Deep Encoders with Auxiliary Parameters for Extreme Classification

Kunal Dahiya, Sachin Yadav, Sushant Sondhi, Deepak Saini, Sonu Mehta, Jian Jiao, Sumeet Agarwal, Purushottam Kar, Manik Varma

IIT Delhi · Microsoft · IIT Kanpur

Goal: Annotate a data point with the *most relevant subset* of labels from an extremely large set.

Extreme Classification (XC)

Applications

- Related web-page recommendation
- Matching user queries to advertiser keywords
- Product-to-product recommendation

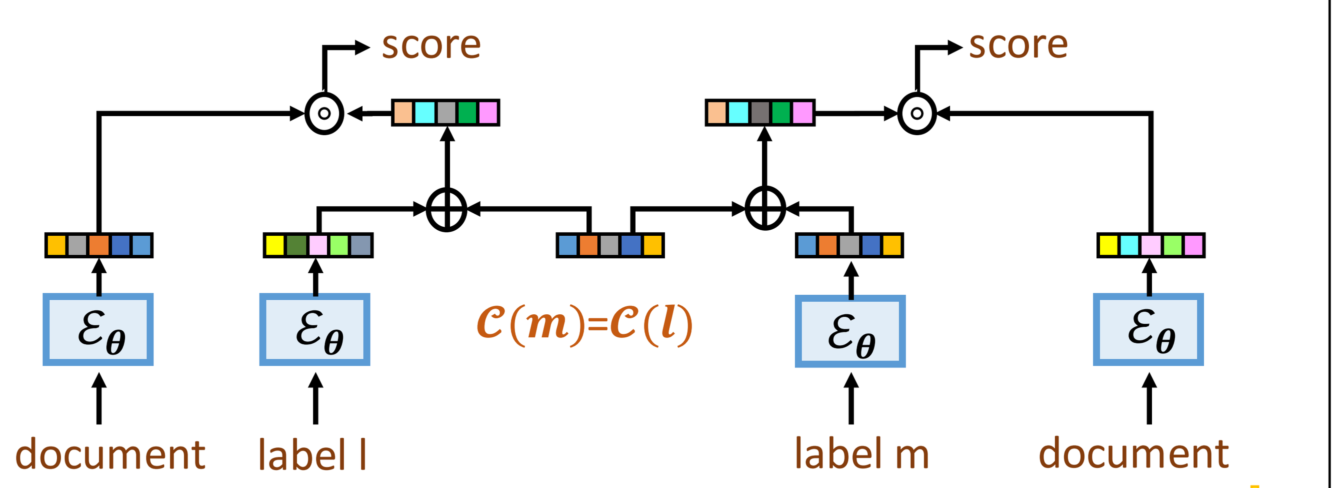
Semantic Gap in Siamese Networks

- Label text is unable to capture full meaning in case of short-text applications
- This leads to distortions in encoder training and sub-optimal accuracies
- Naïve solution to add L free vectors improve accuracies but does not scale

DEXA: Foundations

Dataset: $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}, \{\boldsymbol{z}_l\}_{l=1}^{L} \boldsymbol{y}_i \in \{-1,1\}^L$ where $\boldsymbol{x}_i, \boldsymbol{z}_l \in \boldsymbol{\mathcal{X}}$ are doc/label text respectively

