The Pupil Has Become the Master: Teacher-Student Model-Based Word Embedding Distillation with Ensemble Learning

Bonggun Shin\textsuperscript{1}, Hao Yang\textsuperscript{2}, Jinho D. Choi\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Emory University, Atlanta, GA
\textsuperscript{2}Visa Research, Palo Alto, CA
bonggun.shin@emory.edu, haoyang@visa.com, jinho.choi@emory.edu

Abstract

Recent advances in deep learning have facilitated the demand of neural models for real applications. In practice, these applications often need to be deployed with limited resources while keeping high accuracy. This paper touches the core of neural models in NLP, word embeddings, and presents a new embedding distillation framework that remarkably reduces the dimension of word embeddings without compromising accuracy. A novel distillation ensemble approach is also proposed that trains a high-efficient student model using multiple teacher models. In our approach, the teacher models play roles only during training such that the student model operates on its own without getting supports from the teacher models during decoding, which makes it eighty times faster and lighter than other typical ensemble methods. All models are evaluated on seven document classification datasets and show significant advantage over the teacher models for most cases. Our analysis depicts insightful transformation of word embeddings from distillation and suggests a future direction to ensemble approaches using neural models.

1 Introduction

As deep learning starts dominating the field of machine learning, there have been growing interests in deploying deep neural models for real applications. [Hinton \textit{et al.}, 2014] stated that academic research on model development had mostly focused on accuracy improvement, whereas the deployment of deep neural models would also require the optimization of other practical aspects such as speed, memory, storage, power, etc. To satisfy these requirements, several neural model compression methods have been proposed, which can be categorized into the following four: weight pruning [De-nil \textit{et al.}, 2013; Han \textit{et al.}, 2015; Jurgovsky \textit{et al.}, 2016], weight quantization [Han \textit{et al.}, 2016; Jurgovsky \textit{et al.}, 2016; Ling \textit{et al.}, 2016], lossless compression [Van Leeuwen, 1976; Han \textit{et al.}, 2015], and distillation [Mou \textit{et al.}, 2016]. This paper focuses on distillation methods that can remarkably reduce the model size, resulting in much less memory usage and fewer computations.

Distillation aims to extract core elements from a complex network and transfer them to a simpler network so it gives comparable results to the complex network. It has been shown that the core elements can be transferred to various types of networks i.e., deep to shallow networks [Ba and Caruana, 2014], recurrent to dense networks [Chan \textit{et al.}, 2014], and vice versa [Romero \textit{et al.}, 2014; Tang \textit{et al.}, 2016]. Lately, embedding distillation was suggested [Mou \textit{et al.}, 2016], which transferred the output of the projection layer in the source network as input to the target network, although accuracy drop was expected with this approach. Considering the upper bound of a distilled network, that is the accuracy achieved by the original network [Ba and Caruana, 2014], enough room is left for the improvement of embedding distillation. Distilled embeddings can significantly enhance the efficiency of deep neural models in NLP, where the majority of model space is occupied by word embeddings.

In this paper, we first propose a new embedding distillation method based on three teacher-student frameworks, which is a more advanced way of embedding distillation, because the previous one [Mou \textit{et al.}, 2016] is a standalone embedding distillation (Section 2.4) with limited knowledge transfer. Our distilled embeddings not only enable the target network to outperform the previous state of the art [Mou \textit{et al.}, 2016], but also are eight times smaller than the original word embeddings yet allow the target network to achieve compatible (sometimes higher) accuracy to the source network. We then present a novel ensemble approach which extends this distillation framework by allowing multiple teacher models when training a student model. After learning from multiple teachers during training, the student model runs on its own during decoding such that it performs faster and lighter than any of the teacher models yet pushes the accuracy much beyond them with just 1.25% (50/4000) the size of other typical ensemble models. All models are evaluated on seven document classification datasets; our experiments show the effectiveness of the proposed frameworks, and our analysis illustrates an interesting nature of the distilled word embeddings. To the best of our knowledge, this is the first time that embedding distillation is thoroughly examined for natural language processing and used in ensemble to achieve such promising results.\footnote{\url{https://github.com/bgshin/distill_demo}}
2 Background

Our embedding distillation framework is based on teacher-student models [Ba and Caruana, 2014; Sau and Balasubramanian, 2016; Hinton et al., 2014], where teacher models are trained on deep neural networks and transfer their knowledge to student models on simpler networks. The following subsections describe three popular teacher-student methods applied to our framework. The main difference between these three methods is in their cost functions (Figure 1). The last subsection discusses embedding encoding that is used to extract distilled embeddings from the projection layer.

