ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Multimodal Approach for Harmonized System Code Prediction

Otmane Amel[1*], Sédrick Stassin[1*], Sidi Ahmed Mahmoudi[1] and Xavier Siebert[1 †]

University of Mons - ILIA Unit
Mons - Belgium

Abstract. The rapid growth of e-commerce has placed considerable pressure on customs representatives, prompting advanced methods. In tackling this, Artificial intelligence (AI) systems have emerged as a promising approach to minimize the risks faced. Given that the Harmonized System (HS) code is a crucial element for an accurate customs declaration, we propose a novel multimodal HS code prediction approach using deep learning models exploiting both image and text features obtained through the customs declaration combined with e-commerce platform information. We evaluated two early fusion methods and introduced our MultConcat fusion method. To the best of our knowledge, few studies analyze the feature-level combination of text and image in the state-of-the-art for HS code prediction, which heightens interest in our paper and its findings. The experimental results prove the effectiveness of our approach and fusion method with a top-3 and top-5 accuracy of 93.5% and 98.2% respectively.

1 Introduction

Repeated legislative changes, along with the massive increase in e-commerce flows, have significantly altered declaration procedures and raised the risks that customs representatives must bear. Today, there is a clear need to reduce the declarative risks in the customs field. The tariff classification of products is a crucial component in customs declarations. Customs representatives rely on Harmonized System (HS) codes to classify the products for declaration based on information provided by their clients. The HS codes provide a hierarchical categorization of the products using 6 digits, with each product assumed to belong to a specific category. It starts with two digits (called section or HS2) that define broad categories of products and continues up to six international digits (subheading or HS6) which describes a specific product with precision[1]. Additional digits are added to classify products down to the national level, which determines the exact tax rate for a product. Mistakes made during the customs declaration process can lead to incorrect HS codes being entered into declarations. As a result, an incorrect tax rate may be paid at the end. This emphasizes the importance of the HS code in customs declarations. Therefore, artificial intelligence (AI) systems that seek to verify the classification of goods are perfectly suited

*These authors contributed equally to this work

†The authors thank the support of Infortech institute and the E-origin project funded by the Walloon Region within the pole of logistics in Wallonia.

[1]Example of the Belgian governmental free access database of HS code nomenclature: https://eservices.minfin.fgov.be/extTariffBrowser/browseNomen.xhtml?suffix=80&lang=EN

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

to assist users such as customs representatives. In this paper, we propose a multimodal model using different textual features as well as images, retrieved from customs declarations coupled with information extracted from the e-commerce website related to the purchase of the product. This model classifies six-digit HS codes (HS6) with an accuracy of top-3 and top-5 accuracy of 93.5% and 98.2% respectively. The remainder of this paper is structured as follows. Section 2 presents the related works. Section 3 introduces the proposed approach. Section 4 presents the experimental results. Finally, Section 5 presents a conclusion and gives some perspectives for future work.

2 Related Works

In this Section, we present works that combine text and image modalities in the e-commerce field. Then, we focus in particular on the papers related to the HS code prediction problem itself. However, the necessity for labeled data and the fact that datasets used in research publications come from private sources and remain secret are recurring challenges for results comparison. As a result, it is hard to compare results in the field accurately. The work of Zahavy et al. [1] uses a dataset of 1.2 million items collected from Walmart.com containing the image, the title as well as the shelf (product categories) to predict between 2890 possibilities. Their experiments proved the effectiveness of late fusion using a learned policy based on class probabilities to combine the convolutional neural networks (CNN) decisions reaching 70.2% for text, 56.1% for image and 71.8% for their combination. Chen et al. [2] employed a dataset of 500,000 e-commerce products to predict the category based on Japanese titles and images. Their best result ranged between 72.9 and 81.5% according to the categories predicted, obtained using vision transformers [3] (ViT) and Japanese BERT [4] with the use of cross-modal attention as an early fusion method for the modalities. In the field of multimodal HS code prediction, Turhan et al. [5] proposed a topic modeling approach based on the product description and image. They offer the user the most similar images with the corresponding HS code to facilitate their choice. Their approach achieves a top-10 accuracy of 87.1% and 78.9% for HS4 and HS6, respectively. Another work by Li and Li [6] uses a separately trained text and image CNN. By grouping six similar HS codes into four classes with 2500 data each, they obtain an accuracy of 93.4% for the text model, 76.9% for the image model, and 93.9% for the combination of the two using a late fusion based on weights calculated using the model accuracies. To the best of our knowledge, few studies analyze the combination of text and image in the state-of-the-art, which heightens interest in our paper and its findings. We differ from the previous works in the following way. 1) We study the combination of image and multiple text modalities to enhance HS code prediction. 2) We conduct a comparative analysis of fusion methods at the feature level (early fusion) and we propose our improved early fusion method MultConcat using arithmetic operations inspired by [7]. 3) We examine the impact of the visual modality through a comparative analysis of two transformer-based and CNN feature extractors.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