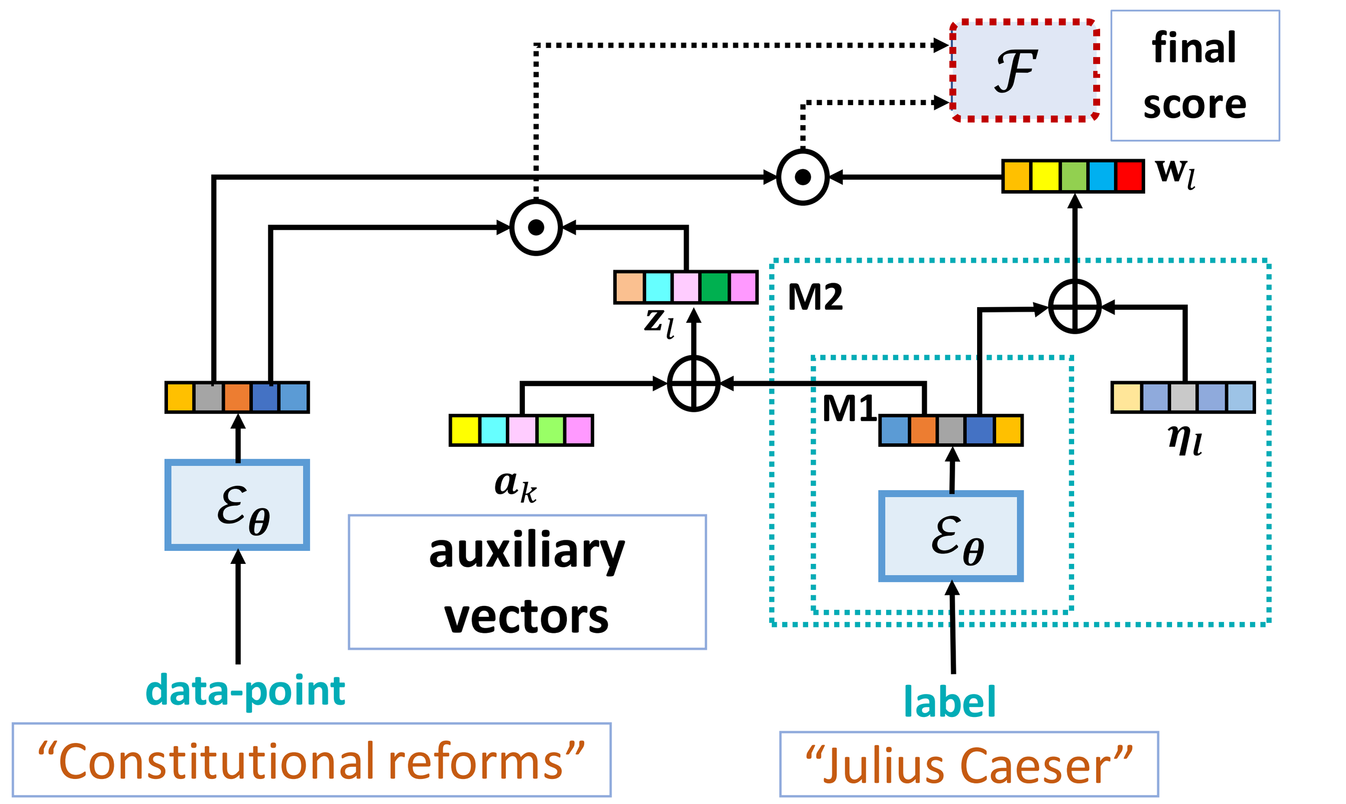
Goal: Learn params $\boldsymbol{\theta}$ for embedding architecture $\mathcal{E}_\theta : \boldsymbol{\mathcal{X}} \to \mathbb{R}^D$, One-vs.-All (OvA) classifiers $\boldsymbol{w}_l \in \mathbb{R}^D$, one for each label, to minimize triplet loss $\mathcal{L}$.



Related labels may have similar correction terms.

DEXA: Deep Encoders with Auxiliary Parameters for Extreme Classification

Architecture



Modular Training & Efficient Prediction

Module I (*Encoder with Auxiliary Parameters*):

- Cluster L labels into $K \ll L$ clusters
- Introduce K auxiliary parameters $(\boldsymbol{a}_k)$
- Use $\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{z}_l) + \boldsymbol{a}_k$ as surrogate for $\boldsymbol{w}_l$ and train embed. arch. $\widehat{\boldsymbol{\theta}}$ using a Siamese loss

Module II (Extreme Classifiers): Train $\boldsymbol{\eta}_l$ in $\mathcal{O}(ND\log L)$ time using +ve & hard –ve labels

Prediction: Efficient procedure, taking $\mathcal{O}(D^2 + D\log L)$ time per point

Illustrative Example

Data Point: Constitutional reforms of Julius Caesar..

