Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition

Allen Schmaltz

aschmaltz@hsph.harvard.edu

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Overview

We propose a new, more actionable view of neural network interpretability and data analysis by leveraging the remarkable matching effectiveness of representations derived from deep networks, guided by an approach for class-conditional feature detection.

We transform a deep network into a simple weighting over exemplar representations and associated labels, yielding an introspectable—and modestly updatable—version of the original model.

Task: Zero-Shot Binary Sequence Labeling

- Training: $\mathbb{D} = \{(\mathbf{x}_d, Y_d) \mid 1 \leq d \leq |\mathbb{D}|\}$
  - Document of N tokens/words: $\mathbf{x} = x_1, \ldots, x_n, \ldots, x_N$
  - Document-level label: $Y \in \{-1,1\}$

    Becomes supervised labeling with access to token-level labels: $\mathbf{y} = y_1, \ldots, y_n, \ldots, y_N$, where $y_n \in \{-1,1\}$
- Inference:
  - Predict token-level labels: $\hat{\mathbf{y}} = \hat{y}_1, \ldots, \hat{y}_n, \ldots, \hat{y}_N$, where $\hat{y}_n \in \{-1,1\}$
- Zero-Shot Grammatical Error Detection:

$y_1 = -1$ $y_2 = 1$ $y_3 = -1$ …

Sentence 1: `The runing example will be grammatical error detection, predicting whether or not each word has a grammatical error.` $Y = 1$

Sentence 2: `See the paper for additional datasets and tasks.` $Y = -1$

Training Convolutional (*Horizontal*) Decomposition (*see panel at right*)

- Cross-entropy against document-level label, $Y' \in \{0,1\}$
- Min-max constraint to encourage sparsity
  - $\mathcal{L}_{min} = -\log(1 - \sigma(s^{+-}_{min}))$
    - $s^{+-}_{min} = \min(s^{+-}_1, \ldots, s^{+-}_n, \ldots, s^{+-}_N)$
  - $\mathcal{L}_{max} = -Y' \cdot \log \sigma(s^{+-}_{max}) - (1 - Y') \cdot \log(1 - \sigma(s^{+-}_{max}))$
    - $s^{+-}_{max} = \max(s^{+-}_1, \ldots, s^{+-}_n, \ldots, s^{+-}_N)$
- Fully-supervised (token-level)
  - $\mathcal{L}_n = -y'_n \cdot \log \sigma(s^{+-}_n) - (1 - y'_n) \cdot \log(1 - \sigma(s^{+-}_n))$

Stronger priors (w.r.t. label distribution)

Zero-Shot Grammatical Error Detection

Model	Sentence-level F_1	Token-level P	Token-level R	Token-level F_1	Token-level $F_{0.5}$
Random	58.30	15.30	50.07	23.44	17.79
MajorityClass	80.88	15.20	100	26.39	18.31
LIME (RoBERTa$_{BASE}$ Transformer)†	84.51	19.06	34.70	24.60	20.95
LSTM+SoftAttention†	85.14	28.04	29.91	28.27	28.40
Transformer (RoBERTa$_{BASE}$) + WeightedSoftAttention†	85.62	20.76	85.36	33.31	24.46
Transformer (BERT$_{BASE}$) + CNNDecomposition	86.29	53.17	35.37	42.48	48.31

FCE zero-shot sequence labeling test set results (Appendix: Table E.1)

†Results from previous works

Model Approximation

$$\hat{y}_n = \text{sgn}\left(f(\mathbf{x}_n)\right) = \text{sgn}\left(s^{+-}_n\right) \approx$$

Original model output (from decomposition)

$$\hat{y}^{KNN}_n = \text{sgn}\left(f(\mathbf{x}_n)^{KNN}\right) = \text{sgn}\left(\beta + \sum_{k \in \arg K \min_{\tilde{n}} ||\mathbf{r}_n - \mathbf{r}_{\tilde{n}}||_2} w_k \cdot \left(\tanh(s^{+-}_k) + \gamma \cdot Y^{(k)}\right)\right)$$

K-NN Approximation

$y_k \in \{-1,1\}$ if token-level labels are available; otherwise, document-level $Y^{(k)} \in \{-1,1\}$

