CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval

Hou Wei Chou[*]
CS Department, Purdue University
West Lafayette, IN, USA

Younghun Lee[†]
CS Department, Purdue University
West Lafayette, IN, USA

Lei Chen[‡]
Rakuten Institute of Technology
Boston, MA, USA
lei.a.chen@rakuten.com

Yandi Xia
Rakuten Institute of Technology
Boston, MA, USA

Wei-Te Chen
Rakuten Institute of Technology
Boston, MA, USA

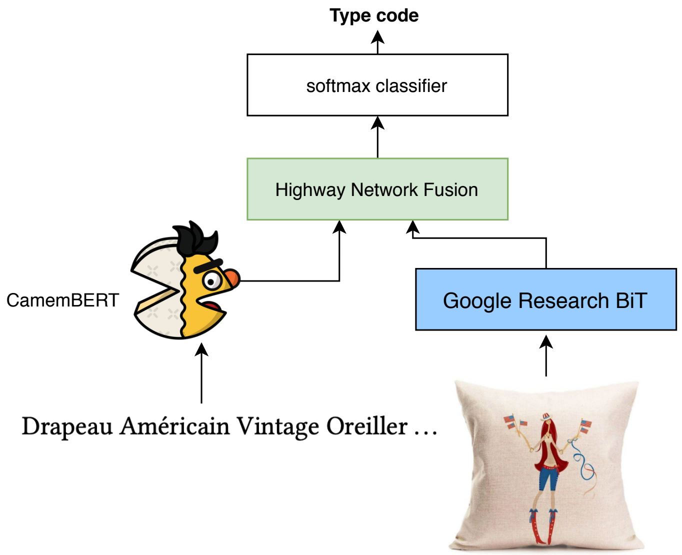

Figure 1: RIT Boston team's CBB-FE multimodal product classification system

ABSTRACT

For SIGIReCom'20 multimodal product classification and retrieval challenge, we proposed a solution that consisting of (a) fine-tuning feature extractors from text and image modalities respectively, (b) fusing multimodal features to train a classifier for the first task on product classification, and (c) using a similarity search method to retrieve product images from their text titles. Utilizing latest technology development on pre-trained models, we used CamemBERT and BiT models for their superior performance in uni-model setups. Then, several fusion mechanism have been compared and the Highway network based fusion was found to be optimal. Similarly, leveraging on the high performance of multimodal features, our cross-modal retrieval system can be built up on top of any similarity-search tool. Our solution has advantages on its modular design and can use pre-trained models' representation learning abilities. On final score board, our multimodal product classification system achieves a macro-F1 90.22%, which is very competitive (lagging the winning solution only 1.3%). For the cross-modal retrieval task, our system ranked 3rd with a recall @1 of 23.30%.

ACM Reference Format:
Hou Wei Chou, Younghun Lee, Lei Chen, Yandi Xia, and Wei-Te Chen. 2020. CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval. In *SIGIReCom'20: SIGIR eCommerce Workshop'20 July 30, 2020, Xi'an, China.* ACM, New York, NY, USA, 5 pages.

[*] While Hou Wei was a summer intern at RIT Boston
[†] While Younghun was a summer intern at RIT Boston
[‡] Correspence author

SIGIReCom'20, July 30, 2020, Xi'an, China

1 INTRODUCTION

Product item categorization (IC), is a core technology in the modern e-Commerce area. Since a very large number of products (millions or even more) are listed in an eCom market, it's important to build an effective and convenient taxonomy so that buyers could quickly locate the products of their interest or need. Consequently, a hierarchy structure with more than $10K$ leaf nodes is utilized to categorize a massive number of products into different nodes.

In spite that IC shares the same setup as a text classification task, it possesses its unique aspects, including (a) a large number of classes, (b) a severe long-tailed distribution of labels, and (c) noisy raw inputs due to the fact that these inputs are generally provided by merchants in a heterogeneous way. These unique aspects make IC be a challenging task in practice.

Products have their titles/descriptions in text modality and images in vision modality. Most of IC studies have been focusing on using text-based cues, but, apparently, images of products provide useful clues for a more accurate IC. For example, in some sub-areas like fashion, information conveyed through photos is richer and more accurate than that through the text channel. [12] proposed a multimodal IC (MIC) system for classifying products on Walmart.com by using titles and photos. A multi-label classification was conducted on about $1.2M$ product. The goal was classifying products to shelves, from 2890 possible choices. Each product is typically assigned to more than one shelf (3 on average). They built a textCNN [4] text model and a VGG-based image model. Then, they designed several policies to fuse two models' decisions. Compared to uni-models' performance, multimodal IC shows promising performance improvements.

