Multimodal Item Categorization Fully Based on Transformers

Lei Chen[*], **Hou Wei Chou**[*], **Yandi Xia, Hirokazu Miyake**
Rakuten Institute of Technology
Boston, MA, USA
{lei.a.chen,houwei.chou,yandi.xia,hirokazu.miyake}@rakuten.com

Abstract

The Transformer has proven to be a powerful feature extraction method and has gained widespread adoption in natural language processing (NLP). In this paper we propose a multimodal item categorization (MIC) system solely based on the Transformer for both text and image processing. On a multimodal product data set collected from a Japanese e-commerce giant, we tested a new image classification model based on the Transformer and investigated different ways of fusing bi-modal information. Our experimental results on real industry data showed that the Transformer-based image classifier has performance on par with ResNet-based classifiers and is four times faster to train. Furthermore, a cross-modal attention layer was found to be critical for the MIC system to achieve performance gains over text-only and image-only models.

1 Introduction

Item categorization (IC) is a core technology in modern e-commerce. Since there can be millions of products and hundreds of labels in e-commerce markets, it is important to be able to map these products to their locations in a product category taxonomy tree efficiently and accurately so that buyers can easily find the products they need. Therefore, IC technology with high accuracy is needed to cope with this demanding task.

Products can contain text (such as titles) and images. Although most IC research has focused on using text-based cues, images of products also contain useful information. For example, in some sub-areas like fashion, the information conveyed through images is richer and more accurate than through the text channel. In this paper, we propose an MIC model entirely based on the Transformer architecture (Vaswani et al., 2017) for achieving a simplification of the model and faster training speed. We conducted experiments on real product data collected from an e-commerce giant in Japan to (a) test the performance of the Transformer-based product image classification, and (b) systematically compare several bi-modal fusion methods to jointly use both text and image cues.

[*] Equal contributor

2 Related works

(Zahavy et al., 2016) is a seminal work on MIC where multi-label classification using both titles and images was conducted on products listed on the Walmart.com website. They used a convolutional neural network to extract representations from both titles and images, then designed several policies to fuse the outputs of the two models. This led to improved performance over individual models separately. Since this work, further research has been conducted on MIC such as (Wirojwatanakul and Wangperawong, 2019; Nawaz et al., 2018).

Recently, a MIC data challenge was organized in the SIGIR'20 e-commerce workshop[1]. Rakuten France provided a dataset containing about $99K$ products where each product contained a title, an optional detailed description, and a product image. The MIC task was to predict 27 category labels from four major genres: books, children, household, and entertainment. Several teams submitted their MIC systems (Bi et al., 2020; Chordia and Vijay Kumar, 2020; Chou et al., 2020). A common solution was to fine-tune pre-trained text and image encoders to serve as feature extractors, then use a bi-modal fusion mechanism to combine predictions. Most teams used the Transformer-based BERT model (Devlin et al., 2019) for text feature extraction and ResNet (He et al., 2016) for image feature extraction, including the standard ResNet-

[1] https://sigir-ecom.github.io/ecom2020/data-task.html

Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), pages 111–115

152 and the recently released Big Transfer (BiT) model (Kolesnikov et al., 2020). For bi-modal fusion, the methods used were more diverse. Roughly in order of increasing complexity, the methods included simple decision-level late fusion (Bi et al., 2020), highway network (Chou et al., 2020), and co-attention (Chordia and Vijay Kumar, 2020). It is interesting to note that the winning team used the simplest decision-level late fusion method.

In other recent work, a cross-modal attention layer which used representations from different modalities to be the key and query vectors to compute attention weights was studied. In (Zhu et al., 2020), product descriptions and images were jointly used to predict product attributes, e.g., color and size, and their values in an end-to-end fashion. In addition, based on the fact that product images can contain information not clearly aligned with or even contradicting the information conveyed in the text, a special gate was used to control the contribution of the image channel. A similar idea was used in (Sun et al., 2020) on multimodal named entity recognition research on Twitter data.

