Resolute Resolutions (a blog), volume 3:

Practical, Real-World Neural Network Interpretability and Deployment: Decomposable Model Approximations for Data-Mediated AI (DMA^2)

December 2021

Allen Schmaltz

Abstract

There are compelling practical reasons to view neural network interpretability as an interactive, human-in-the-loop prediction task at a lower resolution of the input than that for which we typically initially have labels. In this context, we will then aim to move to an *abstain+update/adapt* paradigm in real deployments for the large deep networks. To do so, we will ideally need some properties and behaviors that are not typically associated with the deep networks out-of-the-box: We need some means of analyzing the data under the model, relative to the model's predictions for a given instance; we are going to have to address the domain-shift and uncertainty/reliability issues; we need to relate the global instance-level predictions down to constituent parts (and vice-versa), with flexibility in the approach to be adaptable to various prior information we may have; and we seek some degree of updatability when things (inevitably) go wrong with the model or data without having to re-train the full model.

In this blog post, as a brief overview of our existing work, we motivate these characteristics and describe a practical approach for achieving them via model approximations that decompose the deep networks across their input and across their training sets, using dense representation matching as the bridge. We further introduce the term **Decomposable Model Approximations for Data-Mediated AI (DMA^2)** to encapsulate these ideas.

At a coarse resolution, a large part of the challenge of getting to a realistically viable approach for neural network interpretability—and by extension, deployment—was pinning down exactly what the ideal characteristics for interpretability would even be for the deep networks.

The real-world challenges are daunting. The training sets are massive, and often of unknown quality for training, and potentially unknown distribution at inference. The parameters of the deep networks are non-identifiable, and the annotation schemes backed-out indirectly via attention-style mechanisms can themselves also be non-identifiable, as multiple local-level labeling schemes could be consistent with the global-level labels. The models are effective on in-domain data, but they can unexpectedly go off the rails on domain-shifted data, and it is not clear how to get reasonable measures of reliability/uncertainty in such settings. Further, the models are of a scale that preclude rapid re-training or local modification, so when we find problems, we have limited options to course-correct the models. Ideally we would like to address all of these coupled issues, even though they each seem independently very challenging.

In summary, we seek the following characteristics, in no particular order of precedence as we need them all in practice:

- Class-conditional feature detection, relating global-level predictions to local-level predictions (and vice-versa)
- An explicit connection to the training data for a given prediction, at a given unit of analysis
- Out-of-domain detection and uncertainty estimates
- A degree of updatability/adaptability (without re-training the full network)

We will achieve these characteristics in practice by breaking a given model down to a desired lower unit of analysis, and then approximating those predictions as a weighted vote over the training set. We show this in practice in a series of papers starting with Schmaltz (2021).

In this blog post, we further introduce the new term **Decomposable Model Approximations for Data-Mediated AI (DMA[2])** to encapsulate this particular framework for AI, as an aid to help structure future research endeavors. This will also help to differentiate our proposed approach from alternative approaches, which differ not only on a technical level, but also on foundational epistemological levels. We first motivate the challenges of interpreting and deploying neural networks, and then we briefly review the approach proposed in our existing work. This is the first in a series of posts on this topic.[1]

The Datasets Pose Serious Challenges

Typically when we use the large deep networks it is in the context of massive training datasets, which may have been curated by means opaque to the model designers. This is particularly true for data used for pre-training the networks. To reach scale, such data may be derived via semi-automated means and only lightly curated by annotators. As such, a core theme of our proposed approach for interpretability is that we will want methods for analyzing the data itself, and in particular, relative to a model's prediction for a given instance, at a given unit of analysis.

The other important point on the data front is that often when we think of deploying the deep networks, it is in the context of some type of interactive, dynamic end-user application where we have new data coming in that has the potential to differ rather significantly from what was seen in training. This is referred to as the out-of-domain (OOD) and/or domain/sub-population shift problem[2]. This is an issue of concern since the deep networks tend to be sensitive to relatively modest perturbations in the input data (i.e., they are "brittle").

In summary, when it comes to the data, we face challenges both in terms of knowing what exactly is in the tails of the training data, and in handling input data at inference in the tails of the distribution that could throw the whole endeavor unexpectedly off the rails.

