Incremental Sense Weight Training for In-Depth Interpretation of Contextualized Word Embeddings (Student Abstract)

Xinyi Jiang, Zhengzhe Yang, Jinho D. Choi
Department of Computer Science, Emory University, Atlanta GA 30322
xinyi.jiang2@outlook.com, {zhengzhe.yang, jinho.choi}@emory.edu

Abstract
We present a novel online algorithm that learns the essence of each dimension in word embeddings. We first mask dimensions determined unessential by our algorithm, apply the masked word embeddings to a word sense disambiguation task (WSD), and compare its performance against the one achieved by the original embeddings. Our results show that the masked word embeddings do not hurt the performance and can improve it by 3%.

Introduction
Contextualized word embeddings generate different embeddings for the same word type with different topical senses. In this work, we propose an algorithm that learns the dimension importance in representing sense information by minimizing the distance between sense groups. The effectiveness of our approach is validated by a word sense disambiguation task (WSD) that aims to distinguish the correct senses of words under different contexts, as well as two intrinsic evaluations of embedding groups on the masked embeddings. A full-length paper of our work is available1.

In previous embedding interpretation work, matrix transformation has been widely used (Zobnin 2017; Park, Bak, and Oh 2017). Others apply sparse encoding techniques and map embeddings to sparse vectors to increase vector interpretability (Subramanian et al. 2018; Arora et al. 2018). In this work, a novel idea where the information contained in dimensions of word embeddings is evaluated from a pure machine learning perspective. Three popular word embedding algorithms are used for our experiments: ELMo (Peters et al. 2018), Flair (Akbik, Blythe, and Vollgraf 2018), and BERT (Devlin et al. 2019).

Sense Weight Training (SWT)
With word embedding groups classified by their senses annotated in the SemCor dataset (Miller et al. 1994), the objective is to maximize the average pair-wise cosine similarity in sense groups. A weight matrix (size of embedding) is initialized for each sense and each dimension corresponds to the importance of the embedding dimension to that sense.

Algorithm 1 Algorithm for the incremental Sense Weight Training. n is the number of epochs for exploration, λ the parameter for $l_1$ regularization and $\epsilon$ a small number.


\begin{algorithm}
\begin{algorithmic}
\Input \text{for each sense group SG do}
\Input initialize weights $w$, learning rate $\gamma_0$, Adagrad weights $g_{ti}$
\Input initialize $S_{pre}$
\Input $S_{pre} \leftarrow \sum_{v_i, v_j \in SG, i \neq j} \text{Cosine}(v_i, v_j)$
\Input for each epoch $i$ do
\Input \If {$i < n$ then}
\Input randomly generate N numbers: $D_1, \cdots, D_N$
\Input \Else
\Input generate N numbers based on policy: $D_1, \cdots, D_N$
\InputDialog for $v_i \in SG$
\InputDialog $S_{cur} \leftarrow \sum_{v_i, v_j \in SG, i \neq j} \text{Cosine}(v_i, v_j)$
\InputDialog $\text{grad} = (S_{pre} - S_{cur}) \ast (\text{mask} - 1) - \lambda \ast \text{sign}(w)$
\InputDialog $g_{ti} += \text{grad}^2$
\InputDialog $w \leftarrow \frac{w + \text{grad} \ast \gamma_i}{\epsilon + \sqrt{g_{ti}}}$
\Input end for
\EndInput for
\end{algorithmic}
\end{algorithm}

During training, a mask matrix is applied, which is the size of the weight matrix and has $N$ zeros with the rest ones. The generation of the mask matrix involves first randomly generating $N$ positions of zeros to ensure enough dimensions have been covered and then employing an exploration-exploitation policy: there is a chance of $\alpha$ to randomly generate $N$ numbers and for the rest $1 - \alpha$ probability, the higher weight the dimension number has, the lower probability of the number getting selected. Furthermore, $l_1$ regularization is applied for feature selection purpose, and AdaGrad is used to encourage convergence.

1"https://arxiv.org/abs/1911.01623"
Experiments

SWT algorithm is evaluated by comparing the WSD performances of masked and unmasked embeddings. In this work, the embedding dimensions with weight value ranked below 5% are marked to zero. KNN method is used with an evaluation framework (Raganato, Camacho-Collados, and Navigli 2017). The embeddings from all output layers of ELMo, BERT and Flair are evaluated. Table 1 proves that for ELMo and Flair-2048, masking does not hurt the performance too much and for single layers, it even shows improvements. Figure 1 shows a performance boost for the last 10 layer outputs. Surprisingly, the last layer output score is boosted by 3%.

| Model     | Original | Masked |
|-----------|----------|--------|
| Flair-4096 | 63.7     | 62.1   |
| Flair-2048 | 60.5     | 60.7   |
| BERT      | 67.3     | 64.5   |
| ELMo      | 63.8     | 63.0   |
| ELMo-256  | 61.5     | 62.3   |
| ELMo-512  | 62.7     | 63.0   |
| ELMo-1024 | 62.5     | 63.4   |

Table 1: Results for the original and embeddings with 5% dimensions masked.

Table 2: Correlation coefficient test results

| Model   | Dim | \(N_{\text{masked}}\) | \(\rho_{\text{original}}\) | \(\rho_{\text{masked}}\) |
|---------|-----|------------------------|---------------------------|------------------------|
| BERT    | 768 | 125                    | 0.26814                   | 0.25286                |
| BERT    | 1024| 146                    | 0.27423                   | 0.26575                |
| ELMo    | 256 | 218                    | 0.2852                    | 0.3042                 |
| ELMo    | 512 | 281                    | 0.29577                   | 0.36943                |
| ELMo    | 1024| 608                    | 0.28406                   | 0.30675                |
| Flair   | 2048| 670                    | 0.24891                   | 0.28516                |

Conclusion

This paper demonstrates a novel approach to interpret word embeddings. A conclusion can be drawn from the results that some dimensions can be determined to have little contribution to the representation of sense groups by our algorithm. There are several limitations to this work. First, for the evaluation, the path similarity used may not be the best to fit human judgements. Second, the current tests were limited by the dataset corpus mentioned. For future works, the applications of the algorithm can theoretically be applied to other grouped embeddings, which would require more explorations.

References

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of CICLing*, 1638–1649.

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2018. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *TACL* 6:483–495.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL: HLT*, 4171–4186.

Landes, S.; Leacock, C.; and Fellbaum, C. 1998. *Building Semantic Concordances*. WordNet: An Electronic Lexical Database. 199–216.

Miller, G. A.; Chodorow, M.; Landes, S.; Leacock, C.; and Thomas, R. G. 1994. Using a Semantic Concordance for Sense Identification. In *HLT*.

Park, S.; Bak, J.; and Oh, A. 2017. Rotated Word Vector Representations and their Interpretability. In *Proceedings of EMNLP*, 401–411.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL: HLT*, 2227–2237.

Raganato, A.; Camacho-Collados, J.; and Navigli, R. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL*, 99–110.

Subramanian, A.; Pruthi, D.; Jhamtani, H.; Berg-Kirkpatrick, T.; and Hovy, E. H. 2018. SPINE: SParse Interpretable Neural Embeddings. In *Proceedings of AAAI*, 4921–4928.

Zobnin, A. 2017. Rotations and Interpretability of Word Embeddings: the Case of the Russian Language. *AIST* 10716:116–128.