Although the field has converged on using Transformer-based models for processing text in recent years, ResNet-based image processing is still the dominant approach in MIC research. One immediate difficulty in combining the two types of models is the big gap between the training speeds. Owing to the superior parallel running capability enabled by self-attention in the Transformer architecture, text encoder training is much faster than the image encoder, and the training bottleneck of the MIC system becomes solely the image encoder. In addition, using two different deep learning architectures simultaneously makes building and maintaining MIC systems more complex. One solution is to use Transformers as the encoder of choice for both modalities. Furthermore, a detailed comparison of different fusion methods on large-scale multimodal industry product data is still missing. Our work addresses these two directions of research.

3 Model

Our MIC model is depicted in Figure 1[2]. It consists of feature extraction components using a Transformer on uni-modal channels (i.e., text titles and images), a fusion part to obtain multimodal representations, and a Multi-Layer Perceptron (MLP) head to make final predictions.

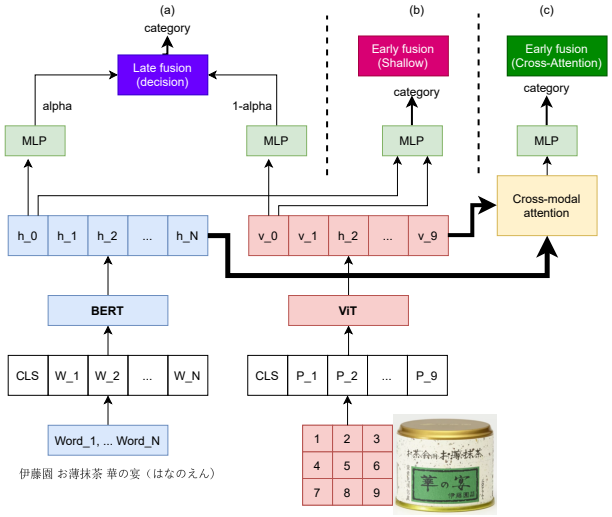

Figure 1: Our Transformer-based MIC system consists of a BERT model to extract textual information and a ViT model to extract visual information. Three different types of multimodal fusion methods are compared, including (a) late fusion, (b) early fusion by concatenating textual and image representations (shallow), and (c) early fusion by using a cross-modal attention. Wide arrows indicate that the entire sequence, e.g., h_0 to h_N, is used in the computation. For illustration we show 3×3 patches for ViT but in our actual implementation a higher P was used.

3.1 BERT text model

We fine-tuned a Japanese BERT model (Devlin et al., 2019) trained on Japanese Wikipedia data. The BERT model encodes a textual product title, $\mathbf{x} = ([CLS], x_1, ..., x_N)$, into text representation sequence $\mathbf{h} = (h_0, h_1, ...h_N)$, where h_i is a vector with a dimension of 768.

3.2 ViT image model

Although originally developed for NLP applications, in recent years the Transformer architecture (Vaswani et al., 2017) has been increasingly applied to the computer vision domain. For example, (Han et al., 2020) is a recent survey paper listing many newly emerging visual models using the Transformer.

Among the many visual Transformer models we used the ViT model (Dosovitskiy et al., 2020), which is a pure Transformer that is applied directly on an image's $P \times P$ patch sequence. ViT utilizes the standard Transformer's encoder part as an image classification feature extractor and adds a MLP head to determine the image labels. The ViT model

[2]The image of the can of tea is from https://item.rakuten.co.jp/kusurinokiyoshi/10016272/

was pre-trained using a supervised learning task on a massive image data set. The size of the supervised training data set impacts ViT performance significantly. When using Google's in-house JFT 300M image set, ViT can reach a performance superior to other competitive ResNet (He et al., 2016) models.

The ViT model encodes the product image. After converting a product image to $P \times P$ patches, ViT converts these patches to visual tokens. After adding a special [CLS] visual token to represent the entire image, the $M = P \times P + 1$ long sequence is fed into a ViT model to output an encoding as $\mathbf{v} = (v_0, v_1, v_2, ...v_M)$, where $M = P \times P$.

