Convolutional Neural Networks for Sentence Classification

Yoon Kim
New York University
yhk255@nyu.edu

Abstract
We report on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. We show that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance. We additionally propose a simple modification to the architecture to allow for the use of both task-specific and static vectors. The CNN models discussed herein improve upon the state of the art on 4 out of 7 tasks, which include sentiment analysis and question classification.

1 Introduction
Deep learning models have achieved remarkable results in computer vision (Krizhevsky et al., 2012) and speech recognition (Graves et al., 2013) in recent years. Within natural language processing, much of the work with deep learning methods has involved learning word vector representations through neural language models (Bengio et al., 2003; Yih et al., 2011; Mikolov et al., 2013) and performing composition over the learned word vectors for classification (Collobert et al., 2011). Word vectors, wherein words are projected from a sparse, 1-of-\(V\) encoding (here \(V\) is the vocabulary size) onto a lower dimensional vector space via a hidden layer, are essentially feature extractors that encode semantic features of words in their dimensions. In such dense representations, semantically close words are likewise close—in euclidean or cosine distance—in the lower dimensional vector space.

Convolutional neural networks (CNN) utilize layers with convolving filters that are applied to local features (LeCun et al., 1998). Originally invented for computer vision, CNN models have subsequently been shown to be effective for NLP and have achieved excellent results in semantic parsing (Yih et al., 2014), search query retrieval (Shen et al., 2014), sentence modeling (Kalchbrenner et al., 2014), and other traditional NLP tasks (Collobert et al., 2011).

In the present work, we train a simple CNN with one layer of convolution on top of word vectors obtained from an unsupervised neural language model. These vectors were trained by Mikolov et al. (2013) on 100 billion words of Google News, and are publicly available.\(^1\) We initially keep the word vectors static and learn only the other parameters of the model. Despite little tuning of hyperparameters, this simple model achieves excellent results on multiple benchmarks, suggesting that the pre-trained vectors are ‘universal’ feature extractors that can be utilized for various classification tasks. Learning task-specific vectors through fine-tuning results in further improvements. We finally describe a simple modification to the architecture to allow for the use of both pre-trained and task-specific vectors by having multiple channels.

Our work is philosophically similar to Razavian et al. (2014) which showed that for image classification, feature extractors obtained from a pre-trained deep learning model perform well on a variety of tasks—including tasks that are very different from the original task for which the feature extractors were trained.

2 Model
The model architecture, shown in figure 1, is a slight variant of the CNN architecture of Collobert et al. (2011). Let \(x_i \in \mathbb{R}^k\) be the \(k\)-dimensional word vector corresponding to the \(i\)-th word in the sentence. A sentence of length \(n\) (padded where

\(^1\)https://code.google.com/p/word2vec/
Figure 1: Model architecture with two channels for an example sentence.

necessary) is represented as
\[ x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n, \]
where \( \oplus \) is the concatenation operator. In general, let \( x_{i:i+j} \) refer to the concatenation of words \( x_i, x_{i+1}, \ldots, x_{i+j} \). A convolution operation involves a filter \( w \in \mathbb{R}^{hk} \), which is applied to a window of \( h \) words to produce a new feature. For example, a feature \( c_i \) is generated from a window of words \( x_{i:i+h-1} \) by
\[ c_i = f(w \cdot x_{i:i+h-1} + b). \]
Here \( b \in \mathbb{R} \) is a bias term and \( f \) is a non-linear function such as the hyperbolic tangent. This filter is applied to each possible window of words in the sentence \( \{x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}\} \) to produce a feature map
\[ c = [c_1, c_2, \ldots, c_{n-h+1}], \]
with \( c \in \mathbb{R}^{n-h+1} \). We then apply a max-over-time pooling operation (Collobert et al., 2011) over the feature map and take the maximum value \( \hat{c} = \max(c) \) as the feature corresponding to this particular filter. The idea is to capture the most important feature—one with the highest value—for each feature map. This pooling scheme naturally deals with variable sentence lengths.

We have described the process by which one feature is extracted from one filter. The model uses multiple filters (with varying window sizes) to obtain multiple features. These features form the penultimate layer and are passed to a fully connected softmax layer whose output is the probability distribution over labels.

In one of the model variants, we experiment with having two ‘channels’ of word vectors—one that is kept static throughout training and one that is fine-tuned via backpropagation (section 3.2). In the multichannel architecture, illustrated in figure 1, each filter is applied to both channels and the results are added to calculate \( c_i \) in equation (2). The model is otherwise equivalent to the single channel architecture.

