Introspection, Updatability, and Uncertainty Quantification with Transformers: Concrete Methods for AI Safety

Allen Schmaltz
Reexpress AI
allen@re.express

Danielle Rasooly
Reexpress AI and Harvard University
danielle@re.express

Abstract

When deploying Transformer networks, we seek the ability to INTROSPECT the predictions against instances with known labels; UPDATE the model without a full re-training; and provide reliable UNCERTAINTY quantification over the predictions. We demonstrate that these properties are achievable via recently proposed approaches for approximating deep neural networks with instance-based metric learners, at varying resolutions of the input, and the associated VENN-ADMIT Predictor for constructing prediction sets. We consider a challenging (but non-adversarial) task: Zero-shot sequence labeling (i.e., feature detection) in a low-accuracy, class-imbalanced, covariate-shifted setting while requiring a high confidence level.

1 Introduction

Deep neural networks present researchers, engineers, and society with a curious dichotomy: The blackbox models offer the potential for point prediction accuracies far beyond alternative options, *but* it is difficult to reliably ascertain over what input data those accuracies can be obtained. The non-identifiable [cf., 5, 6] and extraordinarily large number of parameters suggest a lost cause, which has largely precluded the use of deep networks in higher-risk settings, such as medicine. Irreducible error is unavoidable, and data errors (as with noise in the labels) and model errors (as via mis-specification) are expected; we would not reasonably expect reliable predictions from a blackbox over arbitrary inputs without any monitoring over time. However, the challenge has been that in even rather more closed—but nonetheless, realistic—settings, such as non-adversarial settings without concept shift, the errors from deep networks have been difficult to characterize. Without principled and reliable approaches for such basic settings, it is difficult to have trust in the predictions from neural networks, and difficult to have trust in the ability of end-users to interpret and deploy such models.

More specifically, in a typical natural language processing (NLP) classification task, we have access to a computationally expensive blackbox neural model, F; a training dataset, $\mathcal{D}_{\mathrm{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ of $|\mathcal{D}_{\mathrm{tr}}| = I$ instances paired with their corresponding ground-truth discrete labels, $Y_i \in \mathcal{Y} = \{1, \ldots, C\}$; and a held-out labeled calibration dataset, $\mathcal{D}_{\mathrm{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$ of $|\mathcal{D}_{\mathrm{ca}}| = J$ instances. We are then given a new test instance, X_{N+1}, from an unlabeled test set, $\mathcal{D}_{\mathrm{te}}$. In higher-risk settings, it is insufficient to simply rely on the point prediction. We instead will seek to construct a prediction set, produced by some set-valued function $\hat{\mathcal{C}}(X_{N+1}) \in 2^{C}$, containing the true unseen label with a specified coverage level $1 - \alpha \in (0, 1)$ on average. The challenge then becomes defining that average and achieving the resulting coverage in practice, with the additional complication that labels may be limited at the resolution at which we seek to analyze the data.

In this brief work, we highlight the utility of VENN-ADMIT Predictors [12] for deploying deep networks in real-world settings, providing a proof-of-concept for viewing neural network deployment

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

as a controlled human-in-the-loop prediction task. In the interest of brevity, we will largely defer technical details of the KNN model approximations [11] and the VENN-ADMIT Predictor to their original works. The goal of the present work is to briefly synthesize these existing works, illustrating the potential utility for INTROSPECTION, UPDATABILITY, and UNCERTAINTY, key characteristics necessary for engendering trust in neural networks and in the ability of end-users to interact with neural networks. We provide a concrete implementation of these properties with Transformer networks [13] for distantly-supervised sequence labeling tasks by decomposing document-level predictions via a hard-attention mechanism and approximating the model as a weighted KNN, and then estimating the uncertainty in the predictions via a VENN-ADMIT Predictor after labeling a subset of the word-level predictions. This illustrates that the aforementioned properties can be realistically achieved in relatively challenging task settings.