Throughout this section, a logit refers to a vector representing the layer immediately before the softmax layer in a neural network, where $z_i$ and $v_i$ are the teacher’s and student’s logit values for the class $i$, respectively. Note that the student models are not necessarily optimized for only the gold labels but also optimized for the logit values from the teacher models in these methods.

2.1 Logit Matching (LM)

Proposed by [Ba and Caruana, 2014], the cost function of this teacher-student method is defined by the logit differences between the teacher and the student models ($D$: the total number of classes):

$$L_{LM} = \frac{1}{2} \cdot D \sum_{i=1}^{D} |z_i - v_i|^2$$

2.2 Noisy Logit Matching (NLM)

Proposed by [Sau and Balasubramanian, 2016], this method is similar to Logit Matching except that Gaussian noise is introduced during the distillation, simulating variations in the teacher models, which gives a similar effect for the student to learn from multiple teachers. The cost function takes random noise $\eta$ drawn from Gaussian distribution such that the logit of each teacher model is $z'_i = (1 + \eta) \cdot z_i$. Thus, the final cost function becomes $L_{NLM} = \frac{1}{2D} \sum_{i=1}^{D} |z'_i - v_i|^2$.

2.3 Softmax Tau Matching (STM)

Proposed by [Hinton et al., 2014], this method is based on softmax matching where softmax values are compared between the teacher and the student models instead of logits. Later, [Hinton et al., 2014] added two hyperparameters to further generalize this method. The first hyperparameter, $\lambda$, is for the weighted average of two sub-cost functions, where the first sub-cost function measures a cross-entropy between the student’s softmax predictions and the truth values, represented as $L_1 = -\sum_{i,j} y_i \log p_i$ ($i$ indexes classes, $y$ is the gold label, $p_i \in (0, 1)$ is the prediction for a sample). Another cost function involves the second hyperparameter, $\tau$, that is a temperature variable normalizing the output of the teacher’s logit value:

$$s_i(z, \tau) = \frac{e^{z_i/\tau}}{\sum_{j=1}^{D} e^{z_j/\tau}}$$

Given $s_i$, the second cost function can be defined as $L_2 = -\sum_{i,j} s_i(z, \tau) \log p_i$. Therefore, the final cost function becomes $L_{STM} = \lambda L_1 + (1 - \lambda) L_2$. If $\lambda$ weighs more on $L_1$, the student model values more on the gold labels than teacher's predictions. If $\tau$ is greater, the teacher’s output becomes more uniformed, implying that the probability values are spread out more throughout all classes.

2.4 Embedding Encoding (ENC)

Embedding distillation was first proposed by [Mou et al., 2016] for NLP tasks. Unlike our framework, their method does not rely on teacher-student models, but rather directly trains a single model with an encoding layer inserted between the embedding layer and its upper layer in the network (Figure 2a). Each word $w_i$ is entered to an embedding layer $\phi$ that yields a large embedding vector $\phi(w_i)$. This vector is projected into a smaller embedding space by $W_{enc}$ with an activation function $f$. As a result, a smaller embedding vector $\phi'(w_i)$ is produced for $w_i$ as follows ($b_{enc}$: a bias for the projection):

$$\phi'(w_i) = f(W_{enc} \cdot \phi(w_i) + b_{enc})$$

The smaller embedding $\phi'(w_i)$ generated by this projection contains distilled knowledge from the larger embedding $\phi(w_i)$. The cost function of this method simply measures the cross-entropy between gold labels and the softmax output values.

3 Embedding Distillation

Our proposed embedding distillation framework begins by training a teacher model using the original embeddings. After training, the teacher model generates the corresponding logit value for each input in the training data. Then, a student model that comprises a projection layer is optimized for the logit (or softmax) values from the teacher model. After training the student model, small embeddings are distilled from the projection layer (Figure 3a).
The original large embeddings as well as weights in the projection layer are discarded for deployment such that the small embeddings can be referenced directly from the word indices in the student model during decoding (Figure 3b). Such distillation significantly reduces the model size and computations in the network, which is welcomed in production.

### 3.1 Distillation via Teacher-Student Models

Projecting a vector into a lower dimensional space generally entails information loss, although it does not have to be the case under two conditions. First, the source embeddings comprise both relevant and irrelevant information for the target task, therefore, there is a room to discard the irrelevant information. Second, the projection layer in the target network is capable of preserving the relevant information from the source embeddings. The first condition is met for NLP because most vector space models such as Word2Vec [Mikolov et al., 2013], Glove [Pennington et al., 2014], or FastText [Bojanowski et al., 2017] are trained on a vast amount of text, where only a small portion is germane to a specific task.