Method	Top-5 predictions
DEXA	Acta Senatus ✓ Centuria ✓ Roman Law ✓ Interrex ✓ Byzantine Senate ✓
NGAME	Julius Caeser ✗ Assassination of Julius Caesar ✗ Caesarism ✗ Constitution of the Roman Republic ✗ Caesar's civil war ✗

Theoretical Results: Provable accurate training and crisp generalization bounds

Provable Accurate Training

Lemma: Consider a linear encoder parametrized by $\boldsymbol{E}$ & DEXA with auxiliary params $\boldsymbol{A}$, the gradient norms at optimal value $\boldsymbol{E}^*$:

$$\left\|\Delta_E \boldsymbol{\mathcal{L}}(\boldsymbol{E}^*)\right\|_2 \le 2\|\boldsymbol{E}_*\|_2^2 \sqrt{\frac{1}{L}\sum \|\Delta_l\|_2^2}$$

$$\left\|\Delta_E \boldsymbol{\mathcal{L}}(\boldsymbol{E}^*, \boldsymbol{A})\right\|_2 \le 4\|\boldsymbol{E}_*\|_2^2 \sqrt{\frac{1}{L}\sum \sigma_k^2}$$

where, Δ_l is semantic gap for label l, and σ_k^2 is intra-cluster variance in semantic gap

- DEXA offers smaller encoder gradient, indicating a more faithful recovery of true encoder parameters, if $\sigma_k^2 \ll \sum_{l\in C_k} \left\|\Delta_l\right\|_2^2$
- Even if the individual Δ_l are large in a cluster, DEXA offers faithful encoder recovery so long that semantic gaps are similar to each other

Generalization Bounds

Theorem: Suppose DEXA is used with an encoder parameters θ and auxiliary parameters $\boldsymbol{A}$, then with probability $1-\delta$, we have

$$\ell(\boldsymbol{\theta}, \boldsymbol{A}) \le \hat{\ell}_N(\boldsymbol{\theta}, \boldsymbol{A}) + \epsilon(N) + \sqrt{\frac{\ln\frac{1}{\delta}}{N}} + \frac{\Delta \ln N}{\sqrt{N}}$$

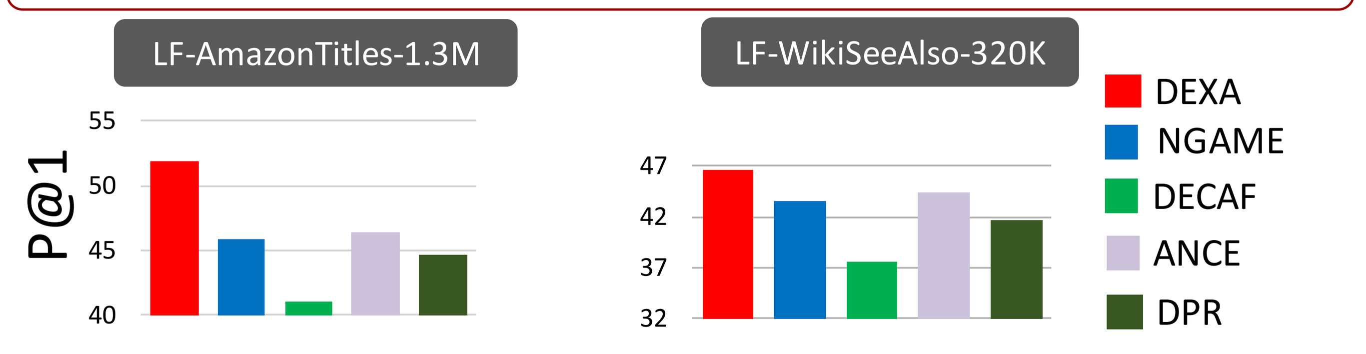
where, $\Delta = \mathcal{O}(\ln(DK))$ apart from numerical constants independent of L

- $\epsilon(N)$ captures the dependence of the excess risk on the encoder parameter characteristics
- Independent of L in favour of $\mathcal{O}(\ln(DK))$

Results: Significant gains in offline evaluation; Readily incorporate with existing architectures

