Research Statement

Tu Vu, University of Massachusetts Amherst https://cs.umass.edu/~tuvu

I develop deep learning algorithms for *resource-constrained* natural language processing (NLP), which involves settings with limited labeled data and/or computational resources. In many real-world applications and scenarios, it is expensive or inconvenient to obtain labeled data (*e.g., translating or summarizing text written in indigenous languages).* During my internships at Google, I have had the privilege to access enormous language models (LMs) and cutting-edge computational resources to conduct research. However, I quickly recognized the unequal distribution of resource power across the world: only a small handful of large and resource-rich corporations can train LMs with billions of parameters. These experiences have motivated me to work on methods for learning with less data and computation to advance and democratize artificial intelligence.

I have made progress on this research vision by employing techniques such as *unsupervised learning*, *semi-supervised learning*, and *transfer learning* that leverage large amounts of unlabeled data or beneficial relationships among tasks to reduce the need for labeled data. Another line of my work aims to *facilitate transfer learning* and *democratize large language models* to the broad research community. I have been working on *parameter-efficient* transfer learning where a single "frozen" pre-trained LM is reused for all tasks, which eases the burden of storing and serving multiple task-specific models. A hallmark of my research is to tackle research problems through *extensive empirical studies* that involve a large number of datasets and experiments. My future research plan is to continue these research directions and seek to advance NLP through *large-scale multi-task* learning from *multilingual* and *multimodal* data.

Motivation & Overview

Large language models show room for improvement: Large LMs (i.e., deep neural networks with millions or billions of parameters pre-trained on large amounts of unlabeled data) have been hailed as one of the most transformative breakthroughs in NLP over the last five years. Scaling up the size of LMs has been shown to not only confer improved performance and sample efficiency (Devlin et al., 2019; Raffel et al., 2020) but also unlock new capabilities (Chowdhery et al., 2022), and thus this trend has continued to push the boundaries of possibility in NLP. However, these models usually suffer from several common weaknesses: *First*, while a few large LMs exhibit impressive few-shot learning ability via prompting—the ability to perform a novel task from only a natural language instruction and/or a few exemplars without any fine-tuning (further training), there is still a large gap between learning with a few demonstrations and with thousands or tens of thousands of labeled examples. In contrast, learning from very little supervision is a hallmark of human intelligence (e.g., a toddler can generalize the concept of "cars" from just a few demonstrations in a book). *Second*, a major limitation associated with large LMs, particularly in low-resource applications, is that it requires significant computational resources for fine-tuning and inference at their scale.

Better learning with limited supervision: Motivated by these limitations, *I work on methods that attain stronger performance with less supervision.* Low-data tasks often occur in real-world applications due to a lack of assets to obtain more labeled data (e.g., in the biomedical domain). It has long been known that transferring knowledge across tasks and domains can confer benefits in such scenarios. However, transfer learning can also hurt performance and factors that determine successful transfer still remain murky, which makes it difficult to exploit synergies between tasks. My EMNLP 2020 paper (Vu et al., 2020) was pioneering work in the exploration and prediction of task transferability in the era of large LMs, which shed light on conditions under which tasks can benefit each other. This work has been highly cited, discussed in different NLP courses,[1] and has influenced an active research area.[2] Additionally, motivated by the findings from this work, I developed novel data augmentation techniques (Vu et al., 2021; Mekala et al., 2022) which can be combined with semi-supervised learning methods (e.g., self-training) to improve few-shot performance.

The appeal of parameter-efficient transfer learning: At Google, I have had the good fortune to access enormous LMs (up to 540B parameters) and cutting-edge computational resources. However, I quickly realized unequal opportunities for practical applicability of these models. Though recent releases of 175B+ parameter models go some way towards democratization, it is still both expensive and inconvenient for small organizations to run models of this scale. *Thus, I have been pursuing parameter-efficient methods (Vu et al., 2022b,a) that facilitate transfer learning and make powerful LMs more accessible to the broad research community.* My work on SPoT (Vu et al., 2022b) proposed an approach that reuses a single "frozen" LM to perform all tasks while only learning minimal task-specific parameters to represent tasks and transfer knowledge between them. Strikingly, across model sizes, SPoT matches or outperforms fine-tuning the whole model (standard fine-tuning) on each task, while being more parameter-efficient (up to 27,000× fewer task-specific parameters). The technique has been adopted for serving recent large language models within Google. This work has also accumulated nearly 100 citations in only several months since its publication (ACL 2022).

Prior & Ongoing Work

In this section, I elaborate on threads of my work to date that form my vision as a researcher.

