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

Allen Schmaltz and Danielle Rasooly

re.express

Overview

With Transformer networks, we demonstrate introspection of the predictions against instances with known labels; updatability of the model without a full re-training; and reliable uncertainty quantification over the predictions. This is possible via KNN-based approximations and the associated VENN-ADMIT Predictor.

Background: Prediction sets for classification / Selective classifiers

- Computationally expensive blackbox (Transformer model): F
- Training dataset: $\mathcal{D}_{\mathrm{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ with $Y_i \in \mathcal{Y} = \{1, \ldots, C\}$
- Held-out labeled calibration dataset: $\mathcal{D}_{\mathrm{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$
- Seek: A prediction set $\hat{\mathcal{C}}(X_{N+1}) \in 2^C$ for a new, unseen test instance X_{N+1} from $\mathcal{D}_{\mathrm{te}}$ containing the true label with proportion $1 - \alpha \in (0,1)$ *on average after stratifying by:*
  - *True label*
  - *Data partition $\mathcal{B}$ (determined by distance & relative similarity to training)*
  - *Set membership (including top label prediction)*

  $\implies$ *Singleton set coverage (a.k.a., well-calibrated selective classification), a quantity useful for typical classification settings*

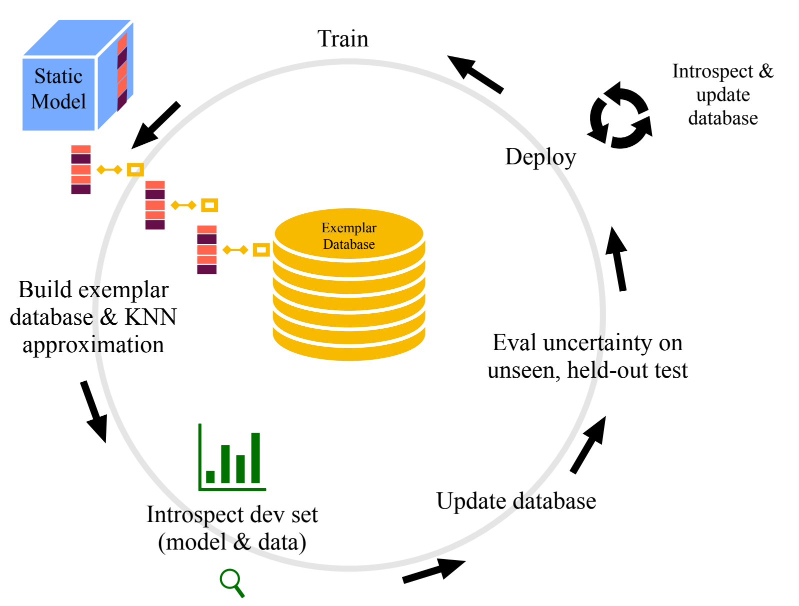
- VENN-ADMIT Predictor: *Approximate conditional* coverage & calibration:

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

Weighted KNN approximations of the deep network encode strong signals for prediction reliability:

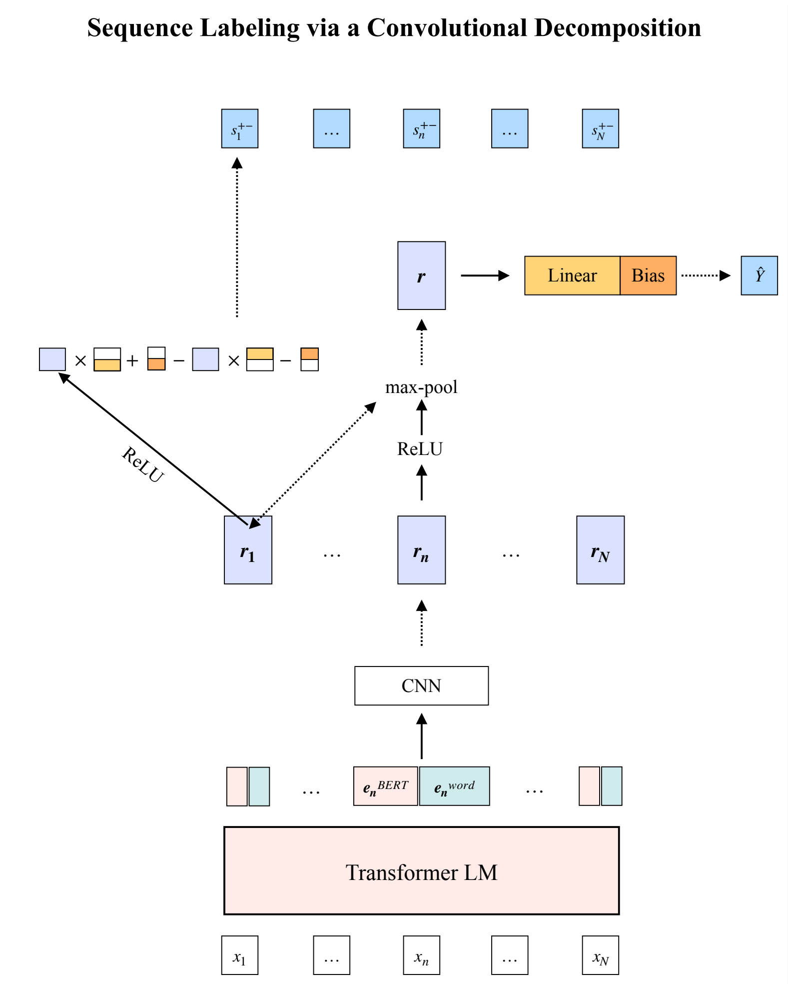
Predictions become less reliable at distances farther from the training set and with increased label and prediction mismatches among the nearest matches.

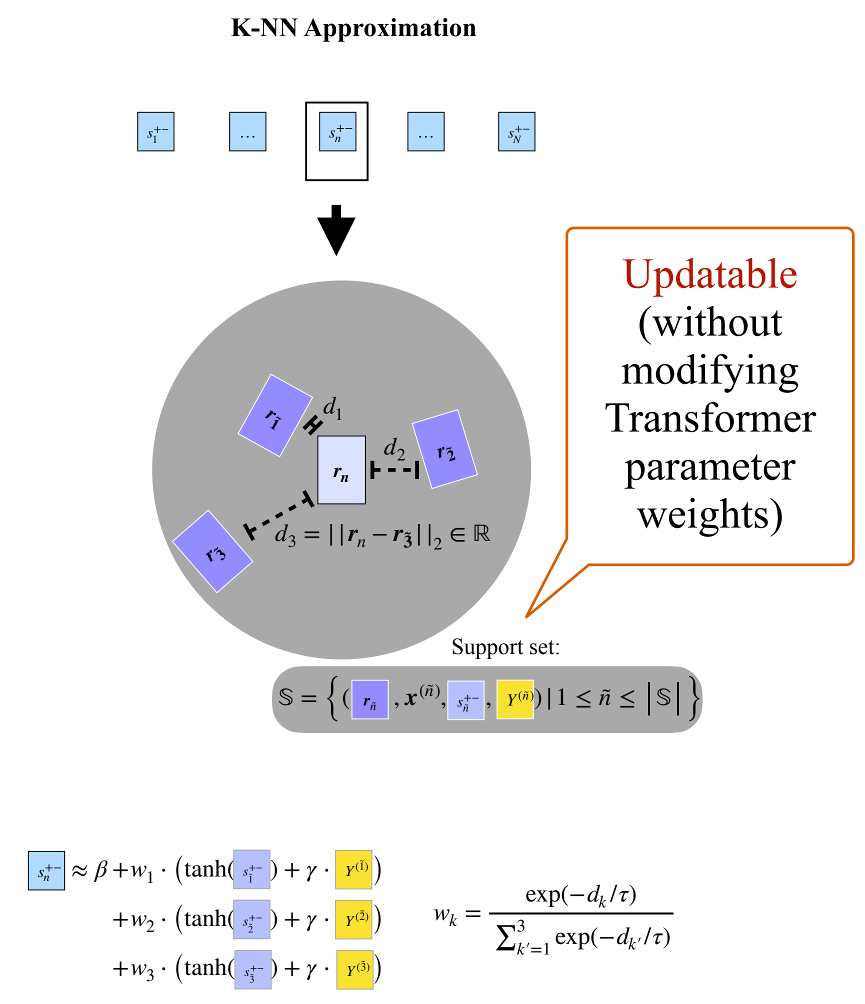
Decompose Transformer into human-understandable parts via instance-based metric learner approximations: Yields properties of Introspection, Updatability, and Uncertainty, with which we can prospectively re-cast neural network interpretability and deployment as a human-in-the-loop prediction task.