2 Classification with Transformers for NLP

We first introduce general notation for sequence labeling tasks. Each instance consists of a document, $\boldsymbol{x} = x_1, \ldots, x_t, \ldots, x_T$, of T words. In the case of **supervised sequence labeling** (**SSL**), we seek to predict $\hat{\boldsymbol{y}} = \hat{y}_1, \ldots, \hat{y}_t, \ldots, \hat{y}_T$, the word-level labels for each word in the document, and we have the ground-truth word labels, y_t, for training. For **zero-shot sequence labeling** (**ZSL**), we also seek to predict $\hat{\boldsymbol{y}}$, but we only have access to the document-level label, y, for training. The **ZSL** task is a proxy for feature detection for document classification; a dataset is created by removing the word-level labels of a standard sequence-labeled dataset. This enables evaluating feature detection against ground-truth labels.

For each task, our base model is a pre-trained Transformer network. We fine-tune a kernel-width 1 CNN (MEMORY LAYER) over the output representations of the Transformer, producing predictions and representative dense vectors at a resolution (e.g., word-level or document-level) suitable for each task. We then approximate these predictions with weighted KNNs. With these approximations, we then construct the VENN-ADMIT Predictor, as described next.

3 Methods

We briefly review the VENN-ADMIT Predictor (Section 3.1), and then highlight the unusual and useful properties it enables for analyzing and deploying Transformer networks (Sections 3.2 and 3.3).

3.1 Inductive Venn-ADMIT Predictor (IVAP)

An Inductive VENN-ADMIT Predictor is a distribution-free frequentist approach for constructing prediction sets over a Transformer model. It can be viewed as a two-stage process of first constructing split-conformal prediction sets [15, 8] followed by calibration via a Venn Predictor [14] and a probability mass correction to obtain coverage stratified by size. Specifically, a VENN-ADMIT Predictor aims to approximate the following quantity:

$$\mathbb{P}\left\{Y_{N+1} \in \hat{\mathcal{C}}(X_{N+1}) \,|\, X_{N+1} \in \mathcal{B}(x), Y_{N+1} = y, \hat{\mathcal{C}}_{N+1} = \mathcal{A}\right\} \geq 1-\alpha,\ P_X(\mathcal{B}(x)) \geq \xi,\ \mathcal{A} \in 2^C \tag{1}$$

That is to say, when evaluating the prediction sets, we seek for the true value to be contained with a proportion of $1-\alpha$ on average after stratifying by the true label, the data partition $\mathcal{B}$, and the membership of the set itself, which also always includes the top-label prediction. This conditioning is rather more stringent than typically considered in the confidence calibration literature, including class- and predicted-label-conditioned variations [7, 4, inter alia]. Previous work [12] has shown that this quantity can be reliably estimated in practice with the VENN-ADMIT Predictor. It is easy to interpret: The set size corresponds to reliability, on average over similar points; sets of size 1 will tend to obtain coverage, unlike standard split-conformal approaches. The price to pay for this simplicity and reliability is a coarser quantity than a single probability.

An unusual and advantageous aspect of a VENN-ADMIT Predictor, and which further distinguishes it from post-hoc Platt-scaling-style calibration [9, 3], is a degree of inherent example-based interpretability: The calibrated distribution for a point is a simple transformation of the empirical probability among similar points, with partitions determined by a KNN that can be readily inspected. Further, the distributions will tend to become more diffuse as the sample sizes decrease, a desirable property.

Finally, the data partitioning separates the reliable predictions from the less reliable predictions; in addition to providing a constraint for higher-risk settings, this provides a decision rule for active learning, noted below.

3.2 Labeling guidance

Labels can be expensive to obtain, especially in higher-risk settings. The data partitions of a VENN-ADMIT Predictor are determined by the depth of true positive matches into a support set of a weighted KNN that approximates the underlying Transformer. We can exploit this as a rule for labeling: Simply refrain from labeling additional training points once the maximum depth has been reached among the available unlabeled points.

3.3 Introspection+Updatability+Uncertainty

Previous work [11] has shown that, despite only having document-level labels for training, the MEMORY LAYER can be decomposed as a hard-attention-style mechanism, with an inductive bias conducive to relatively high-precision feature detection. We can combine this behavior with a KNN model approximation and the VENN-ADMIT Predictor for a straightforward, principled approach for analyzing, deploying, and monitoring a Transformer network for classification:

1. Pre-train and fine-tune the Transformer.
2. Decompose the document-level predictions to the word-level for interpretability and analysis, using the methods of [11].
3. Label the word-level predictions of a held-out calibration set and those of the support set for the KNN approximation (Section 3.2).
4. Construct prediction sets via the VENN-ADMIT Predictor.