Task transferability: Unsupervised pre-training followed by supervised fine-tuning has been the dominant paradigm for developing NLP models. However, the success of supervised pre-training on ImageNet in computer vision (Deng et al., 2009; Russakovsky et al., 2015) hints that supervised pre-training tasks can also be beneficial. *Can further fine-tuning LMs on other supervised source tasks before fine-tuning on the target task improves performance?* While this has been shown to be helpful in prior work (Phang et al., 2019), the conditions for successful transfer remain opaque. Until my work on task transferability (Vu et al., 2020), it has been largely restricted to the use of data-rich source tasks of the same type as the target task. To shed light on what factors influence task transferability, I conducted large-scale transfer learning experiments, which involved 3000+ combinations of source-target tasks and data regimes. I demonstrated conditions under which tasks can benefit each other, of which some were unintuitive or even contrary to prior belief; for example, gains are possible even when the source dataset is small, or the source and target tasks are on the surface very different. Interestingly, factors other than data size, such as task and domain similarity, matter more in low-data regimes. Motivated by these findings, I developed effective task embedding methods (which represent tasks as real-valued vectors) to predict transferability between tasks (by measuring similarity between task vectors). My work has influenced a line of recent important work on large-scale multi-task learning, including T0 (Sanh et al., 2022) and ExT5 (Aribandi et al., 2022), and paved the way for an active research area.[2] In an ongoing project, I am investigating whether transfer learning can confer benefits on other NLP applications, such as text evaluation.

Prompt-based learning: The sheer size of large LMs presents a challenge for their practical application. For a 100B+ parameter model, fine-tuning and deploying a separate instance of the model for each task is prohibitively expensive. To get around the infeasibility of fine-tuning, Brown et al. (2020) introduce "in-context learning" where a "frozen" GPT-3 model is conditioned on a manual text prompt to perform a task. GPT-3's performance, however, still lags far behind state-of-the-art fine-tuned models. Lester et al. (2021) propose "prompt tuning", which learns a

[1] For example, COMP790-101 at UNC Chapel Hill and COMPSCI 685 at UMass Amherst.

[2] https://tl4nlp.github.io

task-specific soft prompt (a sequence of tunable tokens prepended to each training example) to condition a frozen model instead. This approach can only match the performance of standard fine-tuning at scale (when a 11B parameter model is used). My work on SPoT (Vu et al., 2022b) proposes to pre-train a soft prompt on one or more source tasks before tuning it on a target task. Remarkably, across model sizes, SPoT matches or outperforms standard fine-tuning while using only 0.01% or less task-specific parameters. Furthermore, I developed an efficient retrieval approach that uses task prompts to represent tasks and retrieves the best source tasks to transfer onto a novel target task. More importantly, SPoT enables a novel transfer learning paradigm where one can reuse a single frozen model and leverage small soft prompts to represent and transfer knowledge between tasks. This work has motivated a large body of work on parameter-efficient learning methods (Liu et al., 2022; Asai et al., 2022). Given the computational cost of fine-tuning, moving to transfer learning with frozen LMs is appealing. In ongoing work, I am also developing methods to keep these models up-to-date and grounded to factual knowledge without retraining them.

Semi-supervised learning and data augmentation: Another subset of my work focuses on methods that use synthetic and/or pseudo-labeled training data to improve performance. Traditional data augmentation approaches are typically restricted to synthesizing training data for a specific target task. To take advantage of task transferability, I departed from this practice and designed novel "task augmentation" techniques that synthesize large amounts of pseudo-labeled data for auxiliary tasks, such as natural language inference (Vu et al., 2021) or question answering (Mekala et al., 2022), in the domain of the target task.[3] The auxiliary tasks here are those existing tasks with high transferability to the target task and with large training datasets, which allows us to train reliable data generative models. Additionally, my work on STraTA (Vu et al., 2021) studies a traditional semi-supervised learning algorithm, i.e., self-training, and highlights important ingredients for its success in the presence of large and powerful LMs (including training on a broad distribution of pseudo-labeled data). By combining task augmentation and self-training, STraTA achieves state-of-the-art performance across many few-shot benchmarks. I hope this work will enable the wider adoption of generative data augmentation and self-training in NLP (media coverage).