In this paper, we described our submissions designed for solving the SIGIReCom'20 Workshop Multimodal Product Classification and Cross-modal Retrieval Challenge tasks. Our report was organized as follows: Section 2 recaps essential information of the challenge; Section 3 depicts our MIC system in details, including feature extraction (FE) and fusion layer for predicting product type code, and the product image retrieval system using product titles; Section 4 illustrates our experiments during the challenge, including model building at the stage-1 (by July 15th) and final testing results at the stage-2; At last, Section 5 discusses our findings based on experiments.

2 TASK

The Rakuten Multimodal Product Classification and Retrieval Challenge contains two tasks. The first task requires to predict a product to one of 27 product type codepre-defined in Rakuten France. For a product, its title (mostly in French), image, and an optional description (appearing in about 65% titles in the training set) , are provided as raw inputs. In the cross-modal retrieval task, presented with the text of the products, the goal is to retrieve the images corresponding to the products.

For this challenge, Rakuten France has released approximately $99K$ products, including a training set ($84,916$) and test set ($13,812$). The test set was released at two stages. In the stage-1 (before July 15th), only about 10% of the test set was available. The entire test set was released after July 15th. For the classification task, the metric used in the challenge is macro-F1 among all 27 labels. For the retrieval task, the metric is recall at 1 ($R@1$) on held out test samples. Figure 2 shows the number of products among all 27 product type codeclasses. A long-tailed distribution pattern is clearly shown. On all training data, we used Google Sheet's language detection function to identify titles' languages. We found that 85% are in French and about 15% in English and also in other languages.

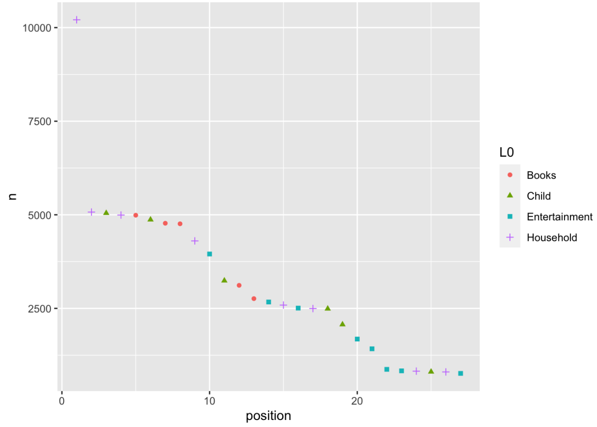

Figure 2: Number of products on 27 type code classes show a long-tailed distribution pattern

3 MODEL

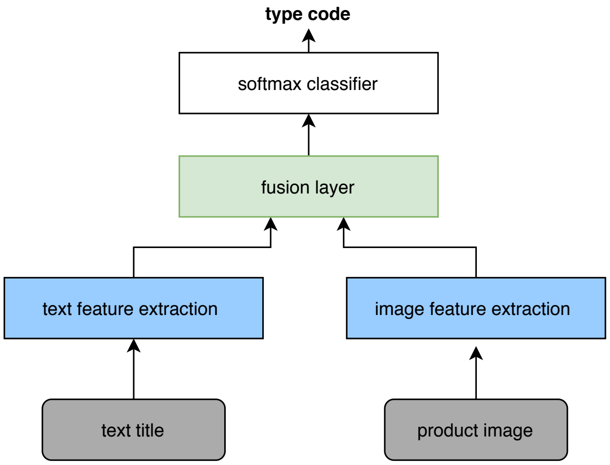

Figure 3: Our proposed MIC system consists of feature extractors on texts and images respectively, a fusion layer to fuse multimodal features, and its final softmax classifier to predict product type code

As shown in Figure 3, our MIC model consists of feature extraction (FE) parts from uni-modal channels (i.e., text titles and images), a fusion part to obtain multimodal features, and a softmax classifier to make final predictions.

3.1 Text FE methods

Regarding the text FE part, we tried two methods, i.e., standard text CNN model [4] and a more recent transformer-based BERT model [1], CamemBERT [9], which is a special version RoBERTa [7] style BERT trained on French data

Similar to BERT, RoBERTa uses masked language modeling objective to pre-train, wherein the model learns to predict masked sections of text from large-scaled raw text samples. It showed several improvements in comparison to BERT, including (a) removing BERT's next-sentence pre-training (NSP) objective, (b) training with much larger mini-batches and learning rates. These improvements allow RoBERTa to enhance the masked language modeling objective compared with BERT and lead to a better downstream task performance.CamemBERT was trained on $136GB$ French data [11].