To meet the second condition, teacher-student models are adapted to our distillation framework, where the projection layer in a student model learns the relevant information for the target task from a teacher model. The output dimension of the projection layer is generally much smaller than the size of the embedding layer in a teacher model. Note that it is possible to integrate multiple projection layers in a student model; Experiment section discusses performance difference by adding different numbers of projection layers in the student model.

To illustrate this concept, consider the following scenario: Suppose we have a large pretrained model and a smaller model that we want to deploy but which requires less space and computation. We can use embedding distillation to transfer the core elements from the large embeddings to the smaller embeddings during training using gold labels. Only the small embeddings are kept for deployment, resulting in less space and fewer computations. Such distillation significantly reduces the model size and computations in the network, which is welcomed in production.

### 3.2 Projection Layer Initialization

Unlike [Mou et al., 2016] who randomly initialized the projection layer, it is initialized with vectors pre-trained by an autoencoder [Hinton and Zemel, 1994] in our framework. This initialization stabilizes and improves optimization of neural networks during training, resulting in more robust models.

### 4 Distillation Ensemble

Ensemble methods generally achieve higher accuracy than a standalone model; however, slow speed is expected due to the runs from multiple models in ensemble. This section presents a novel ensemble approach based on our distillation framework using logit matching that produces a light-weighted student model trained by multiple teachers (Figure 4).

The premise of this approach is that it is possible to have multiple teacher models train a student model by combining their logit values such that the student no longer needs the teachers during decoding because it already learned “enough” from them during training. As a result, our ensemble approach ensures higher efficiency for the student model than for any of the teacher models during decoding. The following sections describe two different ensemble methods applied to our framework.

#### 4.1 Routing by Agreement Ensemble (RAE)

This method gives more weights to the majority, by adopting the dynamic routing algorithm presented by [Sabour et al., 2017]. It first collects the consensus of all teachers, then boosts weights of the teachers who strongly agree with that consensus whereas suppresses the influence of the teachers who do not agree as much. The procedure of calculating the representing logit is described in Algorithm 1.

The squash function in lines 4 and 9 is a non-linear activation function that ensures the norm of the output vector to be in [0, 1] [Sabour et al., 2017]. This vectorized activation is important for a routing algorithm because both magnitude and direction play an important role in the agreement calculation by the dot product, which enforces the consensus into the direction with strong confidence.
Algorithm 1: Get R. LOGIT for RAE and RDE

Input: Teachers’ logits \(Z \in \mathbb{R}^{T \times C}\), and an algorithm selector \(b \in \{\text{RAE, RDE}\}\)

1. **Output:** The representing logit, \(z^{rep} \in \mathbb{R}^C\).
2. \(k \leftarrow 1\) if \(b\) is RAE else \(-1\)
3. for \(t \in \{1, \ldots, T\}\) do
4. \(x^t \leftarrow \text{squash}(z^t), w_t \leftarrow 0\)
5. while \(n\) iterations do
6. \(c \leftarrow \text{softmax}(w)\)
7. \(z^{rep} = \sum_{t=1}^{T} c_t \cdot z^t\)
8. \(s \leftarrow \text{squash}(z^{rep})\)
9. if not last iteration then
10. for \(t \in \{1, \ldots, T\}\) do
11. \(w_t \leftarrow w_t + kx^t \cdot s\)
12. return \(z^{rep}\)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C</th>
<th>TRN</th>
<th>DEV</th>
<th>TST</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR [Pang and Lee, 2005]</td>
<td>2</td>
<td>7,684</td>
<td>990</td>
<td>1,988</td>
</tr>
<tr>
<td>SST-1 [Socher et al., 2013]</td>
<td>5</td>
<td>8,544</td>
<td>1,101</td>
<td>2,210</td>
</tr>
<tr>
<td>SST-2 [Socher et al., 2013]</td>
<td>2</td>
<td>6,920</td>
<td>872</td>
<td>1,821</td>
</tr>
<tr>
<td>Subj [Pang and Lee, 2004]</td>
<td>2</td>
<td>7,199</td>
<td>907</td>
<td>1,894</td>
</tr>
<tr>
<td>TREC [Li and Roth, 2002]</td>
<td>6</td>
<td>4,952</td>
<td>500</td>
<td>1,988</td>
</tr>
<tr>
<td>CR [Hu and Liu, 2004]</td>
<td>2</td>
<td>2,718</td>
<td>340</td>
<td>717</td>
</tr>
<tr>
<td>MPQA [Wiebe et al., 2005]</td>
<td>2</td>
<td>7,636</td>
<td>955</td>
<td>2,015</td>
</tr>
</tbody>
</table>

Table 1: Seven datasets used for our experiments. C: number of classes, TRN/DEV/TST: number of instances in training/development/evaluation set.

