JOINT VERIFICATION-IDENTIFICATION IN END-TO-END MULTI-SCALE CNN FRAMEWORK FOR TOPIC IDENTIFICATION

Raghavendra Pappagari, Jesús Villalba, Najim Dehak

Center for Language and Speech Processing
Department of Electrical and Computer Engineering, The Johns Hopkins University, USA
{rpappag1, jvillal7, ndehak3}@jhu.edu

ABSTRACT

We present an end-to-end multi-scale Convolutional Neural Network (CNN) framework for topic identification (topic ID). In this work, we examined multi-scale CNN for classification using raw text input. Topical word embeddings are learnt at multiple scales using parallel convolutional layers. A technique to integrate verification and identification objectives is examined to improve topic ID performance. With this approach, we achieved significant improvement in identification task. We evaluated our framework on two contrasting datasets: 20 newsgroups and Fisher. We obtained 92.93% accuracy on Fisher and 86.12% on 20 newsgroups, which to our knowledge are the best published results on these datasets at the moment.

Index Terms— BOW, raw text, CNN, verification, identification, topic id, end-to-end

1. INTRODUCTION

Managing the ever increasing information on the Internet is critical for efficient access and utilization. Grouping documents based on topics is one style of top level organization. Each text document can be associated with a broad topic based on its contents. Monitoring calls, document retrieval, and spam detection are some of the applications. In this paper, we consider classifying text documents into a set of predefined classes. We consider two kinds of text namely written text and spoken text. We considered *20 newsgroups* dataset for written text and, the Fisher conversational corpus for spoken text.

A simple document representation can be obtained by computing counts of each word in the document, which is called Bag-Of-Words (BOW) in the literature. Naive Bayes and support vector machines (SVM) [1, 2] were commonly used for classification on this representation. Also, generative modes such as latent Dirichlet Allocation (LDA) [3] and its variations, latent semantic analysis (LSA) [4] and probabilistic LSA [5] were proposed using BOW to model document as a distribution of topics. Authors in [6, 7, 8, 9] applied replicated softmax model, restricted Boltzmann machines (RBM), autoencoders and subspace multinomial Models (SMM) to obtain low dimensional representation of documents from BOW representation. All these models learn to project data into a low dimensional latent space. In this generative model setting, simple linear classifier is used to classify documents on the latent space.

Exploiting semantics is one important aspect of representing text documents which is ignored in BOW representation [10]. Many research groups [11, 12, 13] in the text community apply skip-gram models to compute general word embeddings and use them for any text related task. The drawback of skip-gram models is that they use a small context window to learn semantics, which can exploit only local information. Topic ID task demands for global information of the document along with local context information. This could be one reason why BOW representation works well for topic ID as it provides overall document level statistics of each word in the document.

Ideally, for better topic ID performance, algorithm should consider local context, global statistics of words in that document and possibly topic labels to obtain a good document embedding. Word2vec and doc2vec [12, 13, 14] are word and document representation models which are popularly used to obtain embeddings. In [12], word embeddings are learnt taking small context window and topic into consideration. In [13], context sensitive word embeddings are learnt using skip-gram model i.e, same word will have different representation depending on the context. Authors in [11] use skip-gram model followed by GMM to get word embeddings. In these papers, document embeddings are obtained by weighted averaging of embeddings of words in that document where we may loose some information useful to topic ID. The weights are usually term-frequency inverse-document-frequency (TFIDF) scores of words. We can use CNN and long short-term memory (LSTM) to do better averaging.

In [15, 16], CNN is shown to work well in sentence classification by harnessing the sequence information in the text. In [17], multi-channel CNN is proposed where each channel processes one different word embedding obtained from skip-gram model. Our network is different from this in that each

parallel convolution layer learns topic based word embedding by itself at multiple scales to optimize the objective function.

In this paper, we show that jointly optimizing verification and identification tasks together provides huge gains in the classification performance. We also hypothesize that classification can be improved by exploiting local spatial structure of documents. Minimizing the distance between two documents of same class and maximizing the distance between documents of different class is the main idea of verification task. Optimizing verification and categorical loss objectives improves within class and between class variance. Similar idea was explored in face verification [18] task to improve verification performance which is different to our idea of improving classification performance.

The main contributions of this paper are:

- We present an end-to-end framework for topic ID using multi-scale CNNs.
- An algorithmic approach for integrating verification and identification tasks which improve identification task.
- Our approach works on raw text and do not consider any kind of ad-hoc techniques such as feature weighting, or vocabulary selection for better topic ID performance

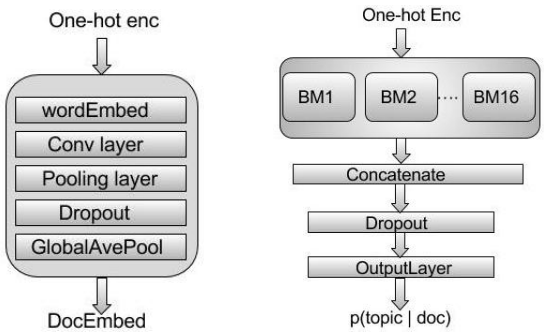
Rest of the paper is organized as follows. In Section 2, the details of our approach are presented. In Section 3, we describe our experimental setup and the details of architecture. Results are discussed in Section 4 followed by conclusion.

