Research Statement

From Parameter Identification *to* Introspection, Constrained Inference, & Updatability: Toward An Actionable Understanding of Neural Networks

December 2020

Allen Schmaltz
Harvard University

In recent years, high-parameter neural networks have exhibited increasingly strong empirical effectiveness across a wide range of prediction tasks. In effect, machine learning has rapidly transitioned from developing per-task hand-annotated features to engineering particular neural architectures, and learned models have started to approach the point at which effectiveness is practical for tasks traditionally considered higher-risk, such as medicine. However, the high-parameter, deeply compositional nature of the successful neural models precludes interpretation via traditional notions of parameter identification, as used with low parameter, few variable linear models. When treated as a black box, the predictions from such neural models are thus difficult to understand when analyzing the parameters in isolation or only considering the output network activations, as via gradient-based or attention-based approaches. Additionally, using a larger network to train a parsimonious, interpretable linear model may be possible for some closed domains, but this may be challenging across all application areas—particularly those in which the input has complex, long-distance dependencies—while maintaining the effectiveness of the original model.

As such, with expressive, high-parameter neural networks, a fundamentally different approach is needed for understanding the models in real-world settings. In a series of papers (Schmaltz, 2019; Schmaltz and Beam, 2020b,a), I have introduced and analyzed an alternative paradigm, which hinges on deriving distilled, dense representations ("exemplars") from parametric networks that can then be used for non-parametric matching. The resulting mappings between representations from training and test end up having properties useful for understanding neural models and data: We can use them as a guide for constraining predictions (for example, only allowing predictions for which the relative distances to true positive training associations are close); we can use them to introspect test-time predictions relative to the training data, including as an aid for uncovering problems with the underlying data; and in a limited sense, we can also use them to "update" a model by modifying the label associated with a particular mapping. In this way, we move from one-off predictions from static networks to an active process of analyzing the distilled representations of the model and data, constraining predictions, and updating the mappings, potentially queuing a full re-training, if necessary, in the background. The challenging, non-admitted predictions can be passed to humans for adjudication. I refer to this particular paradigm of non-parametric memory matching, derived from deep (parametric) neural networks, as *exemplar auditing*.

I have demonstrated this general approach on a variety of NLP tasks, at varying label granularities. Determining where to cut the parametric network to form the distilled representations is paramount. In classification settings, I have demonstrated an attention-style mechanism that can be used to identify exemplar vectors that serve to concisely summarize the representations of the much larger base network. With this approach, it is relatively straightforward to encourage a prior on the activations for classification, as for example when there is some token-level labeled data; when we want to assign one token to represent the global prediction; and/or when we aim to discourage certain spurious features, or annotation artifacts. We can then match these dense representations to analogous representations created over the training set to assess the reliability of the predictions (under the model, relative to the training data); as a means of uncovering dataset artifacts or mistakes; and/or to update the model by establishing new mappings by running over new data, or

updating existing mappings.

I have also extended this idea of non-parametric memory matching to the more general case of retrieval-classification tasks. In this way, a single end-to-end model can be used for tasks that require retrieving multiple sequences, over which a prediction is subsequently made.[1] To do so, I recast these types of retrieval-classification tasks as a search over dense representations (over bi- and cross- encoded sequences), out of which falls out a means of creating representative exemplar vectors across sequences, which can then be used for matching, as in the basic classification setting.

The combination of these approaches yields a flexible set of tools for language and sequence modeling. The model parameters remain non-identifiable, but in this way we can leverage the pattern-matching and compositional strengths of a deep model to introspect the model and data, constraining predictions and performing lightweight updates, as needed. In particular, we have two mechanisms with which we can update a model: We can update a model via exemplar auditing, and for retrieval-classification tasks, we can also update the explicit retrieval datastore. Additionally, we can constrain predictions via distances to the exemplars, as well as distances to the dense representations of the retrieval datastore. We can impose priors to aid interpretability for particular tasks, and we can visualize the token-level prediction contributions and the sequence-level alignments from retrieval, which are useful approaches for debugging, as well as for uncovering patterns at finer-grained resolutions than the available human-annotated labels.

In summary, we now have a number of levers with which to analyze and update a neural model and its associated data, across label resolutions, putting the field on a practical, productive path toward understanding and constraining the behavior of the increasingly large, and/or expressive, models expected in coming years.

[1]The retrieval-classification setting encompasses a number of existing real-world tasks, and re-casting tasks in this manner enables offloading some of the capacity of the model to explicit datastores. Additionally, as generative models become increasingly effective—creating text, images, and videos indistinguishable at the surface level from naturally occurring counterparts—we will need to rely on retrieval to verify the veracity and integrity of data. In this way, synthetic data detection will likely have to move from the identification of surface-level distributional differences to content-level interrogation, and I have proposed one such approach for addressing the latter.

Future Work This line of work lends itself to a number of clear next directions in computer science in the areas of machine learning and AI, as well as re-examining applications in fields such as public policy, public health, and medicine, where traditionally neural networks have been avoided due to challenges in interpretability.

In the near term, I aim to extend the retrieval-classification matching approach to language generation (i.e., classification over a discrete, but very large, label set), and to assess robustness across different input modalities. More generally, applying this set of approaches to additional types of sequence inputs (e.g., proteins or programming languages), structured sequences (e.g., knowledge bases), and altogether different input modalities (e.g., images or videos) will necessitate innovations addressing the efficiency of the underlying parametric networks.

I plan to further explore the utility of this approach as an aid for labeling new—and correcting existing—training data, in cases where the data is otherwise expensive and/or challenging to label, and related active learning settings.

Further down the line, this paradigm could also be used in a life-long learning framework, whereby the model would dynamically update itself by modifying the database of exemplars with new instances seen in the wild that have low distances to existing items in the database, potentially forcing a re-training or fine-tuning of itself on collected exemplars that have been seen sufficient numbers of times, with sufficiently close distances, over time.

References

Allen Schmaltz. 2019. Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition. arXiv.

Allen Schmaltz and Andrew Beam. 2020a. Coarse-to-Fine Memory Matching for Joint Retrieval and Classification. arXiv.

Allen Schmaltz and Andrew Beam. 2020b. Exemplar Auditing for Multi-Label Biomedical Text Classification. arXiv.