Deep Multi-level Boosted Fusion Learning Framework for Multi-modal Product Classification

Ekansh Verma
IIT Madras, India
vermaekansh55@gmail.com

Souradip Chakraborty
Indian Statistical Institute, India
souradip24@gmail.com

Vinodh Motupalli
IIT Madras, India
mvk793@gmail.com

ABSTRACT

In this paper, we present our approach for the 'Multimodal Product Classification' task as a part of the 2020 SIGIR Workshop On eCommerce (ECOM20). The specific objective of this task is to build and submit systems that classify previously unseen products into their corresponding product type codes. We propose a deep Multi-Modal Multi-level Boosted Fusion Learning Framework used to categorize large-scale multi-modal (text and image) product data into product type codes. Our proposed final methodology achieved a macro F1-score of 91.94 on the phase 1 test dataset which is the top-scoring submission and third position on the scoreboard for phase 2 test dataset with macro F1-score of 90.53.

KEYWORDS

product classification, multi-modal, gradient boosting, text classification, image classification, eCommerce

ACM Reference Format:
Ekansh Verma, Souradip Chakraborty, and Vinodh Motupalli. 2018. Deep Multi-level Boosted Fusion Learning Framework for Multi-modal Product Classification. In *SIGIR'20 DC:SIGIR eCom 2020 Data Challenge, July 30, 2020, Xi'an, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Multi-modal product classification with visual and textual product descriptions is one of the most fundamental problems for any eCommerce marketplace with applications ranging from personalised search and recommendations to query understanding. Traditionally the product classification problems have been approached by retail and eCcommerce firms using rule-based methods resulting in a lack of scalability and generalisability when products need to be classified into a large number of classes.

Developing a multi-modal deep learning architecture solution for product classification will not only help the firms in efficiently scaling it for multiple classes but also save a huge amount of time and rule based generalisation errors. The primary objective of this research work is to solve a large scale multi- modal (text and image) product data classification into product type codes. In related work, Zahavy et al. [17] demonstrate that the text classification CNN [6] outperforms the VGG network [12] on a real-world large-scale product to shelf classification problem and propose a decision-level fusion policy to improve over both. [5] show that fusion with discretized features outperforms text-only classification, at a fraction of the computational cost of full multi-modal fusion, with the additional benefit of improved interpretability. Xia et. al. [15] proposed a variation of CNNs called Attention CNNs applied on Japanese product titles for classification

In this work, we investigate the domain adaptability of state of the art text and image modality based architectures for the eCommerce product classification task. We observe that the BERT[1]-based text classification architectures effectively capture the semantic information present in the product title and description yielding high accuracy results. From our experiments, we observe that textual classification models outperform the visual features based models for this specific product classification task. However, incorporating both the modalities helps us leverage the complementary information present across the features, thereby enhancing the overall performance of the classification system. Finally, we propose a Deep Multi-Modal Multi-Level Fusion framework [1] which learns the joint representation using both the modalities simultaneously where these representations are combined with the uni-modal baselines in a probability-fusion strategy to boost the product classification system. Our proposed methodology achieved a macro F1-score of 91.94 on the phase 1 test dataset which is the top-scoring submission and third position on the scoreboard for phase 2 test dataset with macro F1-score of 90.53.

2 DATASET

2.1 Overview

Text and Image dataset is provided with 27 different product type code and a total of 84916 unique products. Entire product catalog is categorized in 4 top level categories which are Child, Household, Entertainment and Books. Text data is in french language and has two fields, one is the title and the other one being description. Across the dataset, the title field has a median length of 11 words with a maximum of 56 words, whereas description field has a median length of 35 words with a maximum of 2068 words. All the products have a title field but 29800 has no description present.

The product images are all squares of dimensions 500 x 500 pixels, which can have white or black borders included. All the products have an image associated with it.