2. END-TO-END TOPIC IDENTIFICATION

Identification is the task of classifying a given test document into one of a predefined set of classes. Verification is the task of deciding whether two given documents are from the same topic. These two tasks are related but have different objectives. We integrated both of these tasks objective functions to improve identification performance. In the document classification process, we obtain several levels of representations: context-dependent word embeddings, learnt using different context windows size; and document embeddings which encapsulate the information from different context scales.

2.1. Architecture

Figure 1 shows the architecture of our network. It has parallel convolutional layers followed by dropout, pooling and fully connected layer with softmax activation. Each parallel convolutional block, shown in Figure 1a and denoted as Base Module (BM), is expected to learn semantics from raw text at multiple scales, for example, word level and sentence level. We used 16 such BMs parallelly. Convolutional layer in each BM



(a) Base Module(BM) (b) Full architecture

Fig. 1: Multi-scale CNN architecture

operates on the word embedding space learned for that specific scale. Here, a transformation matrix is learned to transform one-hot encoding to dense vector which is same as word embeddings. Then, we do local temporal average pooling– half of BMs have pool-size=2 and the other half pool-size=7– , dropout and global temporal average pooling within each BM to summarize the information learnt with that scale. Low dimensional document embedding is obtained by concatenating all the parallel BMs. Note that concatenating the BM embeddings, instead of averaging, allows to propagate more information forward and improve classification. To classify the document, we feed this embedding to dropout layer followed by fully connected layer with softmax activation function. Next, we describe the objective function used to optimize this network.

2.2. Objective function

Objective function in a traditional classification task is categorical cross-entropy

$$H_{\mathrm{A}} = -\sum_{k=1}^{N} y_k \log(p_k) \tag{1}$$

where p_k is the softmax activation at k^{th} output neuron and y_k is 1 if document A belongs to k^{th} class, and it is 0 otherwise; the number of classes is denoted by N.

Generally, in a verification task, using a siamese network [19], contrastive loss or binary cross-entropy is used as objective function. Through experiments we chose binary cross entropy,

$$V(\mathrm{A},\mathrm{B}) = -t_{\mathrm{A-B}} \log(p_{\mathrm{A-B}}) - (1 - t_{\mathrm{A-B}}) \log(1 - p_{\mathrm{A-B}}) \tag{2}$$

where $t_{\mathrm{A-B}}$ is 1 if both documents A and B are from same class and 0 otherwise. $p_{\mathrm{A-B}}$ is same-topic posterior computed as the sigmoid of cosine similarity between document embeddings of A and B denoted by $d(A)$ and $d(B)$.

$$p_{\mathrm{A-B}} = \frac{1}{1 + e^{-\cos(d(A),d(B))}} \ . \tag{3}$$

By optimizing this verification loss on the document embedding, we force the documents from the same class to be more similar and documents from different classes to be more dissimilar.

Finally, we optimize the objective

$$C = \sum_{\mathrm{A}} \sum_{\mathrm{B} \neq \mathrm{A}} H_{\mathrm{A}} + H_{\mathrm{B}} + \lambda V(\mathrm{A}, \mathrm{B}) \tag{4}$$

where λ is a scale factor to balance the weight of classification and verification objectives; and the sum is calculated over all possible document pairs.

To compute verification loss we need to form document pairs, which has significant effect on the system. We obtain each batch of pairs as follows: sample a subset of topics; then sample two document per topic from that subset; finally, form every possible pair between those documents. We describe the details of training in the next section.

3. EXPERIMENTAL SETUP

We used two contrastive topic ID data sets: 20 newsgroups, containing written text, and Fisher Phase 1 corpus, containing conversational spoken text.

3.1. 20 newsgroups

20 newsgroups[1] data set is commonly used in the text processing community. This data set contains approximately 20,000 documents from 20 topics. We used 10,174 documents for training; 1,140 documents for validation; and 7,532 for evaluation. To compare with other papers, we used the 53160 words vocabulary set provided in the dataset web site.

3.2. Fisher

Fisher Phase 1 corpus is commonly used in speech community. It contains 10-minute long telephone conversations between two people discussing a given topic. We used the same training and test splits as [20] in which 1374 and 1372 documents are used for training and testing respectively. We split the 1374 training documents into training and validation sets with 90% and 10% proportions respectively. The number of topics in this data set is 40. We used manual transcriptions for each document.