3 Methodology

We present our proposed architecture for HS code prediction as depicted in Figure 1. In this work, two types of modalities are available: text and image. The visual modality consists of product image denoted I, and the following are textual modalities: invoice description denoted D from the customs declaration, product title denoted T, and product category denoted C. The I, T, and C features are extracted from the e-commerce platform. We employed these encoders for their renowned feature extraction capabilities [1, 2]: Resnet50 [8], ViT [3], and CLIP's image encoder [9]. To obtain the final representation of the features, we extracted the 2048 intermediate features from the *avgpool* layer of Resnet50 and used the classification token T_{cls} for the two transformers models. The textual modalities are fed to the pre-trained model SimCSE [10], a widely used sentence embedding extractor of 768 dimensions. Next, we merge the modalities with different early fusion methods such as a simple concatenation [11] (Concat) of each modality representation M_i or a multimodal low-rank tensor fusion (LMF) [12]. Inspired from [7] we propose an enhanced multiplication fusion approach called MultConcat. After projecting linearly each of the N modalities M_i in the same vector space through a hidden layer resulting in out_i (see Eq. 1 where W_i and b_i are learnable parameters), MultConcat is obtained based on the concatenation $\|$ of two terms (see Eq. 2): a concatenated representation of each out_i called C, and an element-wise multiplication $\odot$ (or called Hadamard product) of each out_i called Z (see Eq. 3).

$$out_i = \mathrm{ReLU}(W_i \times M_i + b_i) \tag{1}$$

$$\mathrm{MultConcat} = C \parallel Z \tag{2}$$

$$C = \|_{i=1}^{N} out_i \qquad\qquad Z = \odot_{i=1}^{N} out_i \tag{3}$$

Finally, the resulting multimodal representation vector is fed to a one-layer classifier for HS code prediction.

4 Experiments

4.1 Dataset

The dataset consists of 2144 customs declarations provided by our project partner e-Origin[2], having a total of 16 distinct HS6 codes along with customs declaration information as well as additional ones provided by the marketplace from where the goods originated. The database underwent a preprocessing step, during which we eliminated punctuation, special characters, and digit numbers from the textual columns. Subsequently, we observed miswritten or concatenated words that needed unique attention. To address this issue, we utilized a

[2]https://eorigin.eu/

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

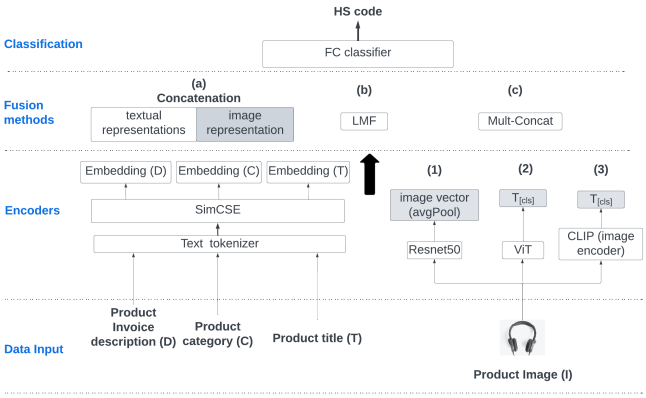

Fig. 1: Proposed multimodal architecture for HS Code prediction.

blend of word-separating tools from the WordSegment library[3] and a Python-based spell-checker[4]. In addition, product images were normalized and resized to fit the image encoders' requirements. Finally, the dataset was split into train, test, and validation sets with a ratio of 80%, 10%, and 10% respectively.

4.2 Setup

The comparative analysis was performed using SimCSE public checkpoint[5], in addition to *vit-base-patch16-224-in21k* weights for ViT[6], and *ViT-B/32* for CLIP image encoder[7]. For fusion methods, we took the implementation of LMF using the framework Multibench [13] with a decomposition rank factor of 16. Our models were trained and evaluated on GPU resources using the PyTorch framework. The Adam optimizer [14] was used with a $1e^{-4}$ learning rate, a total of 100 epochs, and an early stopping set to 10 epochs.