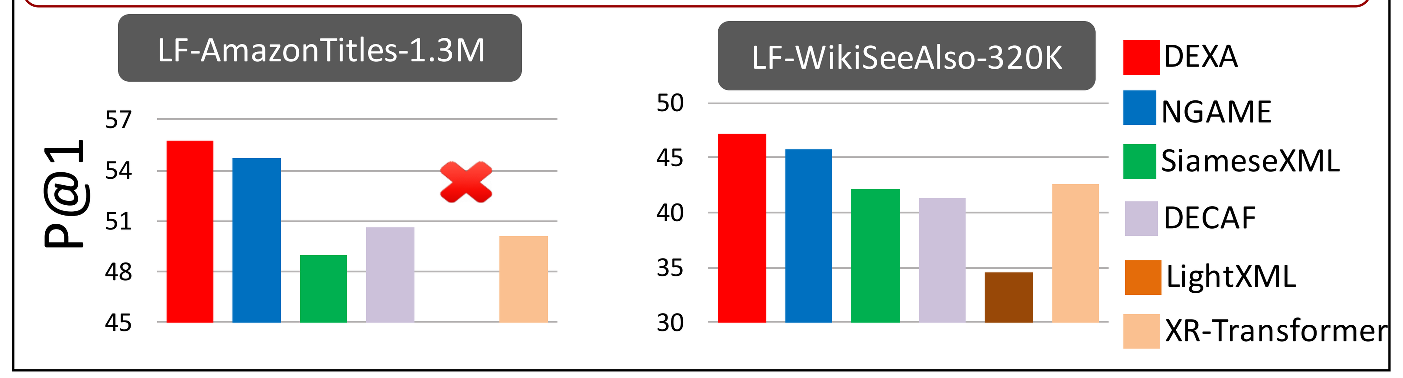
Embeddings on Benchmark Datasets

Up to 11% more accurate embeddings

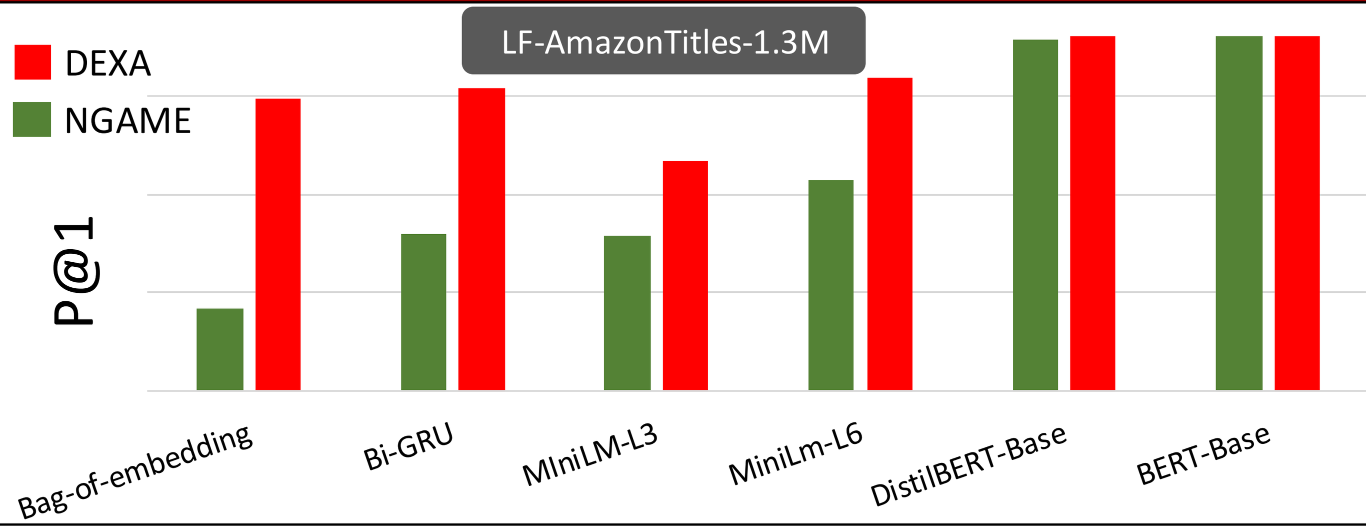


End-to-end Results on Benchmark Datasets

Up to 12% gains over leading existing XC methods



Ablation over architectures



Sponsored Search-40M

7-15% gains in matching queries to keywords