2.1 Regularization
For regularization we employ dropout on the penultimate layer with a constraint on \( l_2 \)-norms of the weight vectors (Hinton et al., 2012). Dropout prevents co-adaptation of hidden units by randomly dropping out—i.e., setting to zero—a proportion \( p \) of the hidden units during forward-backpropagation. That is, given the penultimate layer \( z = [\hat{c}_1, \ldots, \hat{c}_m] \) (note that here we have \( m \) filters), instead of using
\[ y = w \cdot z + b \]
for output unit \( y \) in forward propagation, dropout uses
\[ y = w \cdot (z \circ r) + b, \]
where \( \circ \) is the element-wise multiplication operator and \( r \in \mathbb{R}^m \) is a ‘masking’ vector of Bernoulli random variables with probability \( p \) of being 1. Gradients are backpropagated only through the unmasked units. At test time, the learned weight vectors are scaled by \( p \) such that \( \tilde{w} = pw \), and \( \tilde{w} \) is used (without dropout) to score unseen sentences. We additionally constrain \( l_2 \)-norms of the weight vectors by rescaling \( w \) to have \( ||w||_2 = s \) whenever \( ||w||_2 > s \) after a gradient descent step.

\(^5\)We employ language from computer vision where a color image has red, green, and blue channels.
Table 1: Summary statistics for the datasets after tokenization. c: Number of target classes. l: Average sentence length. N: Dataset size. |V|: Vocabulary size. |Vpre|: Number of words present in the set of pre-trained word vectors. Test: Test set size (CV means there was no standard train/test split and thus 10-fold CV was used).

3 Datasets and Experimental Setup

We test our model on various benchmarks. Summary statistics of the datasets are in table 1.

- MR: Movie reviews with one sentence per review. Classification involves detecting positive/negative reviews (Pang and Lee, 2005).

- SST-1: Stanford Sentiment Treebank—an extension of MR but with train/dev/test splits provided and fine-grained labels (very positive, positive, neutral, negative, very negative), re-labeled by Socher et al. (2013).

- SST-2: Same as SST-1 but with neutral reviews removed and binary labels.

- Subj: Subjectivity dataset where the task is to classify a sentence as being subjective or objective (Pang and Lee, 2004).

- TREC: TREC question dataset—task involves classifying a question into 6 question types (whether the question is about person, location, numeric information, etc.) (Li and Roth, 2002).

- CR: Customer reviews of various products (cameras, MP3s etc.). Task is to predict positive/negative reviews (Hu and Liu, 2004).

- MPQA: Opinion polarity detection subtask of the MPQA dataset (Wiebe et al., 2005).

3.1 Hyperparameters and Training

For all datasets we use: rectified linear units, filter windows \((h)\) of 3, 4, 5 with 100 feature maps each, dropout rate \((p)\) of 0.5, \(l_2\) constraint \((s)\) of 3, and mini-batch size of 50. These values were chosen via a grid search on the SST-2 dev set.

We do not otherwise perform any dataset-specific tuning other than early stopping on dev sets. For datasets without a standard dev set we randomly select 10% of the training data as the dev set. Training is done through stochastic gradient descent over shuffled mini-batches with the Adadelta update rule (Zeiler, 2012).

3.2 Pre-trained Word Vectors

Initializing word vectors with those obtained from an unsupervised neural language model is a popular method to improve performance in the absence of a large supervised training set (Collobert et al., 2011; Socher et al., 2011; Iyyer et al., 2014). We use the publicly available word2vec vectors that were trained on 100 billion words from Google News. The vectors have dimensionality of 300 and were trained using the continuous bag-of-words architecture (Mikolov et al., 2013). Words not present in the set of pre-trained words are initialized randomly.

3.3 Model Variations

We experiment with several variants of the model.

- CNN-rand: Our baseline model where all words are randomly initialized and then modified during training.

- CNN-static: A model with pre-trained vectors from word2vec. All words—including the unknown ones that are randomly initialized—are kept static and only the other parameters of the model are learned.

- CNN-non-static: Same as above but the pre-trained vectors are fine-tuned for each task.