4.3 Discussion

In Table 1 below we provide our comparative results by varying the fusion methods and the image encoders. The top-k accuracies are achieved on the test sets. The goal of this analysis is to assess whether multimodal architectures based on early fusion methods help to boost the HS recommendation results compared to baseline models represented in the bottom part of the table. One can notice an improvement by a large margin with the multimodal networks except those employing the LMF fusion technique, this might be interpreted by a loss of cross-modal information during the decomposition phase. Additionally, our proposed fusion method MultConcat comes ahead of the other fusion methods since it

[3]https://grantjenks.com/docs/wordsegment/
[4]https://pyspellchecker.readthedocs.io/
[5]*sup-simcse-bert-base-uncased* weights from https://github.com/princeton-nlp/SimCSE
[6]https://huggingface.co/google/vit-base-patch16-224-in21k
[7]https://huggingface.co/sentence-transformers/clip-ViT-B-32

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

gives overall better performance than simple concatenation or LMF. Moreover, different image encoders were employed for the visual modality I, and we notice that ViT gives the highest top-1 accuracy of 65.3%. However, ResNet50 yields better results in terms of top-3 and top-5 accuracy metrics. Although by a low margin compared to ViT, this suggests that ResNet50 encoders enhance the flexibility of our multimodal architectures when employed for recommendation tasks. This is the desired characteristic of the model deployed during inference, as customs representatives need the flexibility to select the correct HS code between multiple choices. It is worth noting that adding the visual modality I to the initial invoice description D improves the top-1 accuracy by 8.2% compared to the unimodal approach. When set against the combined textual modalities (T, D, C), this improvement is marginal (0.6%). The limited enhancement might result from noisy image products that aren't directly related to the textual information, hindering overall performance.

Fusion method	**Encoder**		**Modality**	**Top-k**		
	Image	**Text**		**k=1**	**k=3**	**k=5**
MultConcat	ViT	SimCSE	I,T,D,C	**0.653**	0.929	0.977
Concat			I,T,D,C	0.624	0.924	0.977
LMF			I,T,D,C	0.088	0.188	0.347
MultConcat	ResNet50	SimCSE	I,T,D,C	0.612	**0.935**	**0.982**
Concat			I,T,D,C	0.571	0.924	0.977
LMF			I,T,D,C	0.047	0.182	0.241
MultConcat	CLIP	SimCSE	I,T,D,C	0.629	0.918	0.977
Concat			I,T,D,C	0.624	0.924	0.977
LMF			I,T,D,C	0.277	0.359	0.477
MultConcat	/	SimCSE	T,D,C	0.647	0.930	0.970
MultConcat	RestNet50	SimCSE	I,D	0.582	0.870	0.924
baseline (unimodal models)						
/	/	SimCSE	D	0.500	0.829	0.906
/	ViT	/	I	0.394	0.729	0.847
/	RestNet50	/	I	0.388	0.688	0.806
/	CLIP	/	I	0.482	0.806	0.894

Table 1: Top-1, Top-3, and Top-5 accuracy of the model according to fusion methods, encoders, and modalities of the dataset used.

5 Conclusion

In this work, we focused on enhancing HS code prediction by leveraging multimodal auxiliary information. We conducted several experiments by varying the fusion methods and the image encoders to find the optimal combination. Experiments demonstrated that Resnet50 yields better results with a top-3 and top-5 accuracy of 93.5% and 98.2% respectively. In addition, we proposed our

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

MultConcat fusion method that performed better than simple concatenation and LMF [12] methods in all trials. The results underscore the effectiveness of using multimodal data for HS code predictions, as it outperforms unimodal solutions by 8.2% in top-1 accuracy. A possible extension to this work could involve quantifying modality contributions using explainability techniques, as well as developing a fusion method capable of handling missing modalities.

References

[1] Tom Zahavy, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor. Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[2] Lei Chen, Houwei Chou, Yandi Xia, and Hirokazu Miyake. Multimodal item categorization fully based on transformer. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 111–115, 2021.

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

[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] Bilgehan Turhan, Gozde B Akar, Cigdem Turhan, and Cihan Yukse. Visual and textual feature fusion for automatic customs tariff classification. In *2015 IEEE International Conference on Information Reuse and Integration*, pages 76–81. IEEE, 2015.

[6] Guo Li and Na Li. Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electronic Commerce Research*, 19:779–800, 2019.

[7] LUCAS SOUZA RODRIGUES, Kenzo Sakiyama, Edson Takashi Matsubara, José Marcato Junior, and Wesley Nunes Gonçalves. Multimodal fusion based on arithmetic operations and attention mechanisms. *Available at SSRN 4292754*.

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

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[11] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[12] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[13] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.