3.3. Network Details

One-hot encoding was used to represent the input words. Each document was represented as sequence of one-hot vectors and fed to the network. We did not apply any preprocessing like feature weighting or vocabulary selection based on word relevance to topic.

We first tuned the hyperparameters of the network which are dimension of word embeddings, number of filter maps, filter widths, number of parallel convolution layers. Eventually, 300 dimensions for word embeddings was chosen in all parallel layers. 16 parallel convolutional layers each with 3 filter maps were selected. Having more filter maps did not improve the accuracy. We divided these 16 parallel layers into 2 sets. Each set had filter widths linearly spaced between 1 to 22 (1,4,7,10,13,16,19,22). First set of layers were followed by temporal pooling layer with pool size of 2 and second set of layers were followed by pool size of 7. Dropout was used after pooling to avoid over-fitting. After dropout, global temporal pooling layer was used to obtain an embedding vector from each scale. By concatenation of all convolution layers, we got a 48 dimensional dense vector which was fed to another dropout layer followed by softmax output layer.

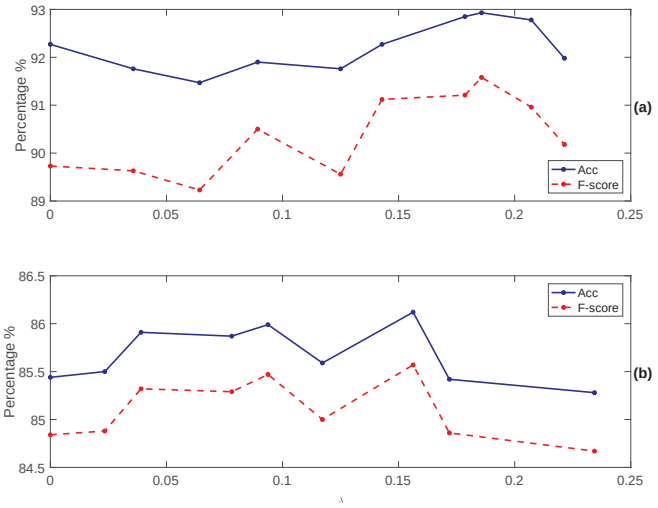


Fig. 2: Accuracy and F-score plots on (a) Fisher and (b) 20 newsgroups for different values of λ

For training, each minibach consisted of a set of documents. To create the minibatches, we used the method in [21]. For 20 newsgroups data set, for each minibatch we selected 8 classes and two documents were randomly sampled from each class resulting a set of 16 samples. Afterwards, we pair each sample with every other sample in the set to have 8 positive pairs (same topic) and 112 negative pairs (different topic). Since we have more negative pairs than positive pairs, we balanced the weight of both pair types in the verification loss function. We did it by multiplying the loss of the negative pairs by the ratio between the number of positive and negative pairs. For Fisher data set, we reduced the minibatch size by sampling from only 4 classes. We did so to avoid GPU memory overflow given the longer size of Fisher documents. Keras with Tensorflow backend was used for our experiments [22]. Accuracy and F-score were used as evaluation metrics. F-score is defined as harmonic average of precision and recall. A good system should have high accuracy and F-score. F-score is more appropriate metric compared to accuracy for Fisher data as the data is unbalanced.

[1] http://qwone.com/ jason/20Newsgroups/

Table 1: Accuracy and F-score on 20newsgroups and Fisher

dataset	Model	Accuracy	F-measure
20 newsgroups	NTSG-1 [13]	82.6	81.2
	SCDV [11]	84.6	84.6
	CNN $\lambda = 0$	85.44	84.84
	CNN $\lambda = 0.16$	**86.12**	**85.57**
Fisher	SVM MCE [20]	91.9	-
	CNN $\lambda = 0$	92.27	89.73
	CNN $\lambda = 0.18$	**92.93**	**91.21**

Table 2: MSCS analysis on document embeddings

	λ	MSCS_W	MSCS_B
20 newsgroups	0	1.00	0.69
	0.16	**1.00**	**0.58**
Fisher	0	0.95	0.63
	0.18	**0.96**	**0.55**

4. RESULTS AND DISCUSSION

Figure 2 shows accuracy for the system with identification+verification objective for different values of λ on both data sets. The system with $\lambda = 0$ corresponds to just identification objective and serves as baseline. For many values of λ, accuracy and F-score significantly improved on both data sets.

Table 1 compares different systems in terms of accuracy and F-score. On 20 newsgroups, our baseline system itself performs better than previously published results with 85.44% accuracy. Adding the verification loss, we improve by 0.68% absolute with $\lambda = 0.16$. Our baseline system is 0.84% better than SCDV [11], and overall our system is 1.52% better.

