Better Late than Never: Late Fusion Techniques for Document Classification

Yawen Zhao
yawen.zhao@student.kuleuven.be

Jordy Van Landeghem
KU Leuven - Contract.fit
firstname@contract.fit

Abstract. In this paper, we propose a novel late fusion technique for document classification by combining fusion mechanisms with feature-to-string encoding methods. After evaluating proposed models against the BERT baseline and several models using traditional encoding methods, we conclude that models using Concatenation Fusion or Attention Fusion combined with *feature-to-string* encoding method have the best performance.

Keywords: Late Fusion, Feature-to-string, Document Classification

1 Introduction

Document classification is one of the most crucial steps in automating email routing. Since BERT is famous for its excellent performance in many NLP tasks, it is the model of choice for practical NLP applications [4]. BERT takes only the text as the input with a limit on the maximum sentence length. However, there is plenty of information in other modalities which can potentially contribute to the performance of models. Because of the limitations mentioned above, valuable features are not always used as input. This paper applies late fusion and feature engineering techniques to merge information from additional modalities with the initial text modality to solve this problem.

2 Methods

Based on fusion mechanisms proposed by previous research [10, 11, 9, 1, 5, 8], we focused on Concatenation Fusion and Attention Fusion mechanisms that we implemented from scratch in our study. Apart from the traditional feature encoding methods of label encoding and one-hot encoding, we also combined fusion mechanisms with feature-to-string encoding [7], which converts all features into a string and exploits the BERT model as a (string-)feature encoder.

The data we use for this study is from an insurance company, a client of Contract.fit. The data set is made up of email documents, including subjects, bodies of email, attachments, and additional features. The attachments are the

OCR text outputs of the images attached to the email. The subjects, bodies of email, and attachments are merged together as inputs to BERT. The sizes of the train, validation, and test data set are 35221, 3914, and 1079 with 60 classes. Due to time and resource constraints, only nine features are selected for experiments. Seven categorical features are the number of tokens and numbers in the merged text, the branch labels, the email receiver, the forward department, and the first, and second digits of policy and claim numbers. Two numerical features are the number of tokens and numerical patterns.

Example 2.1 *the branch is A. the email receiver is abc@defgh.be. the department is B. the first policy digit is 7. the second policy digit is C. the first claim digit is D. the second claim digit is 1. number of tokens is 169, number of numbers is 3.*

3 Results and Discussion

All models are evaluated for document classification task with Accuracy, Brier Loss [3], Negative Log-Likelihood (NLL) [2], and Expected Calibration Error (ECE) [6]. Selected results are given below in Table 3.

Model(encoding)	Accuracy(%)	Brier Loss	NLL	ECE
BERT	85.17	0.238	0.627	0.051
Concatenation(label)	86.01	0.228	0.620	**0.036**
MLP_Attention(label)	85.63	0.222	**0.584**	0.037
noMLP_Attention(label)	**86.65**	**0.220**	0.634	0.049
Concatenation(feature-to-string)	87.12	**0.198**	**0.541**	**0.028**
noMLP_Attention(feature-to-string)	**87.21**	0.205	0.600	0.047

Table 1: Evaluation results of models using label and feature-to-string encoding

The Attention Fusion model using feature-to-string encoding achieved the highest accuracy. The Concatenation Fusion model using feature-to-string encoding has the lowest Brier Loss, NLL, and ECE.

The results indicate that, in terms of feature encoding methods, the feature-to-string encoding outperforms the rest with several benefits. It automatically encodes richer information with the tokenizer of BERT, handles the problem of missing data and allows for automatic feature scaling. Besides, the BERT encoder forms part of the model and is also fine-tuned during training.

To determine the model's performance in production, the automation evaluation is applied. It jointly compares the automation rate and the error rate of each model given a certain confidence threshold. The two models using feature-to-string encoding have the best performance.

Based on these experiments, the feature-to-string encoding method combined with the Concatenation Fusion or the Attention Fusion mechanism is recommended for industrial applications. Future studies could focus mainly on two aspects: testing more advanced fusion mechanisms and improving the quality and quantity of features.

References

[1] Antonios Anastasopoulos, Shankar Kumar, and Hank Liao. Neural language modeling with visual features. *arXiv preprint arXiv:1903.02930*, 2019.

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.maiworkshop-1.10. URL https://aclanthology.org/2021.maiworkshop-1.10.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[7] Chris McCormick. Combining categorical and numerical features with text in bert. https://mccormickml.com/2021/06/29/combining-categorical-numerical-features-with-bert/31-all-features-to-text, 2021.

[8] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

[9] Valentin Vielzeuf, Alexis Lechervy, Stephane Pateux, and Frederic Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[10] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[11] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.