The Models *also* Pose Serious Challenges

Now we know the data itself is going to pose some serious challenges in this context. Unfortunately, it gets even worse from there, since the models themselves pose some intrinsic challenges, which will require some conceptual shifts in how we think about interpretability[3]. In the context of document-level classification, one means of making sense of a document-level prediction is to use some type of learned mechanism of the network from which we can back out predictions/activations for lower-levels of the input (such as words) in an indirect manner. These mechanisms fall under the general heading of "attention mechanisms".[4] These approaches can be useful when the goal is to create a sequence labeler when we do not have word-level labels for training. Provided the sequence labeling effectiveness is good, this already provides a means of analyzing the data (namely, simply run it over your data and visualize the activations).

However, there are some subtle caveats to remember when thinking about these mechanisms in terms of the document-level predictions. The parameters of the deep network are not identifiable, and as a reflection, it is possible to bias the attention mechanisms to have rather different word-level predictions while the document-level predictions nonetheless stay similar. That we now know how to structure the networks with suitable inductive biases/priors (and losses) to create such indirect sequence labelers is good news for practical applications, but the nature of the problem is such that we will have to think carefully about the

[1] Note the new terminology introduced here is currently local to this series of blog posts, but eventually we plan to transition to using it in our future research papers, as well.

[2] The gradation from domain-shifted samples to OOD samples is not well-defined, but in any case, our desired behavior is the same across this spectrum: In real applications, we will aim to abstain from predicting—or similarly, maintain a coverage proportion—over uncertain/unreliable inputs.

[3] Note that we do not try to separate "interpretability" from "explainability". There is not a well-agreed upon distinction between those two terms on a technical level in Computer Science, in part because it has been unclear what exactly our ideal target behavior even really should be with the deep networks.

[4] The original motivation for attention mechanisms in deep learning was not interpretability, per se, but rather to improve overall prediction effectiveness, initially for machine translation. We get the interpretability behavior as a side-effect, and as we show in our work, we can structure and optimize these mechanisms for predicting labels at lower resolutions than our available labels.

inductive biases of both the architecture and the losses for each task and dataset, deciding how the local level relates to the global level, and vice-versa. However, even if our particular attention mechanism produces a strong predictor for our held-out data, there is a sense in which we have just kicked the can down to a lower level: We have turned our uninterpretable document-level prediction into an uninterpretable word-level prediction, and we have not done anything about the domain-shift/reliability issue. As such, we ideally want some type of additional mechanism to relate the test predictions to those for which we have known labels (e.g., those from the original training set), and to have some means of constraining the predictions to at least be similar to what we have observed in training. Finally, if we are analyzing the data, or constraining the predictions, and we find that something is indeed wrong in our training data (e.g., an incorrect label) or to a limited extent, we want to add new data to the training set (for new domains), is our only recourse to retrain the massive model we took weeks and large compute to train? What if an end user wants to make a local update (changing labels, etc.)?

A Partial Solution is Not that Far from No Solution

In short, we have a number of hurdles to overcome in order to have a reasonable handle over interpreting the networks so that we can deploy them in practice. We need some means of analyzing the data under the model. We are going to have to address the OOD and uncertainty/reliability issues to have some sense of not only when the predictions may go off the rails, but also importantly, when our interpretability methods themselves may be going off the rails, if interpretability is itself viewed as a prediction task. We need some means of relating the global instance-level predictions down to constituent parts (and vice-versa), with flexibility in the approach to be adaptable to various prior[5] information we may have. Ideally, we also seek some degree of updatability when things (inevitably) go wrong with the model or data without having to re-train the full model.

Part of the challenge has been simply asking the right questions to determine what characteristics we even need in practice. The whole endeavor is rather more shaky without any one of the aforementioned characteristics. The partial solutions of previously proposed approaches have weaknesses that manifest themselves in real-world applications. For example, you could have a highly effective indirect sequence labeler (via an attention-style mechanism) on your in-domain data, but if you have not addressed the domain-shift problem, it could all unexpectedly fall apart at deployment on the new data that users submit. As another example, perhaps you have some reasonably effective intrinsic measure of reliability/uncertainty, but you lack an explicit connection to the training data, so you have no direct means to analyze and re-label if it turns out that your massive training set has very poor quality labels for certain subpopulations, cutting the data space differently than that of your pre-deployment trials and analyses.