4.2 Routing by Disagreement Ensemble (RDE)

This method focuses on the minority vote instead, because minority opinions may cast important information for the task. The algorithm is the same as RAE, except for the sign of the weight update (line 2 in Algorithm 1).

5 Experiments

5.1 Datasets

All models are evaluated on seven document classification datasets in Table 1. MR, SST-* and CR are targeted at the task of sentiment analysis while Subj, TREC, and MPQA are targeted at the classifications of subjectivity, question types, and opinion polarity, respectively. About 10% of the training sets are split into development sets for SST-* and TREC, and about 10/20% of the provided resources are divided into development/evaluation sets for the other datasets, respectively.

5.2 Word Embeddings

For sentiment analysis, raw text from the Amazon Review dataset is used to train word embeddings, resulting 2.67M word vectors. For the other tasks, combined text from Wikipedia and the New York Times Annotated corpus are used, resulting 1.96M word vectors. For each group, two sets of embeddings are trained with dimensions of 50 and 400 by Word2Vec [Mikolov et al., 2013]. While training, default hyper-parameters are used without an explicit hyper-parameter tuning.

5.3 Network Configuration

Two types of teacher models are developed using Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM); comparing different types of teacher models provides more generalized insights for our distillation framework. All teacher models use 400 dimensional word embeddings, and all student models are based on CNN. The CNN-based teacher and student models share the followings: filter sizes = [2, 3, 4, 5], # of filters = 32, dimension of the hidden layer right below the softmax layer = 50. Teacher models add a dropout of 0.8 to the hidden layer, whereas student models add dropouts of 0.1 to both the hidden layer and the the projection layer. On the other hand, the LSTM-based teacher models use two bidirectional LSTM layers. Only the last output vector is fed into the 50 dimensional hidden layer, which becomes the input to the softmax layer. A dropout of 0.2 is applied to all hidden layers, both in and out of the LSTM. For both CNN and LSTM ensembles, all 10 teachers share the same model structures with different initializations. Although this limited teacher diversity, our ensemble method produces remarkably good results (Section 5.7).

Each student model may consist of one or two projection layers. The one-layered projection adds one 50 dimensional hidden layer above the embedding layer that transfers core knowledge from the original embeddings to the distilled embeddings. The two-layered projection comprises two layers; the size of the lower layer is the same as the size of teacher’s embedding layer, 400, and the size of the upper layer is 50, which is the dimension of the distilled embeddings. This two-layered projection is empirically found to be more robust than various combinations of network architectures including wider and deeper layers from our experiments.

5.4 Pre-trained Weights

An autoencoder comprising a 50-dimensional encoder and a 400-dimensional decoder is used to pre-train weights for the two-layered projection in student models, where the encoder and the decoder have the same and inversed shapes as the upper and lower layers of the projection. Note that results by using pre-trained weights for the one-layered projection are not reported in Table 2 due to the limited space, but we consistently see robust improvement using pre-trained weights for the projection layers.

5.5 Embedding Distillation

Six models are evaluated on the seven datasets in Table 1 to show the effectiveness of our embedding distillation framework: logit matching (LM), noisy logit matching (NLM), softmax tau matching (STM) models with teacher-student

3http://snap.stanford.edu/data/web-Amazon.html

4Among convolutional, relational, and dense-networks with different configurations, the dense-network with the reported configuration produces the best results.
Table 2: Results from our embedding distillation models on the evaluation sets in Table 1 using the CNN-based (the rows 5-10) and the LSTM-based teacher models (rows 13-18) along with the teacher models (*-400), baseline model (*-50) and previous distillation model (ENC). All models are tuned on the development sets and the best performing models are tested on the evaluation sets. Since neural models produce different results at any training due to the random initialization, five models are developed for each approach to avoid (un)lucky peaks, except for *-400 and *-50 where the results are achieved by selecting the best models among ten trials on the development sets. Each score is based on these five trials and represented as a pair of [Average ± Standard Deviation].

![Figure 5: Similarity distributions among sentiment word pairs. Blue and Red colors distinguish histograms from the original and distilled embeddings, respectively. The solid and dashed lines show the distributions from all word pairs regardless of their sentiments. Circles are for sentimentally similar word pairs, that are similarities across positive and negative word pairs such that \((w_i, w_j) \in \tilde{P}_c \times P_c\). Crosses are for sentimentally opposite word pairs, that are similarities across positive and negative word pairs such that \((w_i, w_j) \in P_c \times N_c\).](image)

Based on the table and figure, we can observe that the proposed embedding distillation framework can successfully transfer the core knowledge from the original embeddings with respect to their sentiments. Two models highlights the strength of our distilled models, significantly outperforming them with the same dimensional word embeddings (50-dim).

