Vector of Locally-Aggregated Word Embeddings (VLAWE): A Novel Document-level Representation

Radu Tudor Ionescu and **Andrei M. Butnaru**
University of Bucharest
Department of Computer Science
14 Academiei, Bucharest, Romania
raducu.ionescu@gmail.com
butnaruandreimadalin@gmail.com

Abstract

In this paper, we propose a novel representation for text documents based on aggregating word embedding vectors into document embeddings. Our approach is inspired by the Vector of Locally-Aggregated Descriptors used for image representation, and it works as follows. First, the word embeddings gathered from a collection of documents are clustered by k-means in order to learn a codebook of semnatically-related word embeddings. Each word embedding is then associated to its nearest cluster centroid (codeword). The Vector of Locally-Aggregated Word Embeddings (VLAWE) representation of a document is then computed by accumulating the differences between each codeword vector and each word vector (from the document) associated to the respective codeword. We plug the VLAWE representation, which is learned in an unsupervised manner, into a classifier and show that it is useful for a diverse set of text classification tasks. We compare our approach with a broad range of recent state-of-the-art methods, demonstrating the effectiveness of our approach. Furthermore, we obtain a considerable improvement on the Movie Review data set, reporting an accuracy of 93.3%, which represents an absolute gain of 10% over the state-of-the-art approach. Our code is available at https://github.com/raduionescu/vlawe-boswe/.

1 Introduction

In recent years, word embeddings (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014) have had a huge impact in natural language processing (NLP) and related fields, being used in many tasks including sentiment analysis (Dos Santos and Gatti, 2014; Fu et al., 2018), information retrieval (Clinchant and Perronnin, 2013; Ye et al., 2016) and word sense disambiguation (Bhingardive et al., 2015; Butnaru et al., 2017; Chen et al., 2014; Iacobacci et al., 2016), among many others. Starting from word embeddings, researchers proposed various ways of aggregating word embedding vectors to obtain efficient sentence-level or document-level representations (Butnaru and Ionescu, 2017; Cheng et al., 2018; Clinchant and Perronnin, 2013; Conneau et al., 2017; Cozma et al., 2018; Fu et al., 2018; Hill et al., 2016; Kiros et al., 2015; Kusner et al., 2015; Le and Mikolov, 2014; Shen et al., 2018; Torki, 2018; Zhao et al., 2015; Zhou et al., 2016, 2018). Although the mean (or sum) of word vectors is commonly adopted because of its simplicity (Mitchell and Lapata, 2010), it seems that more complex approaches usually yield better performance (Cheng et al., 2018; Conneau et al., 2017; Cozma et al., 2018; Fu et al., 2018; Hill et al., 2016; Kiros et al., 2015; Torki, 2018; Zhao et al., 2015; Zhou et al., 2016, 2018). To this end, we propose a simple yet effective approach for aggregating word embeddings into document embeddings. Our approach is inspired by the Vector of Locally-Aggregated Descriptors (VLAD) (Jégou et al., 2010, 2012) used in computer vision to efficiently represent images for various image classification and retrieval tasks. To our knowledge, we are the first to adapt and use VLAD in the text domain.

Our document-level representation is constructed as follows. First, we apply a pre-trained word embedding model, such as *GloVe* (Pennington et al., 2014), on all the words from a set of training documents in order to obtain a set of training word vectors. The word vectors are clustered by k-means in order to learn a codebook of semnatically-related word embeddings. Each word embedding is then associated to its nearest cluster centroid (codeword). The Vector of Locally-Aggregated Word Embeddings (VLAWE) representation of a text document is then com-

Proceedings of NAACL-HLT 2019, pages 363–369
Minneapolis, Minnesota, June 2 - June 7, 2019.

puted by accumulating the differences between each codeword vector and each word vector that is both present in the document and associated to the respective codeword. Since our approach considers cluster centroids as reference for building the representation, it can easily accommodate new words, not seen during k-means training, simply by associating them to the nearest cluster centroids. Thus, VLAWE is robust to vocabulary distribution gaps between training and test, which can appear when the training set is particularly smaller or from a different domain. Certainly, the robustness holds as long as the word embeddings are pre-trained on a very large set of documents, e.g. the entire Wikipedia.