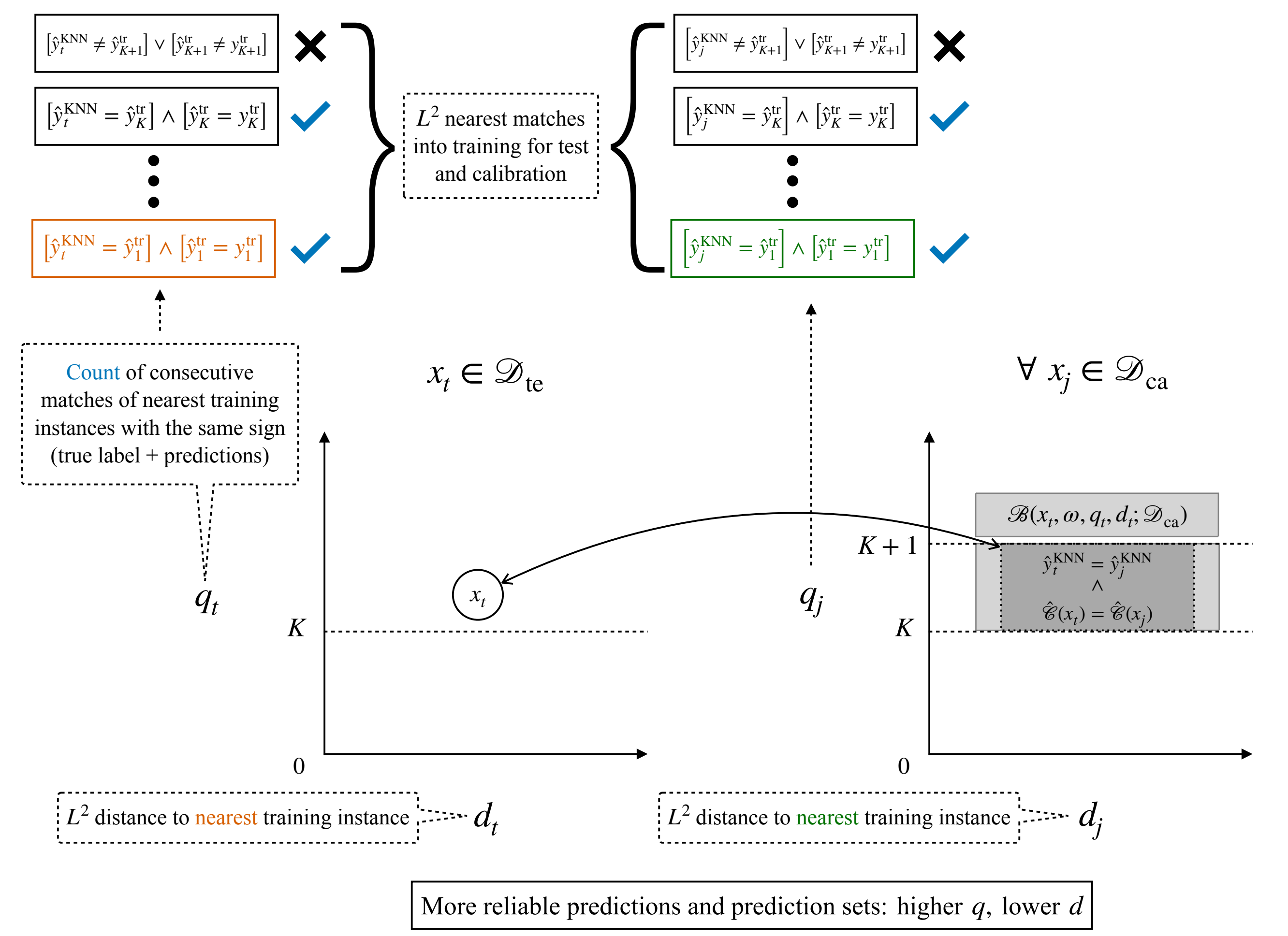
1. Pre-train & fine-tune Transformer using document-level labels
2. Introspect: Decompose the document-level predictions to the word-level for interpretability and analysis
3. Update: Label the word-level predictions of a held-out calibration set and those of the support set for the KNN approximation
4. Quantify uncertainty: Construct prediction sets or selective classifications via the VENN-ADMIT Predictor.
5. Continually monitor and update