3.2 Image FE methods

Regarding the image FE part, we fine-tuned the pre-trained image classification models on product images in the challenge dataset. Two pre-trained models were used in this study. The first model is the ResNet with 152 layers [2], denoted as ResNet-152 model. The ResNet-152 model was pre-trained on the ILSVRC-2012 dataset that contains $1.28M$ images with 1000 classes.

The second model, Google's Big Transfer (BiT) model [6], is a new one released in late May 2020. It reflects recent years' technology developments on pre-training of supervised dataset. The BiT model is based on a standard ResNet model and is trained with the following improvements, i.e., (a) increasing depth and width, replacing BatchNorm (BN) with Group Normalization and Weight Standardization (GNWS), and pre-training on a very large and generic dataset for many more iterations. We used BiT-M, a ResNet-152x4 model, which was pre-trained on the ImageNet $21K$ dataset that contains $14M$ images with $21k$ classes. Compared to the ResNet-152, the BiT-M was trained with more than 10 times more images of about 20 times more classes.

3.3 Multimodal fusion

Regarding combining features from text (f_t) and image domains (f_v) to form a multimodal feature for predicting `product type code`, we considered two options.

The first option uses a highway network [10] to add the concatenated features through its two formats, i.e., one from a linear transformation and the other one from a non-linear transformation. This method has been employed in previous NLP research when integrating different word embedding vectors [5]. The second option uses a tensor fusion to model interactions between two modalities. Simply speaking, the tensor fusion method computed outer products between f_t and f_v and then learned an affine transformation to map onto multimodal feature vector f_m. We applied a special tensor fusion approach, low rank multimodal fusion (LMF) proposed in [8] in this challenge.

3.4 Cross-modal Retrieval Model

In order to retrieve the most relevant image from product title data, we use a simple vector similarity approach.

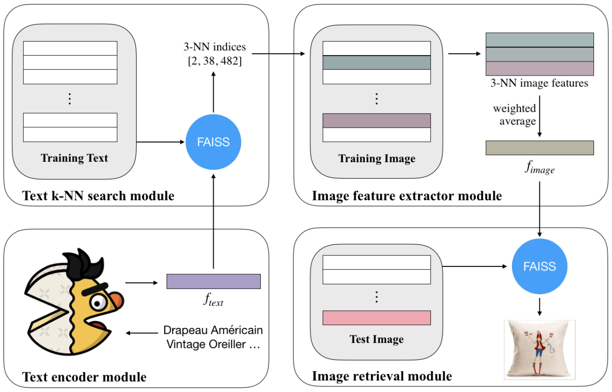

Figure 4: Model structure for cross-modal retrieval task

Using the fine-tuned CamemBERT model, we encode text data from test set and generate feature representations, f_{text}. We then calculate the L2 Euclidian distance between f_{text} and the encoded product title data from the training set. We made use of FAISS [3] to efficiently calculate the L2 distance and the index of k nearest neighbors. After finding the indices of k nearest neighbors to the text input, we get the feature representations of k images from the training set. We then calculate weighted average to get f_{image}, which can be seen as the feature representation of the image that is most relevant to the input text. Weight values are calculated from the inverse of the L2 distances. As a final step, we run another FAISS module to compute L2 distance between f_{image} and the encoded product image data from the test set, and get the closest image representation from the test set.

4 EXPERIMENT

4.1 Setup

The training set was split into the three parts: **FE-train**, **fusion-train**, and **fusion-dev** sets with 80%, 16% and 4% percentages. The FE-train set was used to train or fine-tune uni-modal FE networks. Then, the learned FE networks were applied on both fusion-train and fusion-dev sets to extract multimodal features. The fusion layer and softmax classification layer were trained on the fusion-train set.

Codes in our experiments were developed on PyTorch. For image feature extraction, we used ResNet-152x4 from BiT[1]. For text feature extraction, we used Camembert/Camembert-large from hugging face[2]. The hyper-parameters when running experiments were as follows: CamemBERT uses a learing rate $3e^{-5}$, Adam optimizer, and 20 epochs. BiT (ResNet-152x4) used 10000 steps with default setting. The fusion methods used a learning rate 0.01, Adagrad optimizer, and 20 epochs.