- CNN-multichannel: A model with two sets of word vectors. Each set of vectors is treated as a ‘channel’ and each filter is applied

---

1https://www.cs.cornell.edu/people/pabo/movie-review-data/
2http://nlp.stanford.edu/sentiment/ Data is actually provided at the phrase-level and hence we train the model on both phrases and sentences but only score on sentences at test time, as in Socher et al. (2013), Kalchbrenner et al. (2014), and Le and Mikolov (2014). Thus the training set is an order of magnitude larger than listed in table 1.
3http://cogcomp.cs.illinois.edu/Data/QA/QC/
4http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
5http://www.cs.pitt.edu/mpqa/

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>SST-1</th>
<th>SST-2</th>
<th>Subj</th>
<th>TREC</th>
<th>CR</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-rand</td>
<td>76.1</td>
<td>45.0</td>
<td>82.7</td>
<td>89.6</td>
<td>91.2</td>
<td>79.8</td>
<td>83.4</td>
</tr>
<tr>
<td>CNN-static</td>
<td>81.0</td>
<td>45.5</td>
<td>86.8</td>
<td>93.0</td>
<td>92.8</td>
<td>84.7</td>
<td>89.6</td>
</tr>
<tr>
<td>CNN-non-static</td>
<td>81.5</td>
<td>48.0</td>
<td>87.2</td>
<td>93.4</td>
<td>93.6</td>
<td>84.3</td>
<td>89.5</td>
</tr>
<tr>
<td>CNN-multichannel</td>
<td>81.1</td>
<td>47.4</td>
<td>88.1</td>
<td>93.2</td>
<td>92.2</td>
<td>85.0</td>
<td>89.4</td>
</tr>
<tr>
<td>RAEE (Socher et al., 2011)</td>
<td>77.7</td>
<td>43.2</td>
<td>82.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>86.4</td>
</tr>
<tr>
<td>MV-RNN (Socher et al., 2012)</td>
<td>79.0</td>
<td>44.4</td>
<td>82.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RNTN (Socher et al., 2013)</td>
<td>-</td>
<td>45.7</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DCNN (Kalchbrenner et al., 2014)</td>
<td>-</td>
<td>48.5</td>
<td>86.8</td>
<td>-</td>
<td>93.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Paragraph-Vec (Le and Mikolov, 2014)</td>
<td>-</td>
<td>48.7</td>
<td>87.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CCAE (Hermann and Blunsom, 2013)</td>
<td>77.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.2</td>
</tr>
<tr>
<td>Sent-Parser (Dong et al., 2014)</td>
<td>79.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>86.3</td>
</tr>
<tr>
<td>NBSVM (Wang and Manning, 2012)</td>
<td>79.4</td>
<td>-</td>
<td>93.2</td>
<td>-</td>
<td>81.8</td>
<td>86.3</td>
<td>-</td>
</tr>
<tr>
<td>MNB (Wang and Manning, 2012)</td>
<td>79.0</td>
<td>-</td>
<td>-</td>
<td>93.6</td>
<td>-</td>
<td>80.0</td>
<td>86.3</td>
</tr>
<tr>
<td>G-Dropout (Wang and Manning, 2013)</td>
<td>79.0</td>
<td>-</td>
<td>93.4</td>
<td>-</td>
<td>82.1</td>
<td>86.1</td>
<td>-</td>
</tr>
<tr>
<td>F-Dropout (Wang and Manning, 2013)</td>
<td>79.1</td>
<td>-</td>
<td>-</td>
<td>93.6</td>
<td>-</td>
<td>81.9</td>
<td>86.3</td>
</tr>
<tr>
<td>Tree-CRF (Nakagawa et al., 2010)</td>
<td>77.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>81.4</td>
<td>86.1</td>
<td>-</td>
</tr>
<tr>
<td>CRF-PR (Yang and Cardie, 2014)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM$_S$ (Silva et al., 2011)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>95.0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In order to disentangle the effect of the above variations versus other random factors, we eliminate other sources of randomness—CV-fold assignment, initialization of unknown word vectors, initialization of CNN parameters—by keeping them uniform within each dataset.

4 Results and Discussion

Results of our models against other methods are listed in table 2. Our baseline model with all randomly initialized words (CNN-rand) does not perform well on its own. While we had expected performance gains through the use of pre-trained vectors, we were surprised at the magnitude of the gains. Even a simple model with static vectors (CNN-static) performs remarkably well, giving competitive results against the more sophisticated deep learning models that utilize complex pooling schemes (Kalchbrenner et al., 2014) or require parse trees to be computed beforehand (Socher et al., 2013). These results suggest that the pre-trained vectors are good, ‘universal’ feature extractors and can be utilized across datasets. Fine-tuning the pre-trained vectors for each task gives still further improvements (CNN-non-static).

4.1 Multichannel vs. Single Channel Models

We had initially hoped that the multichannel architecture would prevent overfitting (by ensuring that the learned vectors do not deviate too far from the original values) and thus work better than the single channel model, especially on smaller datasets. The results, however, are mixed, and further work on regularizing the fine-tuning process is warranted. For instance, instead of using an additional channel for the non-static portion, one could maintain a single channel but employ extra dimensions that are allowed to be modified during training.