On Fisher, our system accuracies are 92.27% and 92.93% when λ is 0 and 0.18 which are significantly better than the accuracy 91.9% reported in [20]. Comparison with [20] may not be appropriate here as authors used text produced by Automatic Speech Recognition (ASR) instead of manual transcriptions.

Optimizing verification loss reduces the distances between positive pairs and increases distance between negative pairs. To quantify this operation, we calculated mean squared cosine similarity (MSCS) [8] within topics and between topics. It is calculated as

$$\text{MSCS} = \sqrt{\frac{2}{M(M-1)} \sum_{\text{A,B}} \left(\frac{1 + \cos(\text{A}, \text{B})}{2} \right)^2} \quad (5)$$

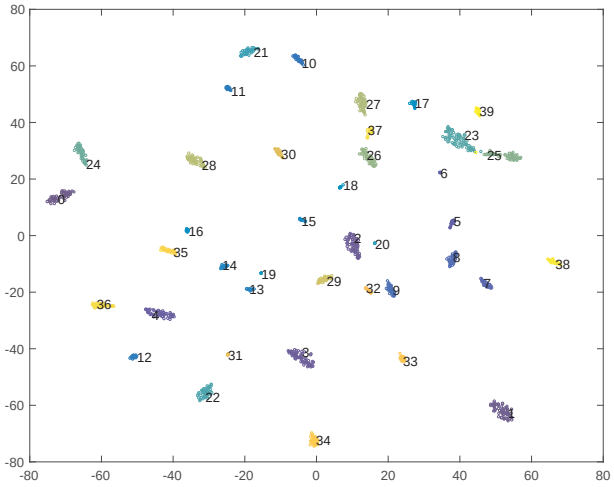


Fig. 3: Visualization of document embeddings on Fisher dataset using t-SNE

where cosine distance is between document embeddings (output of global pooling). Smaller values of MSCS indicate more orthogonality and higher values indicate more similarity between document embeddings. We calculate within and between topic similarity, denoted by MSCS_W and MSCS_B respectively. Table 2 shows that MSCS_B value is smaller when $\lambda > 0$ compared to $\lambda = 0$. It indicates that embeddings are more orthogonal between topics when $\lambda > 0$.

Figure 3 shows t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of document embeddings on fisher dataset. Classes are indicated by different colors and class index. t-SNE is applied after reducing dimensionality of document embeddings to 30 using PCA. It can be observed that classes are well separated into exactly 40 clusters which indicates good classification accuracy.

5. CONCLUSION

In this paper, we presented an end-to-end framework for the topic ID task. Multi-scale CNNs followed by temporal pooling was proposed to compute document embeddings. These embeddings were used for topic identification by applying a fully connected layer with softmax activation. We proposed to combine identification and verification objective functions to train this network. Experiments on two contrasting data sets –20 newsgroups and Fisher– showed that adding the verification objective significantly improved accuracy and F-score w.r.t. just using classification objective. To our knowledge, the results obtained on the 20 newsgroups and Fisher datasets outperformed over the best published results at the moment. As future work, we further explore incorporating sequence dynamics modeling with LSTM into this framework.

6. REFERENCES

[1] István Pilászy, "Text categorization and support vector machines," in *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, 2005.

[2] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] Thomas K Landauer, *Latent semantic analysis*, Wiley Online Library, 2006.

[5] Thomas Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[6] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in neural information processing systems*, 2009, pp. 1607–1614.

[7] Hugo Larochelle and Yoshua Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 536–543.

[8] Yu Chen and Mohammed J Zaki, "Kate: K-competitive autoencoder for text," *arXiv preprint arXiv:1705.02033*, 2017.

[9] Santosh Kesiraju, Lukás Burget, Igor Szöke, and Jan Cernockỳ, "Learning document representations using subspace multinomial model.," in *INTERSPEECH*, 2016, pp. 700–704.

[10] Rie Johnson and Tong Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.

[11] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick, "Scdv: Sparse composite document vectors using soft clustering over distributional representations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[12] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun, "Topical word embeddings.," in *AAAI*, 2015, pp. 2418–2424.

[13] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "Learning context-sensitive word embeddings with neural tensor skip-gram model.," in *IJCAI*, 2015, pp. 1284–1290.

[14] Jey Han Lau and Timothy Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[15] Ye Zhang and Byron Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[16] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[17] Haotian Xu, Ming Dong, Dongxiao Zhu, Alexander Kotov, April Idalski Carcone, and Sylvie Naar-King, "Text classification with topic-based word embedding and convolutional neural networks.," in *BCB*, 2016, pp. 88–97.

[18] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[19] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 539–546.

[20] Timothy J Hazen, "Mce training techniques for topic identification of spoken audio documents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2451–2460, 2011.

[21] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[22] François Chollet et al., "Keras," 2015.