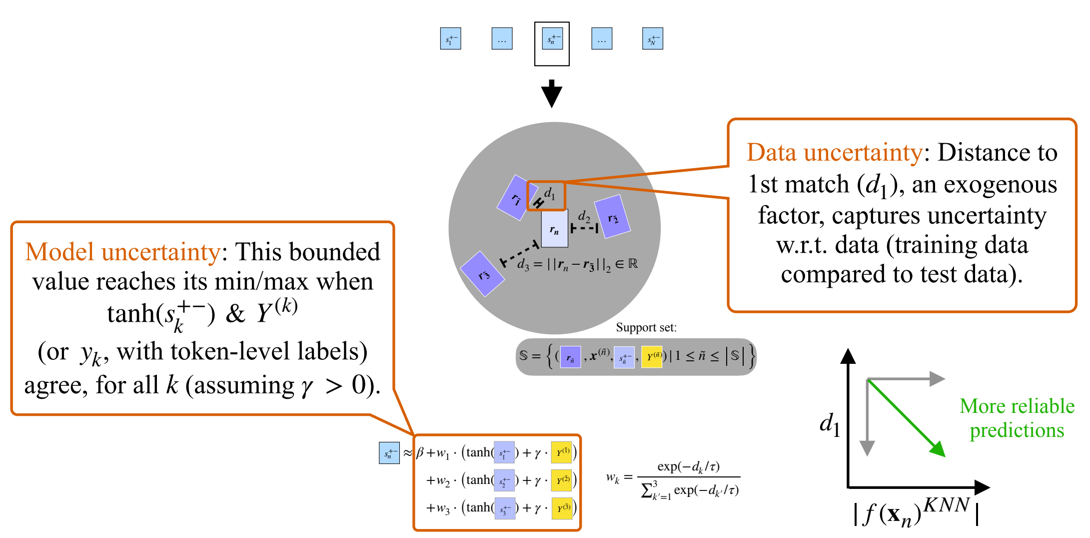
$$w_k = \frac{\exp\left(-||\mathbf{r}_n - \mathbf{r}_k||_2 / \tau\right)}{\sum_{k' \in \arg K \min_{\tilde{n}} ||\mathbf{r}_n - \mathbf{r}_{\tilde{n}}||_2} \exp\left(-||\mathbf{r}_n - \mathbf{r}_{k'}||_2 / \tau\right)}$$

Hyper-parameter: K; Learn β, γ, τ via:

$$\mathcal{L}^{KNN}_n = -\sigma(s^{+-}_n) \cdot \log \sigma\left(f(\mathbf{x}_n)^{KNN}\right) - (1 - \sigma(s^{+-}_n)) \cdot \log\left(1 - \sigma\left(f(\mathbf{x}_n)^{KNN}\right)\right)$$

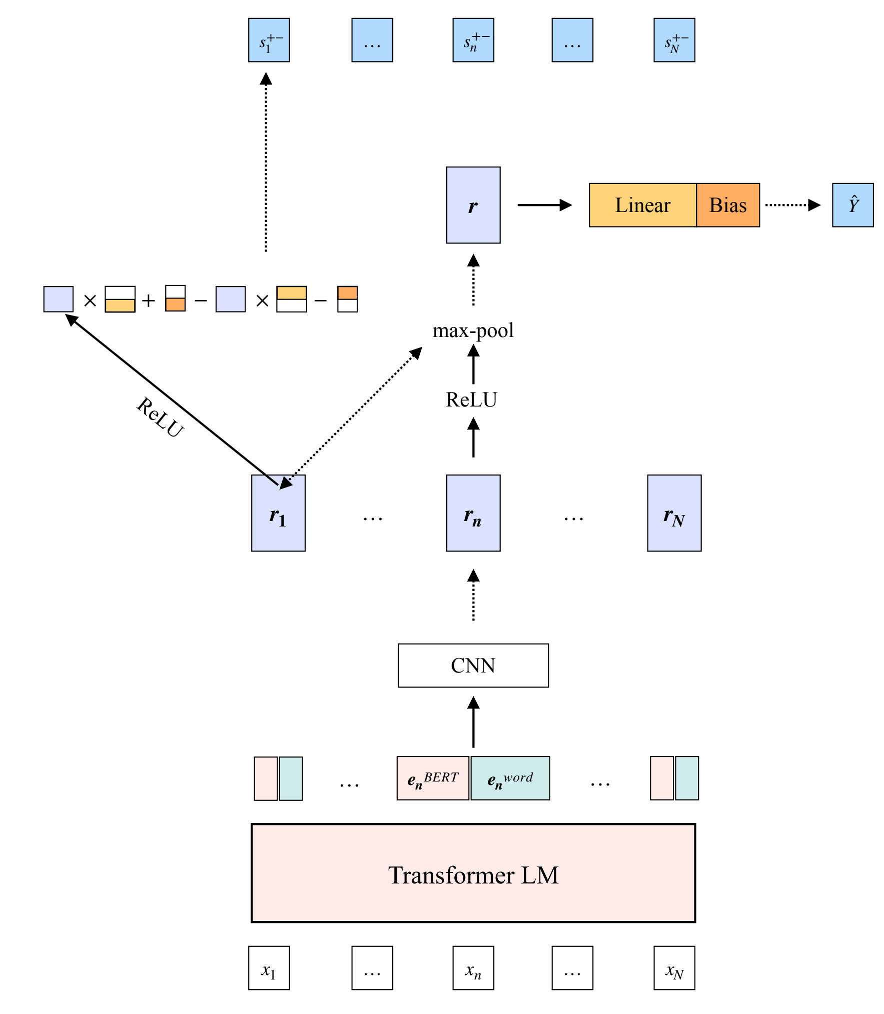
Choose epoch that minimizes: $\delta^{KNN} = \sum_{n \in \text{dev}} [\text{sgn}(s^{+-}_n) \neq \text{sgn}(f(\mathbf{x}_n)^{KNN})]$

