Review of Breaking the Glass Ceiling of Embedding based Classifiers for Large output spaces - NIPS'19[⋆]

Siddhant Katyan[1]

IIIT-Hyderabad, India
siddhant.katyan@research.iiit.ac.in

Abstract. This paper shows that in contrary to the general belief of low-dimensional bottleneck in embedding based methods in extreme classification, there is rather no limitation to using low-dimensional embedding. It claims that overfitting is the root cause of poor performance embedding based methods for extreme classification.

Keywords: Extreme classification · Embedding · Regularization.

1 Introduction

One of the popular methods to approach the problem of extreme classification, which involves large output space is through low-dimensional embeddings. Here, the basic model comprises of an *embedding function* $\phi : R^D \rightarrow R^d$, where D is the original dimension, d is the embedding dimension ($d << D$) and classifier $f : R^D \rightarrow \{0,1\}^K$. Thus, for any input $x \in R^D, f(\phi(x))$ is the label vector of predicted labels.

Key Contributions Three major contributions are:
1. Demonstrates experimentally that the main reason for the poor performance of neural network embedding-based models is overfitting. And prove that there exists a low-dimensional embedding-based linear classifier with perfect accuracy in the limit of infinite expressivity of the embedding map.
2. Propose data augmentation and regularization techniques, including a novel regularizer called GLaS. It is successful in shrinking the generalization gap for embedding based methods.
3. Use stochastic negative mining loss proposed in 2019 NIPS paper.

2 Problem Setting

Given an input $x \in X \subset R^D$, its label $y \in Y \subset \{0,1\}^K$ is a K dimensional vector with multiple non-zero entries, where $y^k = 1$ if and only if label k is relevant for

[⋆] Supported by organization x.

input x. Let L_y denote the set of indices that are non-zero in y. The elements of set the L_y are, hereafter, referrred to as *relevant labels* in y. The goal of all embedding-based methods is to learn a model of the form $f(\phi(x)) : X \rightarrow \{0,1\}^K$ where $\phi(x) \in R^d$ and $d << D, K$ and $f : R^d \rightarrow \{0,1\}^K$ is a classifier on top of the embedding.

The most common form of classifier is a linear classifier. A linear classifier is parameterized by a *label embedding matrix* $\mathbf{V} \in R^{d \times K}$ which is used to predict *scores* for all labels by computing $\phi(x)^T \mathbf{V}$. The set of labels predicted for the input x can be obtained by thresholding the scores at some value τ, i.e., $y : \phi(x)^T \mathbf{v}_y \geq \tau$.

The use of a linear classifier on top of embeddings naturally leads to a low-rank structure for the score vectors of the labels: the set$\phi(x)^T \mathbf{V} : x \in X$ has rank at most d. We should note that the label vectors are generated by either thresholding the scores at some value τ or taking the top m largest scores, which is a highly non linear transformation. Thus, it is not immediately clear if the low-rank structure of the score vectors directly translates to a low-rank structure of the label vectors.

3 Regularizing Embedding-Based Models

3.1 Embedding Normalization

We first apply weight normalization. All label embeddings are l_2-normalized to unit norm, similarly, input embeddings are normalized as well. The proposed regularizer can be easily generalized to cases where the label embeddings are not unit norm.

3.2 GLaS Regularizer

As in large scale mult-label classification, the output space is both large and sparse – most feature vectors are associated with only very few true labels. Thus it may be desirable for an embedding-based classifier to have near-orthogonal label embeddings.

Spread-out Regularization. It brings the inner product of the embeddings of two different labels close to zero, i.e., $v_y^T v_{y'} \approx 0$. if $y \neq y'$.