We plug the VLAWE representation, which is learned in an unsupervised manner, into a classifier, namely Support Vector Machines (SVM), and show that it is useful for a diverse set of text classification tasks. We consider five benchmark data sets: Reuters-21578 (Lewis, 1997), RT-2k (Pang and Lee, 2004), MR (Pang and Lee, 2005), TREC (Li and Roth, 2002) and Subj (Pang and Lee, 2004). We compare VLAWE with recent state-of-the-art methods (Butnaru and Ionescu, 2017; Cheng et al., 2018; Fu et al., 2018; Hill et al., 2016; Iyyer et al., 2015; Kim, 2014; Kiros et al., 2015; Le and Mikolov, 2014; Liu et al., 2017; Shen et al., 2018; Torki, 2018; Xue and Zhou, 2009; Zhao et al., 2015; Zhou et al., 2016, 2018), demonstrating the effectiveness of our approach. Furthermore, we obtain a considerable improvement on the Movie Review (MR) data set, surpassing the state-of-the-art approach of Cheng et al. (2018) by almost 10%.

The rest of the paper is organized as follows. We present related works on learning document-level representations in Section 2. We describe the Vector of Locally-Aggregated Word Embeddings in Section 3. We present experiments and results on various text classification tasks in Section 4. Finally, we draw our conclusion in Section 5.

2 Related Work

There are various works (Butnaru and Ionescu, 2017; Cheng et al., 2018; Conneau et al., 2017; Fu et al., 2018; Hill et al., 2016; Iyyer et al., 2015; Kim, 2014; Kiros et al., 2015; Kusner et al., 2015; Le and Mikolov, 2014; Clinchant and Perronnin, 2013; Shen et al., 2018; Torki, 2018; Zhao et al., 2015; Zhou et al., 2018) that propose to build effective sentence-level or document-level representations based on word embeddings. While most of these approaches are based on deep learning (Cheng et al., 2018; Conneau et al., 2017; Hill et al., 2016; Iyyer et al., 2015; Kim, 2014; Kiros et al., 2015; Le and Mikolov, 2014; Zhao et al., 2015; Zhou et al., 2018), there have been some approaches that are inspired by computer vision research, namely by the bag-of-visual-words (Butnaru and Ionescu, 2017) and by Fisher Vectors (Clinchant and Perronnin, 2013). The relationship between the bag-of-visual-words, Fisher Vectors and VLAD is discussed in (Jégou et al., 2012). The discussion can be transferred to describe the relantionship of our work and the closely-related works of Butnaru and Ionescu (2017) and Clinchant and Perronnin (2013).

3 Method

The Vector of Locally-Aggregated Descriptors (VLAD) (Jégou et al., 2010, 2012) was introduced in computer vision to efficiently represent images for various image classification and retrieval tasks. We propose to adapt the VLAD representation in order to represent text documents instead of images. Our adaptation consists of replacing the Scale-Invariant Feature Transform (SIFT) image descriptors (Lowe, 2004) useful for recognizing object patterns in images with word embeddings (Mikolov et al., 2013; Pennington et al., 2014) useful for recognizing semantic patterns in text documents. We coin the term *Vector of Locally-Aggregated Word Embeddings (VLAWE)* for the resulting document representation.

The VLAWE representation is derived as follows. First, each word in the collection of training documents is represented as a word vector using a pre-trained word embeddings model. The result is a set $X = \{x_1, x_2, ..., x_n\}$ of n word vectors. As for the VLAD model, the next step is to learn a codebook $\{\mu_1, \mu_2, ..., \mu_k\}$ of representative meta-word vectors (codewords) using k-means. Each codeword μ_i is the centroid of the cluster $C_i \subset X$:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_t \in C_i} x_t, \forall i \in \{1, 2, ..., k\}, \quad (1)$$

where $|C_i|$ is the number of word vectors assigned to cluster C_i and k is the number of clusters. Since word embeddings carry semantic information by projecting semantically-related words in the same region of the embedding space, it means that the resulting clusters contain semantically-

related words. The formed centroids are stored in a randomized forest of k-d trees to reduce search cost, as described in (Philbin et al., 2007; Ionescu et al., 2013; Ionescu and Popescu, 2014, 2015a). Each word embedding x_t is associated to a single cluster C_i, such that the Euclidean distance between x_t and the corresponding codeword μ_i is minimum, for all $i \in \{1, 2, ..., k\}$. For each document D and each codeword μ_i, the differences $x_t - \mu_i$ of the vectors $x_t \in C_i \cap D$ and the codeword μ_i are accumulated into column vectors:

$$v_{i,D} = \sum_{x_t \in C_i \cap D} x_t - \mu_i, \quad (2)$$

where $D \subset X$ is the set of word embeddings in a given text document. The final VLAWE embedding for a given document D is obtained by stacking together the d-dimensional residual vectors $v_{i,D}$, where d is equal to the dimension of the word embeddings:

$$\phi_D = \begin{bmatrix} v_{1,D} \\ v_{2,D} \\ \vdots \\ v_{k,D} \end{bmatrix}. \quad (3)$$

Therefore, the VLAWE document embedding is has $k \cdot d$ components.

The VLAWE vector ϕ_D undergoes two normalization steps. First, a power normalization is performed by applying the following operator independently on each component (element):

$$f(z) = sign(z) \cdot |z|^{\alpha}, \quad (4)$$

where $0 \leq \alpha \leq 1$ and $|z|$ is the absolute value of z. Since words in natural language follow the Zipf's law (Powers, 1998), it seems natural to apply the power normalization in order to reduce the influence of highly frequent words, e.g. common words or stopwords, which can corrupt the representation. As Jégou et al. (2012), we empirically observed that this step consistently improves the quality of the representation. The power normalized document embeddings are then L_2-normalized. After obtaining the normalized VLAWE representations, we employ a classification method to learn a discriminative model for each specific text classification task.

4 Experiments

4.1 Data Sets

We exhibit the performance of VLAWE on five public data sets: Reuters-21578 (Lewis, 1997), RT-2k (Pang and Lee, 2004), MR (Pang and Lee, 2005), TREC (Li and Roth, 2002) and Subj (Pang and Lee, 2004).

The Reuters-21578 data set (Lewis, 1997) contains articles collected from Reuters newswire. Following Joachims (1998) and Yang and Liu (1999), we select the categories (topics) that have at least one document in the training set and one in the test set, leading to a total of 90 categories. We use the ModeApte evaluation (Xue and Zhou, 2009), in which unlabeled documents are eliminated, leaving a total of 10787 documents. The collection is already divided into 7768 documents for training and 3019 documents for testing.

The RT-2k data set (Pang and Lee, 2004) consists of 2000 movie reviews taken from the IMDB movie review archives. There are 1000 positive reviews rated with four or five stars, and 1000 negative reviews rated with one or two stars. The task is to discriminate between positive and negative reviews.

The Movie Review (MR) data set (Pang and Lee, 2005) consists of 5331 positive and 5331 negative sentences. Each sentence is selected from one movie review. The task is to discriminate between positive and negative sentiment.

TREC (Li and Roth, 2002) is a question type classification data set, where questions are divided into 6 classes. The collection is already divided into 5452 questions for training and 500 questions for testing.

The Subjectivity (Subj) (Pang and Lee, 2004) data set contains 5000 objective and 5000 subjective sentences. The task is to classify a sentence as being either subjective or objective.

4.2 Evaluation and Implementation Details

In the experiments, we used the pre-trained word embeddings computed with the *GloVe* toolkit provided by Pennington et al. (2014). The pre-trained GloVe model contains 300-dimensional vectors for 2.2 million words and phrases. Most of the steps required for building the VLAWE representation, such as the k-means clustering and the randomized forest of k-d trees, are implemented using the VLFeat library (Vedaldi and Fulkerson, 2008). We set the number of clusters (size of the codebook) to $k = 10$, leading to a VLAWE representation of $k \cdot d = 10 \cdot 300 = 3000$ components. Similar to Jégou et al. (2012), we set $\alpha = 0.5$ for the power normalization step in Equation (4), which consistently leads to near-optimal results on all data sets. In the learning stage, we employ the