Cross-lingual transfer: *Can current transfer learning methods extend successfully to a zero-shot cross-lingual transfer learning setting, i.e., performing a task in a language when labeled data is only available in another language?* In my recent work (Vu et al., 2022a), I conducted the first large-scale empirical investigation of parameter-efficient and standard transfer learning approaches for zero-shot cross-lingual generation (XGen), using summarization as a case study. I found that these approaches struggle on XGen, as a generative multilingual model fine-tuned purely on English catastrophically forgets how to generate non-English. Strikingly, parameter-efficient tuning can provide gains over standard fine-tuning when transferring between less related languages (e.g., from English to Thai). I demonstrated that increasing model scale and decreasing trainable parameter capacity are key for overcoming catastrophic forgetting. Furthermore, I developed methods that explicitly tackle catastrophic forgetting to obtain further gains, including: (1) mixing in unlabeled multilingual data during learning the task, and (2) explicitly factoring soft prompts into "task" and "language" components that can be recombined in novel pairings during inference.[4] My work suggests that robust cross-lingual transfer is within reach and opens up avenues for new research.

Future Work

As the development of ever larger LMs continues, applying these models to NLP applications effectively and efficiently can be more challenging, with existing challenges exacerbated and new

[3] For example, "premise: His acting was really awful. hypothesis: He gave an incredible performance." is an example of the natural language inference task in the movie reviews domain, where the hypothesis contradicts the premise.

[4] To perform summarization in other languages during inference, we simply swap the language components.

challenges emerging. In the next few years, I plan to build on my prior work on *better learning with limited supervision* and *efficient transfer learning* to address these challenges. My long-term vision is to leverage *large-scale* Multi3: **Multi**-task learning from **Multi**lingual and **Multi**modal data.

Better learning with limited supervision: There has been a great deal of effort devoted to facilitating easy use and distribution of pre-trained models and datasets across modalities and languages,[5] which can unlock a multitude of opportunities for large-scale multi-task learning. My work on task transferability (Vu et al., 2020, 2022b) and cross-lingual transfer (Vu et al., 2022a) can serve as a starting point for the following directions:

- Multi3 **pre-training**: Unlike traditional NLP systems, humans typically learn about a new concept through interaction in a real-life environment, which is inherently multi-task oriented and involves learning from multimodal signals. For example, the concept of "motorcycles" formed by humans is also informed by what they look like, what sound they make. I believe NLP models can benefit significantly from multi-task pre-training on multimodal data available across languages. Advanced NLP models should be able to synthesize knowledge from all available modalities during pre-training to solve novel language tasks with limited supervision. For example, information learned from images during pre-training (e.g., *images of cars running on the road in different U.S. states*) should be utilized in answering natural language questions (e.g., *do Americans generally drive on the left or right side of the road?*).

- Multi3 **transfer learning**: With vastly differing amounts of data (both labeled and unlabeled) available across tasks/domains/languages/modalities, there is significant value to developing techniques that can transfer knowledge from higher-resource regimes to improve performance in lower-resource regimes, as suggested by Vu et al. (2020). *Can we devise methods to identify the best existing tasks across modalities and languages to transfer onto a novel language task?*

Efficient transfer learning: I plan to continue my journey to facilitate transfer learning and democratize large LMs to the broad research community (Longpre et al., 2023). I am excited about the following directions:

- **Prompt-based learning with a single frozen** Multi3 **model**: Moving to frozen Multi3 models is appealing with regard to storage and serving costs (Vu et al., 2022b,a). *Can we use a single frozen* Multi3 *model to perform new language tasks based on natural language instructions and/or a few demonstrations only?* This can be seen as an extension of GPT-3 (Brown et al., 2020) to a multilingual and multimodal setting.

- **Parameter-efficient transfer learning**: *Can we develop lightweight modules, e.g., soft prompts, to quickly query and retrieve relevant tasks across modalities and languages and then combine them to solve a novel task with a single frozen* Multi3 *model?* This basically extends my work on SPoT (Vu et al., 2022b) to a multilingual and multimodal setting.

Methodology: I intend to pursue research and innovations that involve *model architecture* (how to represent Multi3 data effectively so one can associate concepts across languages and modalities), *model pre-training* (pre-training objectives to incorporate training signals that span across languages and modalities), *transfer approaches* (how to transfer knowledge from one or more source languages/modalities onto one target language/modality effectively and efficiently).

Taken as a whole, I hope that my future research will have broad impacts that can change the *status quo* not only for NLP, but also for other related fields. Finally, to implement the above interdisciplinary ideas, I hope to collaborate with researchers across fields, including NLP, linguistics, and computer vision.

[5] For example, as of writing, Hugging Face is hosting a plethora of community-contributed resources, with 200K+ models and 34K+ datasets.

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. 2022.

Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. Attentional mixtures of soft prompt tuning for parameter-efficient multi-task knowledge sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. 2022.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. Leveraging qa datasets to improve generative data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. 2022.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022a.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022b.