4.2 Result

4.2.1 Stage 1. Table 1 reports on the evaluation results (macro-F1 among 27 labels) on uni-models (on stage-1 test set). On texts, CamemBERT model has a large macro-F1 improvement than the

[1] https://github.com/google-research/big_transfer
[2] https://huggingface.co/camembert-base

textCNN model (6.42% F-1 increase). On product images, new BiT model has a quite large macro-F1 improvement than the ResNet-152 model (13.32% F-1 increase). This pattern suggests that pre-training on large-scaled dataset improves models' performance on downstream tasks.

Modality	Model	macro-F1 (%)
Text	textCNN	80.58
Text	CamemBERT	87.2
Image	ResNet	60.65
Image	BiT	73.97

Table 1: IC using uni-modal features; This training helps to fine-tune pre-trained models to better fit to the challenge data set.

Table 2 reports on the evaluation results on fusion models jointly using text and image features. If using the highway network fusion layer, the bi-model's performance is 90.51% and suppresses any uni-model result (CamemBERT 87.2% and BiT 73.97%). This showed that using multimodal cues plays an role for improving IC performance. When using low-rank multimodal fusion network to be the fusion layers, the performance was 89.5%, slightly worse than the highway network method.

Ensemble models generally can improve performance and this strategy has been widely used in previous Kaggle data challenges. We ensembled 4 fusion models that were trained on different random seeds. Compared to single model's performance, ensemble models showed further improvements. Highway network fusion can be improved to 90.82% and LMF fusion can be improved to 90.75%.

Models	Fusion method	macro-F1 (%)
CamemBERT + BiT	Highway network	90.51
CamemBERT + BiT	LMF	89.5
Ensemble C+B (n=4)	Highway network	90.82
Ensemble C+B (n=4)	LMF	90.75

Table 2: On top of text feature extractor based on CamemBERT and image feature extractor based on BiT, two fusion methods' performance when using single model or an ensemble of 4

Table 3 reports on the task-2 evaluation result using our simple FAISS [3] solution. When using nearest 3 title's corresponding images to compute an searching vector, we could retrieve 42.34% images from the stage-1 test set.

K	Method	Metric (R@1)
10	ave.	20.95
10	weighted ave.	36.50
3	ave.	24.5
3	weighted ave.	**42.34**

Table 3: Using FAISS [3] to retrieve images from titles

4.2.2 *Stage 2.*

In the stage 2, 10 times larger test set was provided. We selected both classification and retrieval systems with the highest performance in the stage-1 experiments and made our final submissions. Table 4 reports on our final performance metrics, corresponding ranks, and relative loss to top teams. We can find that our multimodal classification method is competitive and its performance is close to the top system. On the retrieval task, our rank is within top 3. However, due to the fact that our retrieval system uses a design principle of fully using vector similarity computation, it still has a large room to improve.

Task	Metric	Rank	loss to top (%)
Classification	90.22	5th	1.3%
Retrieval	23.30	3rd	32%

Table 4: RIT-Boston team's final performance on both product classification and cross-modal retrieval tasks.

5 DISCUSSION

We obtained the following important findings. First, pre-trained models on both text and image domains showed improvements on down-stream tasks. Simply fine-tuning these models on the challenge's about 80K training instances, we could gain proper feature extraction. For example, compared to the textCNN model, fine-tuning CamemBERT showed a noticeable improvement (6.42% F-1 improvement). Second, recently released BiT model demonstrated its larger improvement than ResNet. When fine-tuning on the challenge data set, a substantial F-1 improvement (13.32%) is demonstrated. Third, products' visual information plays important contributions to the IC task. By jointly using multimodal cues, we observed more improved macro-F1 than simply using text inputs. Fourth, dense vectors computed from nowadays pre-trained text and image models are quite informative. Only utilizing vector similarity computation, e.g., FAISS, a simple retrieval system can be built up easily with moderate performance.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).

[4] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1746–1751.

[5] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In *Thirtieth AAAI Conference on Artificial Intelligence.* https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489

[6] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big Transfer (BiT): General Visual Representation Learning. *arXiv:1912.11370 [cs]* (May 2020). http://arxiv.org/abs/1912.11370 arXiv: 1912.11370.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[8] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank

Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2256.

[9] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. *arXiv:1911.03894 [cs]* (May 2020). http://arxiv.org/abs/1911.03894 arXiv: 1911.03894.

[10] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway Networks. *arXiv:1505.00387 [cs]* (Nov. 2015). http://arxiv.org/abs/1505.00387 arXiv: 1505.00387.

[11] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CC-Net: Extracting High Quality Monolingual Datasets from Web Crawl Data. arXiv:1911.00359 [cs.CL]

[12] Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. 2016. Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce. *arXiv preprint arXiv:1611.09534* (2016).