3.3 Multimodal fusion

The fusion method plays an important role in MIC. In this paper we compared three methods, corresponding to Figure 1 (a), (b), and (c).

3.3.1 Late fusion

The simplest fusion method is combining the decisions made by individual models directly (Bi et al., 2020; Chou et al., 2020). We used weights α and $1 - \alpha$ to interpolate the probabilities estimated by BERT and ViT models. The α value was chosen using a held-out set.

3.3.2 Early fusion – shallow

The [CLS] token, the first token of every input sequence to BERT and ViT, is used to provide a global representation. Therefore we can concatenate the two encoded [CLS] tokens to create a multimodal output. The concatenated feature vectors are sent to an MLP head for predicting multi-class category labels. This method is called a *shallow fusion* (Siriwardhana et al., 2020).

3.3.3 Early fusion – cross-modal attention

A cross-modal attention layer provides a more sophisticated fusion between text and image channels (Zhu et al., 2020; Sun et al., 2020). Cross-modal attention is computed by combining Key-Value (K-V) pairs from one modality with the Query (Q) from the other modality. In addition, (Zhu et al., 2020) used a gate to moderate potential noise from the visual channel.

Specifically, the multimodal representation $\mathbf{h}'$ is computed from the addition of the self-attention (SA) version of text representation $\mathbf{h}$ and the cross-modal attention version by considering the visual representation $\mathbf{v}$ as

$$\mathbf{h}' = SA(\mathbf{h}, \mathbf{h}, \mathbf{h}) + VG \odot SA(\mathbf{h}, \mathbf{v}, \mathbf{v}), \quad (1)$$

where

$$SA(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{(W_Q\mathbf{q})(W_K\mathbf{k})^T}{\sqrt{d_k}}\right) W_V\mathbf{v}, \quad (2)$$

$$VG_i = \sigma(W_1 h_i + W_2 v_0 + b), \quad (3)$$

W_Q, W_K, and W_V are trainable query, key, and value parameters, d_k is the dimension of the key vectors, and the visual gate, VG, can be learned from both the local text representations h_i and global visual representation v_0, with W_1, W_2, and b as trainable parameters. The category label prediction $\hat{y}$ is determined as

$$\hat{y} = \text{softmax}\left(W_3 \sum_i h'_i\right), \quad (4)$$

where W_3 is a trainable parameter.

4 Experiment

4.1 Setup

Data set: Our data consisted of about 500,000 products from a large e-commerce platform in Japan, focusing on three major product categories. Our task, a multi-class classification problem, was to predict the leaf-level product categories from their Japanese titles and images. Further details of our data set are shown in the left part of Table 1. We used the macro-averaged F1-score to evaluate model performance.

Models: We compared the following models.

- Text-only: Japanese BERT model[3] fine-tuned on product titles.
- Image-BiT: BiT image model (Kolesnikov et al., 2020) fine-tuned on product images. In particular, we used BiT-M.[4] BiT showed a considerable performance advantage than other conventional ResNet models in the SIGIR'20 MIC data challenge (Chou et al., 2020).

[3] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[4] https://tfhub.dev/google/bit/m-r152x4/1

Root genre	Class size	Train size	Test size	Ma-F1 (BiT)	Ma-F1 (ViT)
Beverages (B)	32	29,269	7,332	**0.666**	0.610
Appliances (A)	280	200,552	50,283	0.574	**0.639**
Men's Fashion (M)	70	228,148	57,077	0.715	**0.733**

Table 1: Summary of our data set obtained from a large e-commerce platform in Japan. Right two columns report image classification macro-F1 values using BiT and ViT models, respectively.

- Image-ViT: ViT image model (Dosovitskiy et al., 2020) fine-tuned on product images. We used ViT-L-16.[5] 16 means that we used 16 × 16 patches when feeding images.
- Fusion: The **late** fusion method described in Section 3.3.1 and depicted in Figure 1 (a), the **early** fusion method described in Section 3.3.2 and depicted in Figure 1 (b), and the **cross-modal** fusion method described in Section 3.3.3 and depicted in Figure 1 (c).