There may be multiple possible avenues for achieving these desired characteristics with the deep neural networks. Henceforth, we will group frameworks that have the desired characteristics under the heading of **Data-Mediated AI (DMA)**, emphasizing the connection between the data and the predictions. More abstractly, "mediation" (as used in the dispute resolution sense) typically involves a back-and-forth (but nonetheless, structured) negotiation process with a third party. Its use here emphasizes that the very nature of the desired properties implies the need for some type of interactive process against the data to reach a suitable settlement for every prediction.

Decomposable Models and Approximations

In Schmaltz (2021)[6] we obtain the aforementioned properties over the deep networks with a particular attention-style mechanism that gives us flexibility in producing word-level predictions from models trained with document-level labels. By structuring the network in this particular way, we are then able to derive dense representations ("exemplar repre-

[5] In this context, we mean "prior information" broadly construed, as opposed to more specifically a "prior distribution" in the typical Bayesian modeling sense. For example, if we have a sequence labeling task, one example of prior information is that we know ahead of time that only 1 of N words should be labeled for every document, so we might structure the network accordingly with a suitable inductive bias to yield very sparse predictions. In some cases, such information may well be readily translatable into a Bayesian framework, but in others, it may not be so obvious how to do so. We leave the connection to Bayesian modeling to future work.

[6] A a 12 minute video overview is here: https://youtu.be/iJ_udvksyqE.

sentations") for each word-level prediction that we use for matching against the training data, or a support set with known labels. We then construct an approximation of those word-level predictions as a K-NN over the matches and their associated labels. So in summary, at a high level, we decompose the model first across the input, down to a lower resolution than that for which we had annotated labels. Next, we decompose each of those predictions as a weighted vote over the training set.

This will give us a tool for analyzing the data at a lower resolution than our available ground-truth labels, and this approach gives us an explicit connection to the training instances for each prediction. Importantly, it also turns out that the approximation gives us strong signals as to the reliability of the predictions: The predictions become more reliable as the output magnitude of the K-NN increases (recall it is a weighted vote of labels and predictions over the training set) and as the distance to the first match decreases. We can then use that to screen instances unlike those seen in training. Interestingly, to a limited extent, we can then update the model. Provided the representations can match to the new data/distribution, we can simply update the labels or instances in the support set, which serves as an updatable database, in effect side-stepping the challenges of catastrophic forgetting typically associated with adaptive-style fine-tuning of the deep networks. Holding other things constant, we have not gained anything in terms of robustness over domain-shifted/OOD data[7], but importantly, at least we can screen such data, to prevent things from going off the rails, sending the input to humans for adjudication, and then in some cases, we can perform an update via the model approximation.

We will henceforth refer to approaches that break a deep network apart into the relevant constituent features and then approximate those predictions as a weighting over the training set with the general term **Decomposable Model Approximations (DMA)**.

[7] In fact, there do not appear to be any approaches to date that have consistently improved robustness, ceteris paribus, *holding the model and training data fixed.*

Discussion

For interpretability of a deep network's predictions and its underlying training data, there are compelling practical reasons to structure the deep network in such a way as to be able to back-out predictions at a lower resolution than that of the training labels, and then approximate those predictions as a weighted vote over the training set. Dense matching constraints against the observed data yield signals for prediction reliability/uncertainty, and the approximations additionally provide a mechanism for updating the model in some settings without a full re-training of the underlying deep network.

In this way, we can view interpretability as an interactive, human-in-the-loop prediction task at a lower resolution of the input than that for which we have labels. For higher risk settings, we can move to an *abstain+update/adapt* paradigm, wherein we abstain from predicting when the model is uncertain, sending the prediction to a human for further adjudication. In effect, with an approximation at least as effective as the original model, the approximation itself becomes the model we use in practice. We may then be able to update the model (via the weighted vote over the database of exemplar representations and labels) by modifying the existing labels, or by adding altogether new labeled instances to the database.

The Future is Distant, but Probably Near Enough

With Decomposable Model Approximations for Data-Mediated AI (DMA^2), both in the abstract and with our particular implementations, the field has rapidly transitioned from very loose notions of interpreting the current generation of deep networks, to concrete approaches that may well be sufficient for many real-world applications. Importantly, we gain not only a means of constraining the predictions, but also a means of analyzing the data itself.

References

Allen Schmaltz. 2021. Detecting Local Insights from Global Labels: Supervised and Zero-Shot Sequence Labeling via a Convolutional Decomposition. *Computational Linguistics*, pages 1–45.