Method	Reuters-21578	RT-2k	MR	TREC	Subj
Average of word embeddings (baseline)	85.3	84.7	77.4	80.0	89.5
BOW (baseline)	86.5	84.1	77.1	89.3	89.3
TF + FA + CP + SVM (Xue and Zhou, 2009)	87.0	-	-	-	-
Paragraph vectors (Le and Mikolov, 2014)	-	-	74.8	91.8	90.5
CNN (Kim, 2014)	-	83.5	81.5	93.6	93.4
DAN (Iyyer et al., 2015)	-	-	80.1	-	-
Combine-skip (Kiros et al., 2015)	-	-	76.5	92.2	93.6
Combine-skip + NB (Kiros et al., 2015)	-	-	80.4	-	93.6
AdaSent (Zhao et al., 2015)	-	-	83.1	92.4	95.5
SAE + embs. (Hill et al., 2016)	-	-	73.2	80.4	89.8
SDAE + embs. (Hill et al., 2016)	-	-	74.6	78.4	90.8
FastSent + AE (Hill et al., 2016)	-	-	71.8	80.4	88.8
BLSTM (Zhou et al., 2016)	-	-	80.0	93.0	92.1
BLSTM-Att (Zhou et al., 2016)	-	-	81.0	93.8	93.5
BLSTM-2DCNN (Zhou et al., 2016)	-	-	82.3	96.1	94.0
DC-TreeLSTM (Liu et al., 2017)	-	-	81.7	93.8	93.7
BOSWE (Butnaru and Ionescu, 2017)	87.2	89.7	-	-	-
TreeNet (Cheng et al., 2018)	-	-	79.8	91.6	92.0
TreeNet-GloVe (Cheng et al., 2018)	-	-	83.6	96.1	95.9
BOMV (Fu et al., 2018)	-	90.2	-	-	90.9
SWEM-average (Shen et al., 2018)	-	-	77.6	92.2	92.5
SWEM-concat (Shen et al., 2018)	-	-	78.2	91.8	93.0
COV + Mean (Torki, 2018)	-	-	80.2	90.3	93.1
COV + BOW (Torki, 2018)	-	-	80.7	91.8	93.3
COV + Mean + BOW (Torki, 2018)	-	-	81.1	91.6	93.2
DARLM (Zhou et al., 2018)	-	-	83.2	96.0	94.1
VLAWE (ours)	89.3	94.1	93.3	94.2	95.0

Table 1: Performance results (in %) of our approach (VLAWE) versus several state-of-the-art methods (Butnaru and Ionescu, 2017; Cheng et al., 2018; Fu et al., 2018; Hill et al., 2016; Iyyer et al., 2015; Kim, 2014; Kiros et al., 2015; Le and Mikolov, 2014; Liu et al., 2017; Shen et al., 2018; Torki, 2018; Xue and Zhou, 2009; Zhao et al., 2015; Zhou et al., 2016, 2018) on the Reuters-21578, RT-2k, MR, TREC and Subj data sets.

Support Vector Machines (SVM) implementation provided by LibSVM (Chang and Lin, 2011). We set the SVM regularization parameter to $C = 1$ in all our experiments. In the SVM, we use the linear kernel. For optimal results, the VLAWE representation is combined with the BOSWE representation (Butnaru and Ionescu, 2017), which is based on the PQ kernel (Ionescu and Popescu, 2013, 2015b).

We follow the same evaluation procedure as Kiros et al. (2015) and Hill et al. (2016), using 10-fold cross-validation when a train and test split is not pre-defined for a given data set. As evaluation metrics, we employ the micro-averaged F_1 measure for the Reuters-21578 data set and the standard classification accuracy for the RT-2k, the MR, the TREC and the Subj data sets, in order to fairly compare with the related art.

4.3 Results

We compare VLAWE with several state-of-the-art methods (Butnaru and Ionescu, 2017; Cheng et al., 2018; Fu et al., 2018; Hill et al., 2016; Iyyer et al., 2015; Kim, 2014; Kiros et al., 2015; Le and Mikolov, 2014; Liu et al., 2017; Shen et al., 2018; Torki, 2018; Xue and Zhou, 2009; Zhao et al., 2015; Zhou et al., 2016, 2018) as well as two baseline methods, namely the average of word embeddings and the standard bag-of-words (BOW). The corresponding results are presented in Table 1.