2.2 Preprocessing

For text data, we performed the initial experiments with product title and descriptions as separate fields to train distinct models

[1]Code to reproduce the framework can be found in https://github.com/depshad/Deep-Learning-Framework-for-Multi-modal-Product-Classification

SIGIR'20 DC, July 30, 2020, Xi'an, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

for text classification. Later, we combined both into one field by concatenating description and title together for each product which leads to a better performing model. Our further experiments use this concatenated field as the input for all the text based models. We have used the text tokenizers corresponding to the pre-trained model.

Images are resized to a uniform size of 224×224 and normalized per channel. We augment the training data for image classification by randomly rotating, flipping and extracting random crops from the original images.

3 IMAGE NETWORK

He et al. [2] introduced the deep residual network (ResNet). Xie et al. [16] this paper proposed a modularized network architecture for image classification adopting a multi-branch aggregation transform residual network called ResNext. ResNeXt won 2nd place in ILSVRC 2016 classification task and also showed performance improvements in Coco detection and ImageNet-5k set than their ResNet counter part. Squeeze-and-Excitation Networks (SENet) [3] introduces Squeeze-and-Excitation (SE) block that adaptively recalibrates channel-wise feature responses by explicitly modelling inter-dependencies between channels. These blocks can be stacked together to form SENet architectures that generalise extremely effectively across different datasets. Further, the SENet and multi-branch aggregation transformation are integrated to form the SE-ResNeXt.

4 TEXT NETWORK

BERT stands for bidirectional encoder representations from Transformers, is designed to learn deep bidirectional representations by jointly conditioning on both left and right context in all layers. Since the first release of BERT, many non-English BERT based language models have been released. Language Models (LM) provide an efficient way to fine-tune continuous word representations to create competitive models for a wide range of downstream tasks, such as sequence classification, named entity recognition, and question answering. We used two BERT inspired French language models : CamemBERT [10] and FlauBERT [7] as baseline models in our methodology.

4.1 CamemBERT

The first French BERT-like language model was that of CamemBERT. It was trained over OSCAR corpus [13] corpus, a French section of the CommonCrawltext[2] dataset. The architecture is largely inspired by an evolution of BERT called RoBERTa [8]. In the CamemBERT paper, authors evaluate the language model in four different downstream tasks for French: part-of-speech (POS) tagging, dependency parsing, named entity recognition (NER) and natural language inference (NLI); improving the state of the art for most tasks over previous monolingual and multilingual approaches.

4.2 FlauBERT

In the paper[7], authors introduce and share FlauBERT, a model learned on a very large and heterogeneous French corpus. FlauBERT has the same model architecture as BERT, which consists of a multi-layer bidirectional Transformer . Their results show that a French language model improves the results compared to similar BERT (multi-lingual) models as well as to other French-based models. In sequential classification tasks namely, constituency parsing and POS tagging, the combination FlauBERT and CamemBERT gives the best results indicating that the two models are complimentary.

[2]https://commoncrawl.org/about/

5 UNI-MODAL BASELINE METHOD

5.1 SE-ResNeXt

We fine-tuned SE-ResNeXt with 50 layers pre-trained on the ImageNet weights using the product images. Visual features from SE-ResNeXt model, $Z_{im} \in R^{p \times p \times d_1}$, where $p \times p$ denotes the size of the output feature maps and d_1 represents the dimension of feature channel, are passed to a adaptive average pooling layer to output, X_{im}, where $X_{im} \in R^{d_1}$. A linear layer is added on top with output dimension equal to the number of classes using a linear layer.

5.2 FlauBERT

We extracted the first token of the FlauBERT's hidden-states output sequence from the last layer, $Z_{t_1} \in R^{m \times d_2}$ where m denotes the sequence length of the text inputs and d_2 represents the model's hidden size. Extracted token embedding, represented by $X_{t_1} \in R^{d_2}$ is projected to the output dimension equal to the number of classes using a linear layer. Flaubert model along with the classification head is fine-tuned for the task of product classification

5.3 CamemBERT

We fine-tuned the CamemBERT for product classification by adding classification head on top of the hidden-states output of the last layer of the model, $Z_{t_2} \in R^{m \times d_2}$. Output Z_{t_2} is flattened and passed through a linear layer which outputs $X_{t_2} \in R^{d_2}$ followed by tanh activation and a linear layer of output dimension equal to the number of classes.

