE RESEARCH AGENDA

One crucial method of science is utilizing scientific reasoning over observations. As the volume and the complexity of quantified observations have reached an unprecedented level, how to make sense of them becomes a growing problem. To this end, deep learning methods stand out with their unique abilities to process complicated data. Yet, their excessive resource consumption has hindered various research and application developments, due to the inherent resource limitation of real-world problems. Thus, I believe both of my research directions have significant potential—they are not only immediately needed but can open new possibilities for various disciplines.

Broaden Supervision Sources. I aim to extend the scope of learning systems to include data acquisition and supervision acquisition. For example, I plan to integrate data augmentation with the resource-label conversion. Since both processes can effectively incorporate additional information and reduce the reliance on human annotations, combining them allows them to enhance each other and further minimize the reliance on human annotations. Eventually, this process can be further integrated with active learning and lifelong learning.

Simplify the Training Workflow. To further simplify the training workflow, I aim to explore the assumptions and principles behind successful heuristics. For example, it has been observed that RoBERTa training converges faster with a larger batch size. This phenomenon contradicts the common wisdom that small batches would converge faster, which indicates that existing stochastic optimization algorithms are sub-optimal. Therefore, identifying its underlying pattern can open new possibilities to further accelerate model training.

Longtime Goal and Collaboration. My research enthusiasm is driven most by the unique opportunity to probe human learning through developing machine learning models. Towards this goal, I will engage with scholars across disciplines, and I believe integration among different domains would allow them to enhance each other. Also, since improving productivity requires not only practical effectiveness but also theoretical analysis, my work has already positioned me to integrate empirical and theoretical perspectives of deep learning, and I look forward to continuing these integrations. During my Ph.D. study, my research was supported by NSF, ARL, and Microsoft Research. In the future, I will keep seeking collaborations and funding opportunities from multiple funding agencies and industries.