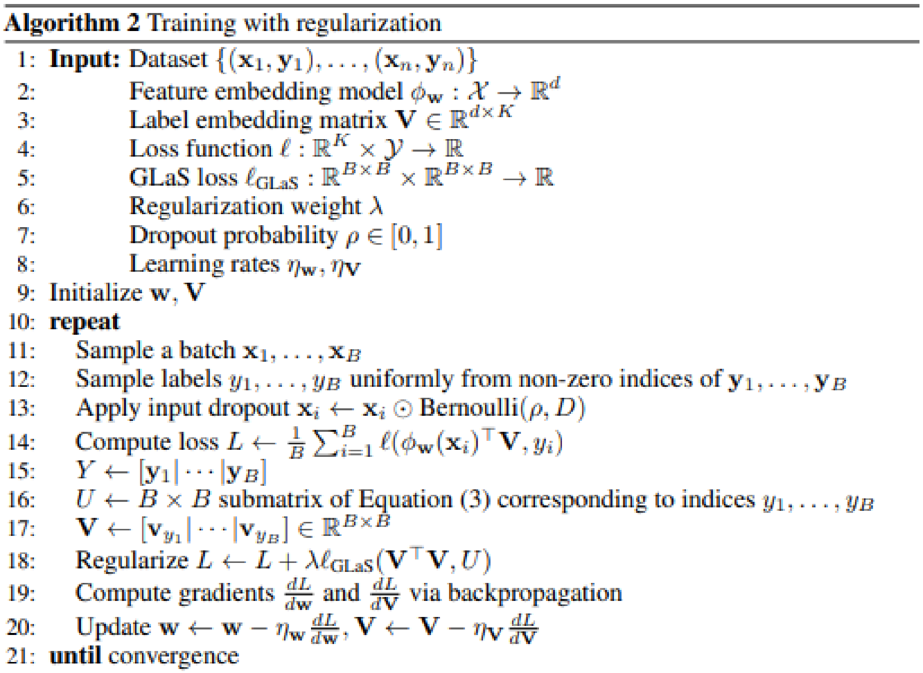
$$l_{spreadout} = \frac{1}{K^2} \sum_{y=1}^{K} \sum_{y'=1}^{K} \left(v_y^T v_{y'}\right)^2 \tag{1}$$

Drawback of this regularizer is that it over-penalizes the embeddings of two different labels that co-occur frequently together (e.g., *apple* and *fruit* tend to co-occur for many inputs). Label embeddings of labels that co-occur frequently are also encouraged to be far away, which is clearly undesirable.

Correcting Over-penalization: GLaS Regularization. Let $Y \in \{0,1\}^{n \times K}$ be the training set label matrix where each row corresponds to single training example. Let $A = Y^T Y$ so that $A_{y,y'}$ = number of times label y and y' co-occur, and let $Z = diag(A) \in R^{K\times K}$ be the matrix containing the diagonal component of A. Observe that AZ^{-1} represents the conditional frequency of observing one label given the other. Similarly, $Z^{-1}A = (AZ^{-1})_T$ contains the conditional frequencies in reverse.

These conditional frequencies encode the degree of co-occurence between labels y and y', and we would like their embeddings v_y and $v_{y'}$ to reflect this co-occurence pattern:

$$l_{GLaS} = \frac{1}{K^2}\|\mathbf{V}^T V - \tfrac{1}{2}(AZ^{-1} + Z^{-1}A)\|_F (2)$$