Intermediate vector and tensor representations introduced above are utilised as inputs to simultaneously learn from image and textual features for the task of product classification. Our end-to-end multi-modal methodology is described in the section below.

6 MULTI-MODAL JOINT REPRESENTATION LEARNING

In this section, we emphasize on the multi-modal joint learning of the visual and textual features extracted from the data for the product classification task.

The visual feature extracted by passing the input images through the SE-ResNeXt-50 model is represented by X_{im}, where $X_{im} \in R^{d_1}$. Similarly, the textual features extracted by passing the available textual titles and/or descriptions (whenever available) through the CamemBERT & FlauBERT models are represented by X_{t_1} & X_{t_2} respectively, where both X_{t_1} & $X_{t_2} \in R^{d_2}$. We use a Convolution 1D layer to project the image embeddings from $X_{im} \in R^{d_1}$ to $X'_{im} \in R^{d_2}$ i.e in the same dimensional space as text embeddings R^{d_2}. We have tried multi-modal fusion using two primary ways : *Addition Ensemble* and *Concatenation Ensemble*.

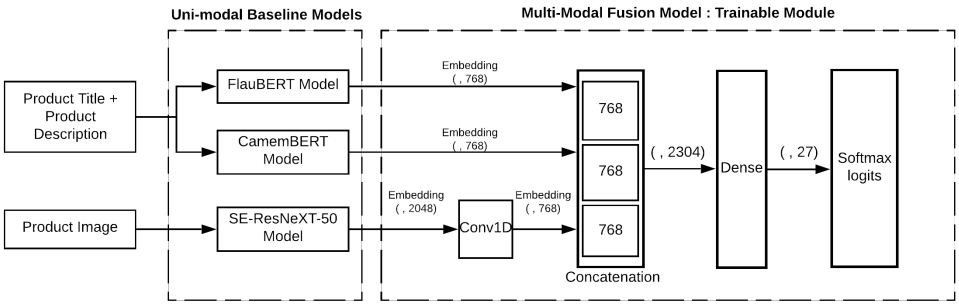

Figure 1: Multi-Modal Representation: Concatenation Ensemble

Addition Ensemble: In this method, we add the textual representations X_{t_1} & $X_{t_2} \in R^{d2}$ and the reduced image representation $X'_{im} \in R^{d2}$ across the dimension d_2. The final feature $X_{avg} = (X_{t_1} + X_{t_2} + X'_{im})$, where $X_{avg} \in R^{d_2}$

Concatenation Ensemble : In this method, we concatenate the textual representations X_{t_1} & $X_{t_2} \in R^{d2}$ and the reduced image representation $X'_{im} \in R^{d2}$ to obtain the final feature $X_{conc} =< X_{t_1}, X_{t_2}, X'_{im} >$, where $X_{conc} \in R^{d_2}$ and <> is the concatenation operation. This architecture is presented in Figure 1.

7 BOOSTED LATE-FUSION MODEL

In the Boosted Late-Fusion Method, a lightGBM [4] gradient boosting multi-class classification model was used to learn the weighted contribution for class probability outputs from each model. Late-fusion method is the second level model on top of the trained uni-modal baselines and multi-modal fusion method. Inputs to this model are the final softmax logits/probability outputs from first level methods namely FlauBERT, CamemBERT, SE-ResNext-50, and the multi-modal concatenation ensemble model. Late-fusion boosting model learns the suitable weights to combine the probability outputs per product instance from the four different models mentioned above to output the final product classification. We further study the weight effects of this model in the ablation study section.