4 Experiments

We examine the key behavior using the data of the GRAMMAR task of [12], which seeks to label whether each word in a sentence has ($y = 1$) or does not have ($y = 0$) a grammatical error. The test set is challenging for two reasons. First, the $y = 1$ class appears with a proportion of 0.07 of all of the words. Second, the in-distribution task, consisting of college essays from second-language learners, itself is relatively challenging, but it is made yet harder by adding newswire text, resulting in low point accuracies for the minority class.

The GRAMMAR task is a fully-supervised sequence labeling task. We compare against a variation, FEATURE, in which the model only has access to sentence-level labels at training. At test time, the model must label at the word level, and thus the task is analogous to feature detection. To construct the prediction sets, the FEATURE model is provided $\leq (K+1) \cdot |\mathcal{D}_{\rm ca}| + K \cdot |\mathcal{D}_{\rm te}|$ word-level labels to construct the model approximations of the VENN-ADMIT Predictor, corresponding to the true word-level labels of the calibration set, $|\mathcal{D}_{\rm ca}|$, and the word-level labels of the nearest K words in the training set for both the calibration set and the test set, $K \cdot (|\mathcal{D}_{\rm ca}| + |\mathcal{D}_{\rm te}|)$, some of which could be duplicates. The word-level calibration labels are necessary to determine the quantile thresholds for the conformal prediction sets and the empirical probabilities of the assigned categories of the Venn Predictor, and the word-level training labels are necessary to determine the data partitions of the VENN-ADMIT Predictor [see 12]. The FEATURE task thus illustrates KNN-based updating of the model without modifying the parameters of the Transformer: We fine-tune the Transformer only with document-level labels. Word-level labels are then only introduced as part of the KNN approximations, which contain a total of 6 real-valued parameters. As a consequence, catastrophic forgetting becomes less of an issue, as the dense representations (each of dimension 1000), let alone the > 300 million parameters of the full Transformer, are never altered. Further, potentially fewer word-level labels are needed overall.

For all experiments, we set $\alpha = 0.05$ and otherwise follow the experimental setup in [12], comparing to the same baselines. As a distribution-free **baseline** of comparison we consider the size- and adaptiveness-optimized RAPS algorithm of [1], $\text{RAPS}_{\text{SIZE}}$ and $\text{RAPS}_{\text{ADAPT}}$, which combine regularization and post-hoc Platt-scaling calibration [9, 3], on the output of the MEMORY LAYER. Using stratification of coverage by cardinality as a metric, RAPS$_{\text{ADAPT}}$, in particular, was reported to more closely approximate conditional coverage than the alternative APS [10], with smaller sets. Conf$_{\text{BASE}}$ is a split-conformal point of reference for simply using the output of the base KNN approximation without further conditioning, nor post-hoc calibration. Local$_{\text{CONF}}$ is a localized conformal baseline [2] using a KNN localizer. Across methods, the point prediction is included in the set, which ensures conservative (but not necessarily exact/upper-bounded) coverage by eliminating null sets. We evaluate coverage, $\overline{y \in \mathcal{C}}$, and cardinality, $\overline{|\mathcal{C}|}$, across training distance, class, and cardinality stratifications.

5 Results

Figures 1 and 2 summarize the key takeaway in this challenging setting: Only the Venn-ADMIT sets (and their first-stage non-mass-corrected ADMIT conformal prediction sets) obtain acceptable coverage for the minority class.