Introspection: Decompose prediction via CNN (hard attention) & then approximate with a KNN over the training set





Uncertainty Quantification: A VENN-ADMIT Predictor calibrates the output as the empirical probability of similar points via dense matching



Empirical behavior: Proof-of-concept using zero-shot sequence labeling (i.e., feature detection) in a low-accuracy, class-imbalanced, covariate-shifted setting while requiring a high confidence level

$(1 - \alpha = 0.95,\ N = 93k,\ y \in \{0,1\})$

*Train model with document-level labels & then **update** via KNN with word-level labels*

Method	$y = 0$: $\overline{y \in \mathcal{C}}$	$y = 0$: n/N	$y = 1$: $\overline{y \in \mathcal{C}}$	$y = 1$: n/N
KNN ACC.	0.97	0.93	0.23	0.07
$\mathrm{CONF}_{\mathrm{BASE}}$	1.00	0.66	0.16	0.03
$\mathrm{RAPS}_{\mathrm{ADAPT}}$	0.94	0.40	0.40	0.03
$\mathrm{RAPS}_{\mathrm{SIZE}}$	0.94	0.40	0.40	0.03
APS	0.94	0.40	0.40	0.03
$\mathrm{LOCAL}_{\mathrm{CONF}}$	1.00	0.72	0.17	0.04
→ VENN-ADMIT	0.99	<0.01	1.00	<0.01

Fully-supervised model

Method	$y = 0$: $\overline{y \in \mathcal{C}}$	$y = 0$: n/N	$y = 1$: $\overline{y \in \mathcal{C}}$	$y = 1$: n/N
KNN ACC.	0.98	0.93	0.27	0.07
$\mathrm{CONF}_{\mathrm{BASE}}$	0.99	0.77	0.30	0.04
$\mathrm{RAPS}_{\mathrm{ADAPT}}$	0.98	0.60	0.42	0.03
$\mathrm{RAPS}_{\mathrm{SIZE}}$	0.98	0.60	0.43	0.03
APS	0.98	0.59	0.42	0.03
$\mathrm{LOCAL}_{\mathrm{CONF}}$	1.00	0.77	0.21	0.04
→ VENN-ADMIT	0.99	0.02	0.97	<0.01

→ Well-calibrated selective classification, with a sharpness suitable even for highly imbalanced, low-accuracy settings, with robustness to covariate shifts

→ Prospectively provides safeguard when using fewer labels (and/or weaker models, in general)

- Behavior holds for in-distribution tasks, as well, with majority of points (n/N) admitted (see `https://arxiv.org/abs/2205.14310`)

Contact: allen@re.express, drasooly@bwh.harvard.edu