8 EXPERIMENTAL SETUP

In this section, we describe the implementation details used to train the proposed single-modal and multi-modal approaches. Our implementation is based on pytorch [11] framework,we build our text based models using Transformers [14] for our experiments involving FlauBERT and CamemBERT. We used FlauBERT base architecture with cased vocabulary which has 138M parameters and CamemBERT base architecture which contains 110M parameters. FlauBERT and CamemBERT models were trained with a batch size of 32 and a sequence length of 256 for 12 epochs. SE-ResNeXt-50 model was trained with a batch size of 64 for 10 epochs. The multi-modal representation learning module was fine-tuned using the pre-trained weights from the trained uni-modal baseline models. The uni-modal network layers were freezed and we trained our proposed multi-modal architecture using the feature fusion methods described in the earlier section, followed by the softmax layer to perform product classification. We used categorical cross entropy minimization objective as loss function and AdamW optimizer with the learning rate value of 2e-5 for the text models, Adam optimizer and 1e-4 for image model to train the networks. For training the first level models we used 90% of the labelled data and the rest 10% was kept aside for validation. We report and compare the performance of our uni-modal text and image baseline networks along with multi-modal joint learning models on this 10% validation dataset which comprises of 8492 product instances as shown in Table 1. Further, we used this 10% validation dataset (8492 products) to train and validate our second level late-fusion boosted model.

Table 1: Macro F1-Score of Models on Validation dataset

Model Type	Model	Validation Set
Base Models	FlauBERT	89.37
	CamemBERT	89.21
	SE-ResNeXt	61.44
Multi-Modal	Addition Ensemble	90.26
	Concatenation Ensemble	**90.93**

9 ABLATION STUDY AND ERROR ANALYSIS

Our proposed multi-modal concatenation ensemble method corrects the misclassified products from the text-based models in two ways: predicting the correct top-level category which constitutes 38% of the total corrected products and also assigning the right product type code within the same top level category which is 62% of the total corrected products. This proves that relying only on textual features alone is insufficient. Hence, adding the image features helps us to leverage the complementary information, thereby enhancing the overall product classification results.

In our study, we analyse the learned weights of late-fusion boosted model to understand the contribution from each input model which are FlauBERT, CamemBERT, SE-ResNeXt, and multi-modal concatenated model. SHapley Additive exPlanations [9] (SHAP) value is the average marginal contribution of a feature value across all possible coalitions. We utilise SHAP to assign the importance values for probability outputs from the four input models used in late-fusion method per product instance.

Table 2: Qualitative analysis of the proposed methodology. S_{t_1}, S_{t_2}, S_{im} & S_{con} denote the SHAP values for probability outputs of FlauBERT, CamemBERT, SE-ResNeXt and multi-modal concatenated model respectively.

Integer_id	Product Title	True-label	Prediction	Analysis
47740	Marvel Super-Héros : La Collection Officielle Hors-Série # 7 : Le Rhino (Spider-Man) Eaglemoss 2008	1140	1140	Text and image uni-modal networks predict books and child category respectively. Only multi-modal concatenation (MMC) model predicts the right category,entertainment and correct product type code. Highest S_{con} = 2.54 suggests that MMC model drives the prediction.
51782	Super Cassette Vision Wheelie Racer	2060	2060	FlauBERT and MMC model predict child category. Image model predicts correct category, entertainment but wrong product type code. In contrast, CamemBERT model precisely predicted the product code and the category. S_{t_2} = 2.82 being the highest value implies that CamemBERT model dictates the prediction.
38476	Colonne Suspendue Couleur Gris Mat Elona	50	50	All models except image model predict the correct category, household. However, MMC model inaccurately predicts the product type code. FlauBERT and CamemBERT predict the true product type code and category. Combined contribution from text models, S_{t_1} + S_{t_2} = 2.87 pushes prediction to the correct product type code, evident from Figure 2.
76704	Le Téléphone Le Microphone Et Le Phonographe	2705	2705	All models predict the category as books accurately. However, only image model gets the correct product type code. Dominant S_{im} = 3.25 suggests that correct prediction for this product instance is driven by image model.

