Local Matching Networks for Confidential Documents Classification

Ying Li *, Haoqi Zhu
NetEase, Inc

1 Background

Assign security level or access control to documents is prevalent in organizations. However, following detailed rules lacks flexibility, and the manual designation is arbitrary and security risks could be introduced. Referring to historical document that already have a security level is an excellent way to go, considering that an organization's information has a certain continuity within a limited period. With this, we pro-pose the Local Matching Networks in this paper that consist of two parts. First, similar existing documents with labels were found for the query example. And then, a matching function parametrized by the neural network renders the label of security level based on the metric learned from a tailored form of datasets

2 Introduction

The non-parametric model can offer a simple solution to refer to historical data, e.g., the Nearest Neighbor model. This method is quickly adapted to new data since it needs no training, but it often suffers from computation inefficient and lack of flexibility in generalization.

the Neural network is another choice, the problem is that the model has to be retrained for the new data. This will cost enormous resources as the number of documents grows. And as data grows, the limitation of model capacity may be reached.

Because the more similar the documents in the historical data are to the current one, the more guidance information they can provide. We propose to treat this problem as finding differences in similarities rather than a simple classification problem.

3 Methods

There are two main parts of our approach.

First, for every query example, we use a text similarity algorithm to find similar documents in the history data and then combine them to form our model's training set.

Second, we build a model similar to the Matching Networks which we call the Local Matching Networks (LMN). This model takes in the new form dataset and decides the label of the query sample based on its similarity, which is measured by the metrics that it learned, to the examples in the support set

4 Results

	Bert	LMN
random split dataset	91.3%	88.7%
chronologically split dataset	87.4%	88.1%

The Bert baseline trained on the random split dataset performs the best. Since the baseline Bert is very powerful because it has a complete view of the dataset. Our model is better at the chronological split dataset. We assume the primary reason is that similar data collected from historical data provide helpful information.

	1	2	3	4	5	6	7	8
Bert	85.3%	89.7%	89.2%	80.2%	**87.6%**	88.5%	87.9%	**82.8%**
LMN	70.7%	83.7%	75.5%	71.7%	**89.3%**	71.9%	80.2%	**83.6%**

Although the Bert baseline is mighty in terms of the single indicator (i.e., accuracy), our model can occasionally shine when it comes to specific similar document clusters. Here we randomly pick 8 similar document clusters, each with the same number of samples, and let the model predict the examples in these clusters, and the results are shown in

	Bert	LMN
Train on just D_{test1}	86.8%	88.5%
Train on $\{D_{train}, D_{test1}\}$	87.9%	88.7%

As we can see, when retaining the model on the expanded training data and testing on the smaller test set, the performance increases with no surprise. However, when trained only on the new data, the accuracy of the baseline model drops. We assume this is because the fewer new training data may cause the model a bit of overfitting, and the information that can support the test set contained in the old training set is affected by the newly emerging data. Nevertheless, the performance of our model in this situation increased. Such a behavior could be expected since instead of learning the distribution character that the baseline model does, our model is learning a proper way of representing similar documents. This meta-learning-like property, combined with the fact that when trained on the new data, the similar data found are from the older training set, could alleviate the forgetting problem.

5 Conclusion

We propose the Local Matching Network in this paper, which combines the advantage of the quick adaption from the non-parametric method and the better generalization and running efficiency from the parametric neural network while avoiding their disadvantages. It is suitable for determining the security level of confidential documents and considering both efficiency and effectiveness. A drawback of our model is that the number of similar documents in each class is fixed, which lacks some flexibility since some found documents may not be relative enough. We consider this as an opportunity for improvement in future work.

* References

1. Vinyals, Oriol, et al. "Matching networks for one shot learning." Advances in neural information processing systems 29 (2016): 3630-3638.Insert your text here. Change the font size to fit your text in the space.
2. Pérez-Iglesias, Joaquín, et al. "Integrating the probabilistic models BM25/BM25F into Lucene." arXiv preprint arXiv:0911.5046 (2009).
3. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).