Table 2 shows the results achieved by all models. While the two existing models, *-50 and ENC, show marginal differences, our proposed models, *+TSED, outperform the previous embedding distillation SOTA (ENC). Our final models, *+PT+2L, outperform all the other models, reaching similar (or even higher) accuracy to the teacher models (*-400), which confirms that the proposed embedding distillation framework can successfully transfer the core knowledge from the original embeddings with respect to the target tasks, independent from the network structures of the teacher models.

The fact that the best model for each task comes from a different teacher-student strategy appears to be random. However, considering that it is the nature of neural network models whose accuracy deviates at every training, this is not surprising although it signifies the need of ensemble approaches to take advantage of multiple teachers and produce a more robust student model.

5.6 Lexical Analysis

Embedding distillation for a task such as sentiment analysis can be viewed as a vector transformation that adjusts similarities between word embeddings with respect to their sentiments. Thus, we hypothesize that distilled word embeddings from sentiment analysis should bring similar sentiment words together while disperse opposite ones in vector space.

To verify this, lexical analysis on both the original and the distilled word embeddings is conducted using the four sentiment datasets: SST-1, SST-2, MR, and CR. First, positive and negative words are collected from two publicly available lexicons, the MaxDiff Twitter Sentiment Lexicon [Kiritchenko et al., 2014] and the Bing Liu Opinion Lexicon [Hu and Liu, 2004]. Then, two groups of sentiment word sets are con-
structured, \((P_t, N_t)\) and \((P_o, N_o)\), where \(P_s\) and \(N_s\) compose positive and negative words, and \(w_1\) and \(w_2\) are collected from the Twitter and the Opinion lexicons, respectively. Next, the intersection between each type of sentiment word sets is found, that are \(P_t = P_t \cap P_o\) and \(N_t = N_t \cap N_o\), where \(|P_t| = 72\) and \(|N_t| = 89\). Finally, the intersections between \(P_o\) and the vocabulary set from each of the four sentiment datasets are found (e.g., \(P_t = A_k \cap P_o\), where \(k\) is one of the four datasets and \(A_k\) is the set of all words in \(k\)):

- \(\ell = \text{SST} \Rightarrow |P_\ell| = 66 \) and \(|N_\ell| = 83\).
- \(\ell = \text{SST} \Rightarrow |P_\ell| = 67 \) and \(|N_\ell| = 83\).
- \(\ell = \text{CR} \Rightarrow |P_\ell| = 19 \) and \(|N_\ell| = 57\).

For each \(\ell\), cosine similarities are measured for all possible word pairs \((w_1, w_2) \in P_N \times P_N\) where \(P_N = P_t \cup N_t\), using the original and distilled embeddings. Figure 5 illustrates the similarity distributions. It is clear that similar sentiment words generally give high similarity scores with the distilled embeddings (the plots drawn by red circles), whereas opposite sentiment words give low similarity scores (the plots drawn by red crosses). On the other hand, low similarity scores are shown for any case with the original embeddings (the plots drawn by blue circles and crosses). The normal distributions derived from the distilled embeddings (the red lines) are more symmetric and spread than the ones from the original embeddings (the blue lines), implying that distilled embeddings are more stable for the target task.

### 5.7 Distillation Ensemble

All ensemble models are based on our distillation framework using logit matching (LM) where the teacher models compose of 10 CNN-based or 10 LSTM-based models. Two ensemble methods are evaluated: Routing by Agreement (RAE) and Routing by Disagreement (RDE), and Figure 6 shows comparisons between these ensemble models against the teacher models.

The most notable finding is that RDE significantly outperforms the teacher, if the dataset is big. For example, RDE outperforms the teacher models on average, except for CR and MPQA, whose training set is relatively small (Table 1). The insight behind this trend might be that if there are many data samples, then the probability of exploring different knowledge from minor opinions could be increased, which would positively affect to the ensemble.

### 6 Conclusion

This paper proposes a new embedding distillation framework based on several teacher-student methods. Our experiments show that the proposed distillation models outperform the previous distillation model and give compatible accuracy to the teacher models, yet they are significantly smaller. Lexical analysis on sentiments reveals the comprehensiveness of the distilled embeddings. Moreover, a novel distillation ensemble approach is proposed, which shows huge advantage in both speed and accuracy over any teacher model. Our distillation ensemble approach consistently shows more robust results when the size of training data is sufficiently large.
References