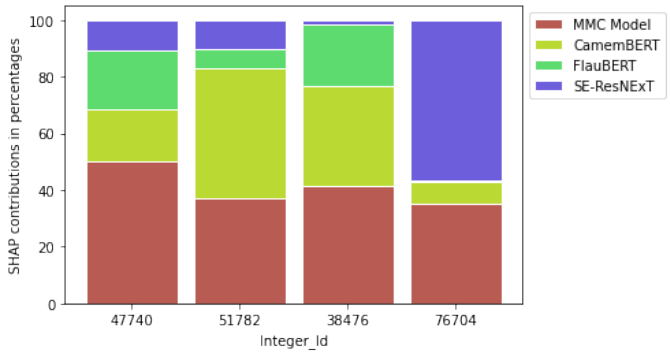

Figure 2: SHAP value distribution for Ablation study

10 RESULTS AND DISCUSSION

SE-ResNext-50 achieves an F1-score of 61.44 on the validation dataset. Product text classification models, FlauBERT and CamemBERT attain 89.37 and 89.21 F1-score respectively on the same validation dataset. For this product classification task, the greater performance of the text based models with a difference in F1-score of around 30% illustrates that product title and description features are more important than product image features for a larger part of the product catalog. We evaluate the contribution of our proposed multi-modal fusion methods for eCommerce product classification by comparing the performance with the uni-modal text and image networks. Addition and concatenation fusion methods reach an F1-score of 90.26 and 90.93 respectively, which indicates that merging the two modalities boosts the performance compared to uni-modal networks. Finally, late-fusion boosted model achieved a macro F1-score of 91.96 on the validation set and further 91.94 and 90.53 on the phase 1 test dataset and phase 2 test dataset respectively.

11 CONCLUSION

In this work, we propose an approach for eCommerce product classification using feature fusion method to leverage the multi-modal information and the late-fusion boosting method to selectively learn from the uni-modal and multi-modal models. From our ablation study, we infer that late fusion method assigns higher weightage to multi-modal method when uni-modal information is insufficient to determine the product category and product type code, and vice-versa. The proposed architecture achieves state of the art result on the benchmark dataset. Our proposed methodology achieved a macro F1-score of 91.94 on the phase 1 test dataset which is the top-scoring submission and third position on the scoreboard for phase 2 test dataset with macro F1-score of 90.53.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

arXiv preprint arXiv:1810.04805 (2018).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[3] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-Excitation Networks. *CoRR* abs/1709.01507 (2017). arXiv:1709.01507 http://arxiv.org/abs/1709.01507

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NIPS*.

[5] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient Large-Scale Multi-Modal Classification. *CoRR* abs/1802.02892 (2018). arXiv:1802.02892 http://arxiv.org/abs/1802.02892

[6] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR* abs/1408.5882 (2014). arXiv:1408.5882 http://arxiv.org/abs/1408.5882

[7] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. FlauBERT: Unsupervised Language Model Pre-training for French. arXiv:1912.05372 [cs.CL]

[8] 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. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[9] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[10] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. arXiv:1911.03894 [cs.CL]

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[12] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[13] Pedro Javier Ortiz Suarez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures *(Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019)*, Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi (Eds.). Leibniz-Institut für Deutsche Sprache, Mannheim, 9 – 16. https://doi.org/10.14618/ids-pub-9021

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[15] Yandi Xia, Aaron Levine, Pradipto Das, Giuseppe Di Fabbrizio, Keiji Shinzato, and Ankur Datta. 2017. Large-Scale Categorization of Japanese Product Titles Using Neural Attention Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 663–668. https://www.aclweb.org/anthology/E17-2105

[16] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. arXiv:1611.05431 [cs.CV]

[17] 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:1611.09534 [cs.CV]