Coverage is obtained for the Venn-ADMIT sets for the Feature task, even though the model is relatively weak, with the memory layer only having been trained with document-level labels. The difference with the stronger fully-supervised model of the Grammar task is reflected in fewer singleton sets over the weaker model. Specifically, among the Venn-ADMIT sets for the Grammar task, there are 4,285 words with singleton sets for class 0 (with a coverage of 1.00) and 381 words with singleton sets for class 1 (coverage of 0.93). In contrast, among the Venn-ADMIT sets for the Feature task, there are only 194 words with singleton sets for class 0 (with a coverage of 0.96) and 111 words with singleton sets for class 1 (coverage of 0.97). Critically, approximate conditional coverage (including size-stratified) is retained, despite the model only having a point accuracy of 0.23 on a minority class that only occurs in 7% of all the words.

6 Discussion

In higher-risk settings, indirect feature detection—as with attention-style mechanisms—is not sufficient alone and labeled data is necessary for trustworthy uncertainty quantification. In comparing the Venn-ADMIT sets for the Grammar and Feature tasks, we have demonstrated the feasibility of initially training a model with document-level labels and then subsequently labeling data at the word level according to the matches of the KNN approximation of the Venn-ADMIT Predictor, after decomposing the document-level predictions to the word level.

Prospectively, as the labeled set is expanded, the Venn-ADMIT sample size checks (the hard threshold, κ, which is 100 here, and the impact of the weighted localizer and test-point augmentation) within each data partition provide additional checks on small sample sizes (high variances). In higher-risk settings, we can further restrict sets to the most reliable data partition.[1] We are not aware of other calibration approaches that have these unusual—but highly advantageous—properties.

7 Conclusion

A Venn-ADMIT Predictor provides more robust approximate conditional coverage over Transformer networks than extant alternatives, and enables a structured approach for labeling data to obtain such coverage. We have briefly summarized the implications of recent works and provided additional results using zero-shot sequence labeling, providing a demonstration of the key properties needed for analyzing, deploying, and monitoring a Transformer network in non-adversarial settings.

[1] For example, restricting to the partition $q = K$, as discussed in [12], only results in rejecting 3 singleton sets for the Feature task and 7 singleton sets for the Grammar task. Labeling budgets can be conserved by not labeling the full KNN support for points outside the top partition, since they will typically not be singleton sets, in any case.

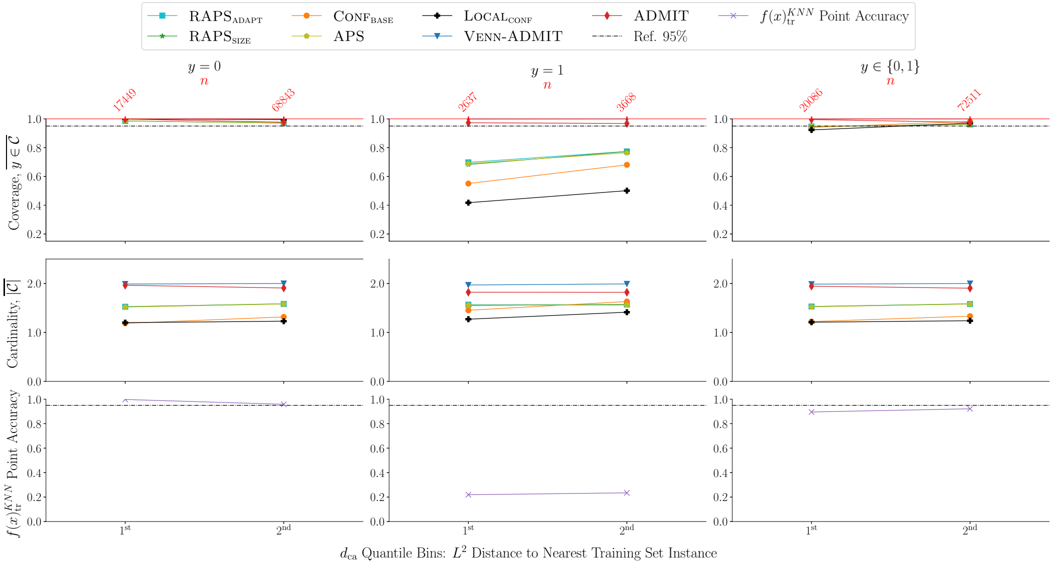

Figure 1: Coverage, cardinality, and point accuracy for the FEATURE task, $\alpha = 0.05$.