Algorithm 2 Training with regularization

```
Input: Dataset $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$
        Feature embedding model $\phi_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}^d$
        Label embedding matrix $\mathbf{V} \in \mathbb{R}^{d\times K}$
        Loss function $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$
        GLaS loss $\ell_{\text{GLaS}} : \mathbb{R}^{B\times B} \times \mathbb{R}^{B\times B} \to \mathbb{R}$
        Regularization weight $\lambda$
        Dropout probability $\rho \in [0, 1]$
        Learning rates $\eta_{\mathbf{w}}, \eta_{\mathbf{V}}$
Initialize $\mathbf{w}, \mathbf{V}$
repeat
    Sample a batch $\mathbf{x}_1, \ldots, \mathbf{x}_B$
    Sample labels $y_1, \ldots, y_B$ uniformly from non-zero indices of $\mathbf{y}_1, \ldots, \mathbf{y}_B$
    Apply input dropout $\mathbf{x}_i \leftarrow \mathbf{x}_i \odot \text{Bernoulli}(\rho, D)$
    Compute loss $L \leftarrow \frac{1}{B}\sum_{i=1}^{B} \ell(\phi_{\mathbf{w}}(\mathbf{x}_i)^\top \mathbf{V}, y_i)$
    $Y \leftarrow [\mathbf{y}_1 | \cdots | \mathbf{y}_B]$
    $U \leftarrow B \times B$ submatrix of Equation (3) corresponding to indices $y_1, \ldots, y_B$
    $\mathbf{V} \leftarrow [\mathbf{v}_{y_1} | \cdots | \mathbf{v}_{y_B}] \in \mathbb{R}^{B\times B}$
    Regularize $L \leftarrow L + \lambda\ell_{\text{GLaS}}(\mathbf{V}^\top\mathbf{V}, U)$
    Compute gradients $\frac{dL}{d\mathbf{w}}$ and $\frac{dL}{d\mathbf{V}}$ via backpropagation
    Update $\mathbf{w} \leftarrow \mathbf{w} - \eta_{\mathbf{w}}\frac{dL}{d\mathbf{w}}, \mathbf{V} \leftarrow \mathbf{V} - \eta_{\mathbf{V}}\frac{dL}{d\mathbf{V}}$
until convergence
```

Fig. 1. Training with GLaS Regularizer

4 Doubts

Doubt 1: The paper claims that non-regularized embedding-based method severely overfits to the training data as show below:

The justification of measuring the overfitting via accuracy doesn't seems to be valid in this case as extreme classification datasets follow a power-law distribution. Computing accuracy is valid in those datasets where the number of samples for each of the labels is approximately same, which is not the case in extreme classification datasets.

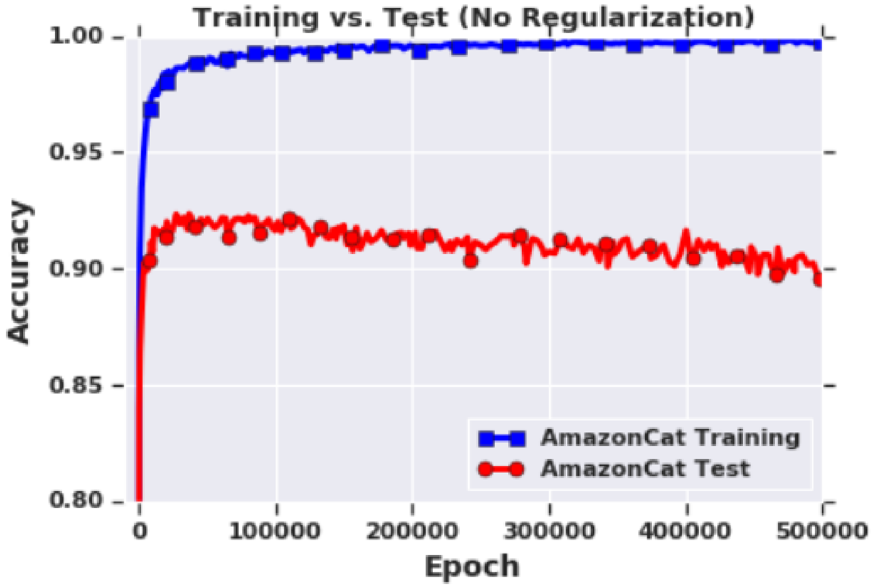


Fig. 2. Overfitting of embedding based method AMAZONCAT-13K dataset

Doubt 2 It is not immediately clear if the low-rank structure of the score vectors directly translates to a low-rank structure on the label vectors. There has not been any definitive proof that the inherent problem with embedding-based methods is their use of low-dimensional representations for the score vectors.