<table>
<thead>
<tr>
<th></th>
<th>Static Channel</th>
<th>Non-static Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bad</strong></td>
<td>good</td>
<td>terrible</td>
</tr>
<tr>
<td></td>
<td>terrible</td>
<td>horrible</td>
</tr>
<tr>
<td></td>
<td>horrible</td>
<td>lousy</td>
</tr>
<tr>
<td></td>
<td>lousy</td>
<td>stupid</td>
</tr>
<tr>
<td><strong>good</strong></td>
<td>great</td>
<td>nice</td>
</tr>
<tr>
<td></td>
<td>bad</td>
<td>decent</td>
</tr>
<tr>
<td></td>
<td>terrific</td>
<td>solid</td>
</tr>
<tr>
<td></td>
<td>decent</td>
<td>terrific</td>
</tr>
<tr>
<td><strong>n’t</strong></td>
<td>os</td>
<td>not</td>
</tr>
<tr>
<td></td>
<td>ca</td>
<td>never</td>
</tr>
<tr>
<td></td>
<td>ireland</td>
<td>nothing</td>
</tr>
<tr>
<td></td>
<td>wo</td>
<td>neither</td>
</tr>
<tr>
<td><strong>!</strong></td>
<td>2,500</td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td>entire</td>
<td>lush</td>
</tr>
<tr>
<td></td>
<td>jez</td>
<td>beautiful</td>
</tr>
<tr>
<td></td>
<td>changer</td>
<td>terrific</td>
</tr>
<tr>
<td><strong>,’</strong></td>
<td>decasia</td>
<td>but</td>
</tr>
<tr>
<td></td>
<td>abysmally</td>
<td>dragon</td>
</tr>
<tr>
<td></td>
<td>demise</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>valiant</td>
<td>and</td>
</tr>
</tbody>
</table>

Table 3: Top 4 neighboring words—based on cosine similarity—for vectors in the static channel (left) and fine-tuned vectors in the non-static channel (right) from the multichannel model on the SST-2 dataset after training.

4.2 Static vs. Non-static Representations

As is the case with the single channel non-static model, the multichannel model is able to fine-tune the non-static channel to make it more specific to the task-at-hand. For example, good is most similar to bad in word2vec, presumably because they are (almost) syntactically equivalent. But for vectors in the non-static channel that were fine-tuned on the SST-2 dataset, this is no longer the case (table 3). Similarly, good is arguably closer to nice than it is to great for expressing sentiment, and this is indeed reflected in the learned vectors.

For (randomly initialized) tokens not in the set of pre-trained vectors, fine-tuning allows them to learn more meaningful representations: the network learns that exclamation marks are associated with effusive expressions and that commas are conjunctive (table 3).

4.3 Further Observations

We report on some further experiments and observations:

- Kalchbrenner et al. (2014) report much worse results with a CNN that has essentially the same architecture as our single channel model. For example, their Max-TDNN (Time Delay Neural Network) with randomly initialized words obtains 37.4% on the SST-1 dataset, compared to 45.0% for our model. We attribute such discrepancy to our CNN having much more capacity (multiple filter widths and feature maps).

  - Dropout proved to be such a good regularizer that it was fine to use a larger than necessary network and simply let dropout regularize it. Dropout consistently added 2%–4% relative performance.

  - When randomly initializing words not in word2vec, we obtained slight improvements by sampling each dimension from $U[-a, a]$ where $a$ was chosen such that the randomly initialized vectors have the same variance as the pre-trained ones. It would be interesting to see if employing more sophisticated methods to mirror the distribution of pre-trained vectors in the initialization process gives further improvements.

- We briefly experimented with another set of publicly available word vectors trained by Collobert et al. (2011) on Wikipedia, and found that word2vec gave far superior performance. It is not clear whether this is due to Mikolov et al. (2013)’s architecture or the 100 billion word Google News dataset.

- Adadelta (Zeiler, 2012) gave similar results to Adagrad (Duchi et al., 2011) but required fewer epochs.

5 Conclusion

In the present work we have described a series of experiments with convolutional neural networks built on top of word2vec. Despite little tuning of hyperparameters, a simple CNN with one layer of convolution performs remarkably well. Our results add to the well-established evidence that unsupervised pre-training of word vectors is an important ingredient in deep learning for NLP.

Acknowledgments

We would like to thank Yann LeCun and the anonymous reviewers for their helpful feedback and suggestions.
References