First, we notice that our approach outperforms both baselines on all data sets, unlike other related methods (Le and Mikolov, 2014; Hill et al., 2016). In most cases, our improvements over the baselines are higher than 5%. On the Reuters-21578 data set, we surpass the closely-related approach of Butnaru and Ionescu (2017) by around 2%. On

Method	MR
VLAWE ($k = 2$)	93.0
VALWE (PCA)	93.2
VLAWE (full, $k = 10$)	93.3

Table 2: Performance results (in %) of the full VLAWE representation (with $k = 10$) versus two compact versions of VLAWE, obtained either by setting $k = 2$ or by applying PCA.

the RT-2k data set, we surpass the related works of Fu et al. (2018) and Butnaru and Ionescu (2017) by around 4%. To our knowledge, our accuracy of 94.1% on RT-2k (Pang and Lee, 2004) surpasses all previous results reported in literature. On the MR data set, we surpass most related works by more than 10%. To our knowledge, the best accuracy on MR reported in previous literature is 83.6%, and it is obtained by Cheng et al. (2018). We surpass the accuracy of Cheng et al. (2018) by almost 10%, reaching an accuracy of 93.3% using VLAWE. On the TREC data set, we reach the third best performance, after methods such as (Cheng et al., 2018; Zhou et al., 2016, 2018). Our performance on TREC is about 2% lower than the state-of-the-art accuracy of 96.1%. On the Subj data set, we obtain an accuracy of 95.0%. There are two state-of-the-art methods (Cheng et al., 2018; Zhao et al., 2015) reporting better performance on Subj. Compared to the best one of them (Cheng et al., 2018), our accuracy is 1% lower. Overall, we consider that our results are noteworthy.

4.4 Discussion

The k-means clustering algorithm and, on some data sets, the cross-validation procedure can induce accuracy variations due to the random choices involved. We have conducted experiments to determine how large are the accuracy variations. We observed that the accuracy can decrease by up to 1%, which does not bring any significant differences to the results reported in Table 1.

Even for a small number of clusters, e.g. $k = 10$, the VLAWE document representation can grow up to thousands of features, as the number of features is $k \cdot d$, where $d = 300$ is the dimensionality of commonly used word embeddings. However, there are several document-level representations that usually have a dimensionality much smaller than $k \cdot d$. Therefore, it is desirable to obtain a more compact VLAWE representation. We hereby propose two approaches that lead to more compact representations. The first one is simply based on reducing the number of clusters. By setting $k = 2$ for instance, we obtain a 600-dimensional representation. The second one is based on applying Principal Component Analysis (PCA), to reduce the dimension of the feature vectors. Using PCA, we propose to reduce the size of the VLAWE representation to 300 components. In Table 2, the resulting compact representations are compared against the full VLAWE representation on the MR data set. Although the compact VLAWE representations provide slightly lower results compared to the VLAWE representation based on 3000 components, we note that the differences are insignificant. Furthermore, both compact VLAWE representations are far above the state-of-the-art method (Cheng et al., 2018).

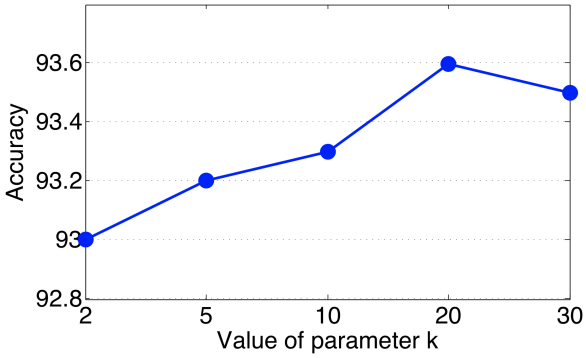

Figure 1: Accuracy on MR for different numbers of k-means clusters.

In Figure 1, we illustrate the performance variation on MR, when using different values for k. We notice that the accuracy tends to increase slightly, as we increase the number of clusters from 2 to 30. Overall, the VLAWE representation seems to be robust to the choice of k, always surpassing the state-of-the-art approach (Cheng et al., 2018).

5 Conclusion

We proposed a novel representation for text documents which is based on aggregating word embeddings using k-means and on computing the residuals between each word embedding allocated to a given cluster and the corresponding cluster centroid. Our experiments on five benchmark data sets prove that our approach yields competitive results with respect to the state-of-the-art methods.

Acknowledgments

We thank the reviewers for their useful comments. This research is supported by University of Bucharest, Faculty of Mathematics and Computer Science, through the 2019 Mobility Fund.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised Most Frequent Sense Detection using Word Embeddings. In *Proceedings of NAACL*, pages 1238–1243.