Implementation details: Our models were implemented in PyTorch using a GPU for training and evaluation. The AdamW optimizer (Loshchilov and Hutter, 2017) was used. Tokenization was performed with MeCab.[6]

4.2 Result

Table 1 reports on macro-F1 values for the three genres using the ResNet-based BiT vs. Transformer-based ViT. ViT shows higher performance compared to BiT on two of the three genres. In addition, consistent with the speed advantage reported in (Dosovitskiy et al., 2020), we also observed that the training for ViT is about four times faster than BiT. This is critical for an MIC system deployable in industry since image model training time is the main bottleneck.

Model	F1 (B)	F1 (A)	F1 (M)
Text-BERT	0.718	0.733	0.802
Image-ViT	0.610	0.639	0.733
Fusion-late	0.725	0.709	0.814
Fusion-early	0.714	0.726	0.788
Fusion cross-modal	**0.729**	**0.740**	**0.815**

Table 2: Macro-F1 on the three product genres. Uni-modal models, i.e., BERT text model and ViT image model, and different fusion models are compared.

Table 2 reports on uni-modal model performance, i.e., text-BERT and image-ViT separately, as well as the results of fusing these models in various ways. We found that the early (shallow) fusion method leads to poor model performance. One possible reason is that product images used in e-commerce product catalogs sometimes do not appear to be clearly related to its corresponding titles. For example, a bottle of wine may be packaged in a box and its image only shows the box. We also found that late (decision) fusion does not lead to consistent gains. In the appliance genre, we found that the fused model was worse than the text model. On the other hand, the cross-modal attention fusion method showed consistent gains over both the text and image models separately on all three genres.

5 Discussion

Although various approaches have been explored in MIC research, we found that a MIC system built entirely out of the Transformer architecture was missing. Combining the well-established BERT text model and the newly released ViT image model, we proposed an all-Transformer MIC system on Japanese e-commerce products. From experiments on real industry product data from an e-commerce giant in Japan, we found that the ViT model can be fine-tuned four times faster than BiT and can have improved performance. Furthermore, fusing both text and image inputs in an MIC setup using the cross-modal attention fusion method led to model performance better than each model separately, and we found that this fusion method worked better than late fusion and the early (shallow) fusion of simply concatenating representations from the two modalities.

There are several directions to extend the current work in the future, including (1) considering jointly modeling texts and images in one Transformer model like FashionBERT (Gao et al., 2020), and (2) using self-training to go beyond the limit caused by the size of labeled image data for the image model.

[5] https://github.com/asyml/vision-transformer-pytorch

[6] https://taku910.github.io/mecab/

References

Ye Bi, Shuo Wang, and Zhongrui Fan. 2020. A Multimodal Late Fusion Model for E-Commerce Product Classification. *arXiv preprint arXiv:2008.06179*.

V. Chordia and B.G. Vijay Kumar. 2020. Large Scale Multimodal Classification Using an Ensemble of Transformer Models and Co-Attention. In *Proc. SIGIR'20 e-Com workshop*.

H. Chou, Y.H. Lee, L. Chen, Y. Xia, and W.T. Chen. 2020. CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval. In *Proc. SIGIR'20 e-Com workshop*.

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]*. ArXiv: 1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, and Yixing Xu. 2020. A Survey on Visual Transformer. *arXiv preprint arXiv:2012.12556*.

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*, pages 770–778.

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]*. ArXiv: 1912.11370.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Muhammad Umer Anwaar, and Ignazio Gallo. 2018. Learning fused representations for large-scale multimodal classification. *IEEE Sensors Letters*, 3(1):1–4.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Pasawee Wirojwatanakul and Artit Wangperawong. 2019. Multi-Label Product Categorization Using Multi-Modal Fusion Models. *arXiv preprint arXiv:1907.00420*.

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*.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal Joint Attribute Prediction and Value Extraction for E-commerce Product. *arXiv preprint arXiv:2009.07162*.