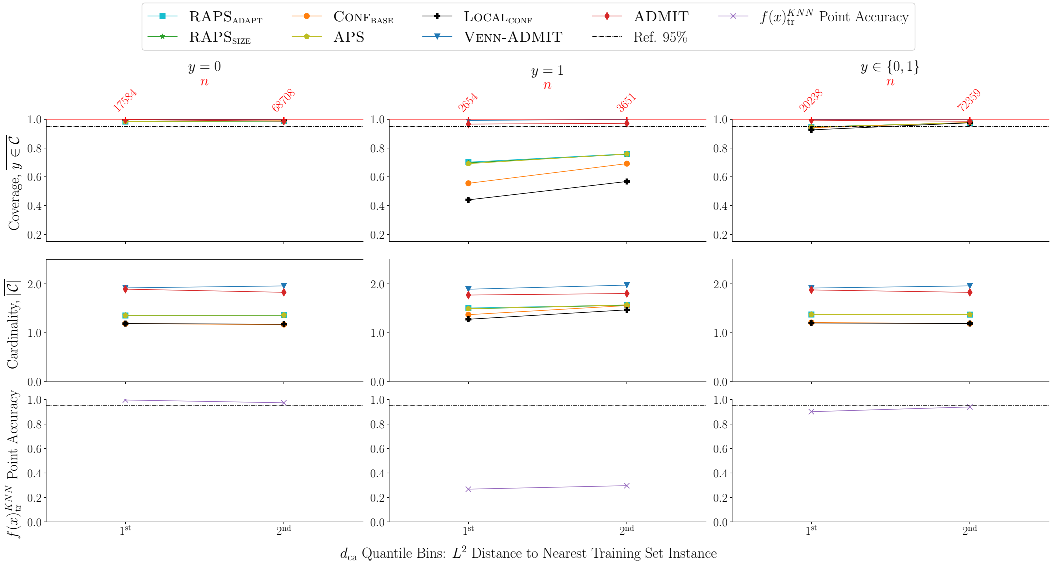

Figure 2: Coverage, cardinality, and point accuracy for the GRAMMAR task, $\alpha = 0.05$.

References

[1] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=eNdiU_DbM9`.

[2] L. Guan. Localized Conformal Prediction: A Generalized Inference Framework for Conformal Prediction. *Biometrika*, 07 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac040. URL `https://doi.org/10.1093/biomet/asac040`. asac040.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330. JMLR.org, 2017.

[4] C. Gupta and A. Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=WqoBaaPHS-`.

[5] J. T. G. Hwang and A. A. Ding. Prediction Intervals for Artificial Neural Networks. *Journal of the American Statistical Association*, 92(438):748–757, 1997. ISSN 01621459. URL `http://www.jstor.org/stable/2965723`.

[6] S. Jain and B. C. Wallace. Attention is not Explanation. *CoRR*, abs/1902.10186, 2019. URL `http://arxiv.org/abs/1902.10186`.

[7] M. Kull, M. Perello-Nieto, M. Kängsepp, T. S. Filho, H. Song, and P. Flach. *Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[8] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML'02, pages 345–356, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540440364. doi: 10.1007/3-540-36755-1_29. URL `https://doi.org/10.1007/3-540-36755-1_29`.

[9] J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[10] Y. Romano, M. Sesia, and E. J. Candès. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[11] A. Schmaltz. Detecting Local Insights from Global Labels: Supervised and Zero-Shot Sequence Labeling via a Convolutional Decomposition. *Computational Linguistics*, 47(4):729–773, Dec. 2021. doi: 10.1162/coli_a_00416. URL `https://aclanthology.org/2021.cl-4.25`.

[12] A. Schmaltz and D. Rasooly. Approximate Conditional Coverage via Neural Model Approximations. *arXiv:2205.14310*, 2022. doi: 10.48550/ARXIV.2205.14310. URL `https://arxiv.org/abs/2205.14310`.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[14] V. Vovk, G. Shafer, and I. Nouretdinov. Self-calibrating Probability Forecasting. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL `https://proceedings.neurips.cc/paper/2003/file/10c66082c124f8afe3df4886f5e516e0-Paper.pdf`.

[15] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.