Leveraging Model Approximations for Prediction Reliability Heuristics & Screening Input Dissimilar to the Support Set

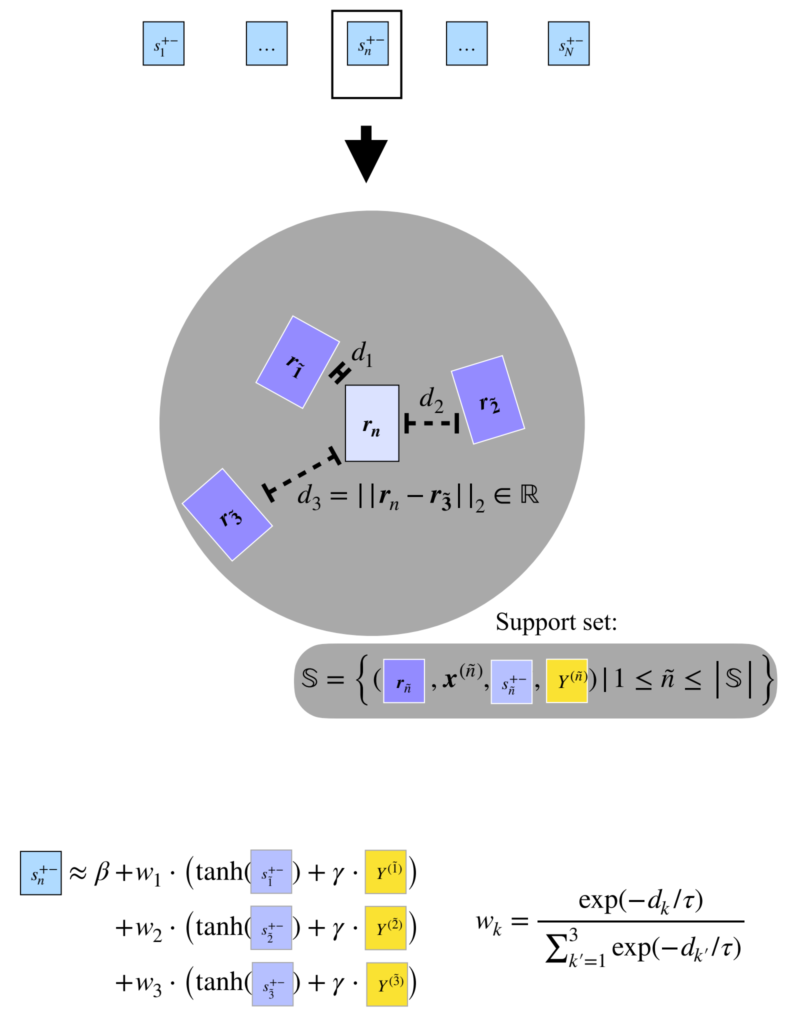


Horizontal (across the input) & *Vertical* (across the support set) Model Decompositions

Sequence Labeling via a Convolutional Decomposition



K-NN Approximation



Support Set Can be Viewed as an Updatable Database

$$\mathbb{S} = \left\{\left(\mathbf{r}_{\tilde{n}}, \mathbf{x}^{(\tilde{n})}, s^{+-}_{\tilde{n}}, Y^{(\tilde{n})}\right) \mid 1 \leq \tilde{n} \leq |\mathbb{S}|\right\}$$

Robustness to OOD data remains challenging, but we can detect such data and abstain from predicting:

Model: K-NN Approx. of Transformer (BERT$_{LARGE}$) +CNNDecomposition+MinMaxLoss

L^2 distance max constraint (Class -1, Class 1)	K-NN Output min threshold (Class -1, Class 1)	Admitted n	n/N	$F_{0.5}$
		92597	1.0	27.0
	(-1.2, 0.8)	38110	0.41	45.9
(34.2, 53.3)		7879	0.09	53.5
(34.2, 53.3)	(-1.2, 0.8)	4180	0.05	75.8

…and then update the support set:

Model	Training set	Support set	$F_{0.5}$
K-NN Approx.	FCE	FCE	27.0
K-NN Approx.	FCE	FCE+OOD	46.3
Original Model	FCE		25.8
Original Model	FCE+OOD		33.3

Token-level FCE+News2k (domain-shifted) test set results (Main text: Tables 5 & 6)