Andrei Butnaru and Radu Tudor Ionescu. 2017. From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings. In *Proceedings of KES*, pages 1784–1793.

Andrei Butnaru, Radu Tudor Ionescu, and Florentina Hristea. 2017. ShotgunWSD: An unsupervised algorithm for global word sense disambiguation inspired by DNA sequencing. In *Proceedings of EACL*, pages 916–926.

Chih-Chung Chang and Chih-Jen Lin. 2011. LibSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of EMNLP*, pages 1025–1035.

Zhou Cheng, Chun Yuan, Jiancheng Li, and Haiqin Yang. 2018. TreeNet: Learning Sentence Representations with Unconstrained Tree Structure. In *Proceedings of IJCAI*, pages 4005–4011.

Stéphane Clinchant and Florent Perronnin. 2013. Aggregating continuous word embeddings for information retrieval. In *Proceedings of CVSC Workshop*, pages 100–109.

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of ICML*, pages 160–167.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of EMNLP*, pages 670–680.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of ACL*, pages 503–509.

Cícero Nogueira Dos Santos and Maira Gatti. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING*, pages 69–78.

Mingsheng Fu, Hong Qu, Li Huang, and Li Lu. 2018. Bag of meta-words: A novel method to represent document for the sentiment classification. *Expert Systems with Applications*, 113:33–43.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *Proceedings of NAACL*, pages 1367–1377.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of ACL*, pages 897–907.

Radu Tudor Ionescu and Marius Popescu. 2013. Kernels for Visual Words Histograms. In *Proceedings of ICIAP*, pages 81–90.

Radu Tudor Ionescu and Marius Popescu. 2014. Objectness to improve the bag of visual words model. In *Proceedings of ICIP*, pages 3238–3242.

Radu Tudor Ionescu and Marius Popescu. 2015a. Have a SNAK. Encoding Spatial Information with the Spatial Non-alignment Kernel. In *Proceedings of ICIAP*, pages 97–108.

Radu Tudor Ionescu and Marius Popescu. 2015b. PQ kernel: a rank correlation kernel for visual word histograms. *Pattern Recognition Letters*, 55:51–57.

Radu Tudor Ionescu, Marius Popescu, and Cristian Grozea. 2013. Local Learning to Improve Bag of Visual Words Model for Facial Expression Recognition. In *Proceedings of WREPL*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of ACL*, pages 1681–1691.

Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *Proceedings of CVPR*, pages 3304–3311.

Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. 2012. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716.

Thorsten Joachims. 1998. Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In *Proceedings of ECML*, pages 137–142, London, UK, UK. Springer-Verlag.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of EMNLP*, pages 1746–1751.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Proceedings of NIPS*, pages 3294–3302.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of ICML*, pages 957–966.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of ICML*, pages 1188–1196.

David Lewis. 1997. The Reuters-21578 text categorization test collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of COLING*, pages 1–7.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Dynamic compositional neural networks over tree structure. In *Proceedings of IJCAI*, pages 4054–4060.

David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales. In *Proceedings of ACL*, pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*, pages 1532–1543.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of CVPR*, pages 1–8.

David Powers. 1998. Applications and explanations of Zipf's law. In *Proceedings of NeMLaP/CoNLL*, pages 151–160.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of ACL*, pages 440–450.

Marwan Torki. 2018. A Document Descriptor using Covariance of Word Vectors. In *Proceedings of ACL*, pages 527–532.

Andrea Vedaldi and B. Fulkerson. 2008. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/.

Xiao-Bing Xue and Zhi-Hua Zhou. 2009. Distributional features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 21(3):428–442.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR*, pages 42–49.

Xin Ye, Hui Shen, Xiao Ma, Răzvan Bunescu, and Chang Liu. 2016. From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of ICSE*, pages 404–415.

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-Adaptive Hierarchical Sentence Model. In *Proceedings of IJCAI*, pages 4069–4076.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. In *Proceedings of COLING*, pages 3485–3495.

Qianrong Zhou, Xiaojie Wang, and Xuan Dong. 2018. Differentiated attentive representation learning for sentence classification. In *Proceedings of IJCAI*, pages 4630–4636.