Link the head to the “peak”: Zero Shot Learning from Noisy Text description at Part Precision

Supplementary Materials

Mohamed Elhoseiny\textsuperscript{1,2}, Yizhe Zhu\textsuperscript{1}, Han Zhang\textsuperscript{1}, and Ahmed Elgammal\textsuperscript{1}

elhoseiny@fb.com, yizhe.zhu@rutgers.edu, \{han.zhang, elgammal\}@cs.rutgers.edu

\textsuperscript{1}Rutgers University, Department of Computer Science, \textsuperscript{2}Facebook Research

This supplementary document includes the following sections. For reproducing the results, our code, data, and models are publicly available link [1].

1. \(\lambda_1\) and \(\lambda_2\) Setting in Experiments.
2. Gradient Derivations
3. More Qualitative Examples
4. More Figures and Detailed Results

\section{1. \(\lambda_1\) and \(\lambda_2\) Setting}

Similar to methods in the literature (e.g., \cite{3, 2}), we learn the hyper-parameters by cross validation (CV) on the validation set, with the grid search in the range of 10\([3:6]\) for both \(\lambda_1\) and \(\lambda_2\) (i.e., 4 \times 4 grid).

1. \textit{CUB (Easy Split)}: \(\lambda_1 = 10^5\) and \(\lambda_2 = 10^4\).
2. \textit{NABirds (Easy Split)}: \(\lambda_1 = 10^5\) and \(\lambda_2 = 10^4\).
3. \textit{CUB (Hard Split)}: \(\lambda_1 = 10^6\) and \(\lambda_2 = 10^4\).
4. \textit{NABirds (Hard Split)}: \(\lambda_1 = 10^6\) and \(\lambda_2 = 10^5\).

We find it intuitive to see higher values lamdas after cross-validation for the \textit{Hard Split} since regularization becomes more important as shared information gets smaller. Moreover, we did not find the method very sensitive to the hyper parameters. For instance, the performance on \textit{CUB (Easy Split)} with \(\lambda_1 = 10^5\), the performance of \(\lambda_2 = 10^3\), \(\lambda_2 = 10^4\), and \(\lambda_2 = 10^5\) are 35.4\%, 37.2\%, and 35.9\%, respectively.

\section{2. Gradient Derivations}

\subsection{2.1. Gradients for Equation 5: Fix \(W_t\), and optimize over \(W_x\)}

We name the loss in Equation 3 in the paper as \(L\).

Let \(X = [X^{(1)}, X^{(2)}, \ldots, X^{(P)}] \in \mathbb{R}^{d \times P \times d} \).

Let \(W_x^T = [W_x^1, W_x^2, \ldots, W_x^P], \) where \(W_x^p \in \mathbb{R}^{d \times d}; \) and \(W_t \in \mathbb{R}^{d \times d T}.\)

(a) The first term:

\[ \| (\sum_{p=1}^{P} X^{(p)^T} W_x^p) W_t T - Y \|^2_F = \| X^T W_x^T W_t T - Y \|^2_F = \text{Tr}((X^T W_x^T W_t T - Y)(X^T W_x^T W_t T - Y)^T) \quad (1) \]

\* Both authors contributed equally to this work
We can get the derivative of the first term in the objective function w.r.t. $W_p$:

$$\frac{\partial ||X^TW_x^TW_tT - Y||_F^2}{\partial W_p^{(p)}} = 2W_tTT^TW_t^TW_xX^{(p)}X^{(p)T} - 2W_tTY^TX^{(p)T}$$ (2)

(b) The derivative of the second term in the objective function w.r.t. every part $W_p$:

$$\frac{\partial \lambda_1||W_x^TW_tT||_F^2}{\partial W_p^{(p)}} = 2\lambda_1W_tTT^TW_t^TW_p$$ (3)

(c) For the third term in the objective function, we do the partial derivative for each part:

$$\lambda_2Tr(W_p^pW_t^pD_t^pW_t^TW_p^T)$$ (4)

The derivative of the third term in the objective function w.r.t. every part $W_p^p$:

$$\frac{\partial \lambda_2Tr(W_p^pW_t^pD_t^pW_t^TW_p^T)}{\partial W_p^{(p)}} = 2\lambda_2W_p^pW_t^pD_t^pW_t^T$$ (5)

Therefore, the partial derivative of the loss function w.r.t. $W_p^p$ is:

$$\frac{\partial L}{\partial W_p^{(p)}} = 2W_tTT^TW_t^TW_xX^{(p)}X^{(p)T} - 2W_tTY^TX^{(p)T} + 2\lambda_1W_tTT^TW_t^TW_p^p + 2\lambda_2W_p^pW_t^pD_t^pW_t^T$$ (6)

2.2. Gradients for Equation 4: Fix $W_p^p$, and optimize over $W_t$

The loss function is rewritten as:

$$L = ||X^TW_x^TW_tT - Y||_F^2 + \lambda_1||W_x^TW_tT||_F^2 + \lambda_2 \sum_{p=1}^P Tr(W_p^pW_t^pD_t^pW_t^TW_p^T)$$ (7)

The partial derivative over $W_t$:

$$\frac{\partial L}{\partial W_t^T} = 2W_xXT^TW_x^T - 2W_xYT + 2\lambda_1W_xX^TW_tTT^T + 2\lambda_2 \sum_{i=1}^P W_p^T^TW_p^pW_x^T^TD_t^p$$ (8)

3. More Qualitative Results

We show more qualitative examples in this section.

Figure 1: Part-to-Term connectivity demonstrated on falsely labeled samples
Figure 2: Part-to-Term connectivity demonstrated on correctly labeled samples
4. More Figures and Detailed Results

More Generalized Zero-Shot Learning Curves; see the captions for the corresponding benchmark.

(a) CUBirds Seen-Unseen accuracy Curve on SCS split

(b) NABirds Seen-Unseen accuracy Curve on SCS split

(c) CUBirds Seen-Unseen accuracy Curve on SCE split

(d) NABirds Seen-Unseen accuracy Curve on SCE split

Figure 3: Result comparison with Seen-Unseen accuracy Curves on different split settings.

References

