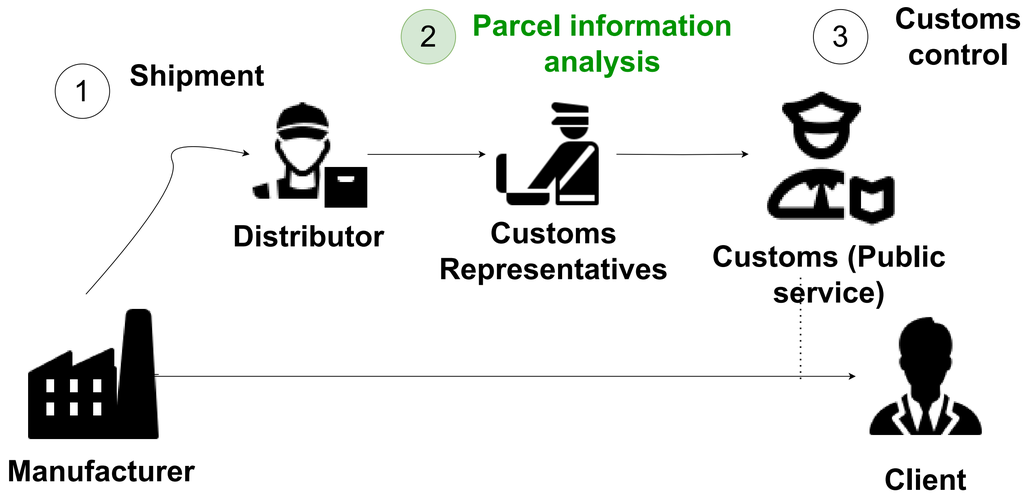
Multimodal Approach for Harmonized System Code Prediction

Otmane Amel, Sédrick Stassin, Sidi-Ahmed Mahmoudi, Xavier Siebert

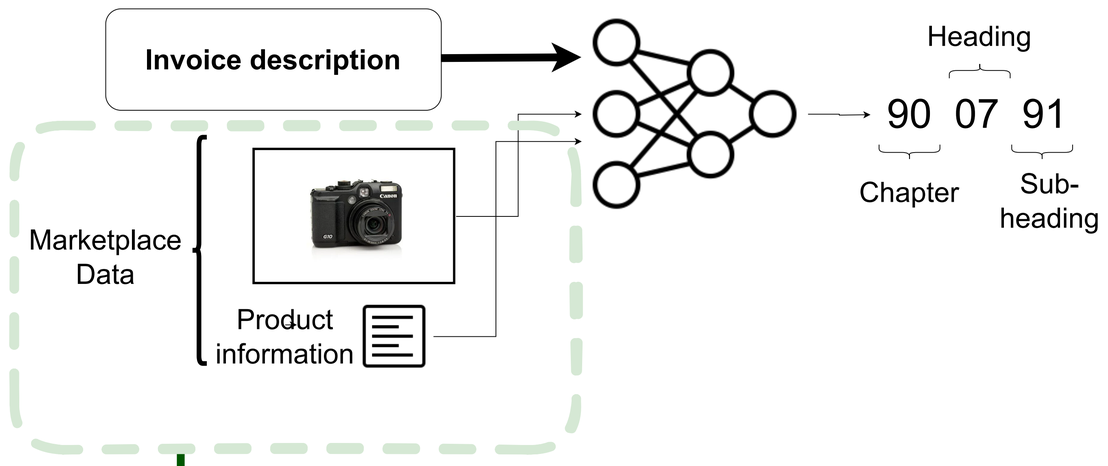
University of Mons
inforTech — INSTITUT DE RECHERCHE EN TECHNOLOGIES DE L'INFORMATION ET SCIENCES DE L'INFORMATIQUE DE L'UMONS
POLYTECH MONS
e-Origin

1 INTRODUCTION



Problem statement : is it feasable for a multimodal AI-assisted system to aid customs representatives with the analysis of parcel information, such as Harmonized System code prediction ?

2 TASK



Marketplace data, like product images, titles, and categories, can provide more context and details about the primary invoice description, helping the model to identify and categorize those products more thoroughly.

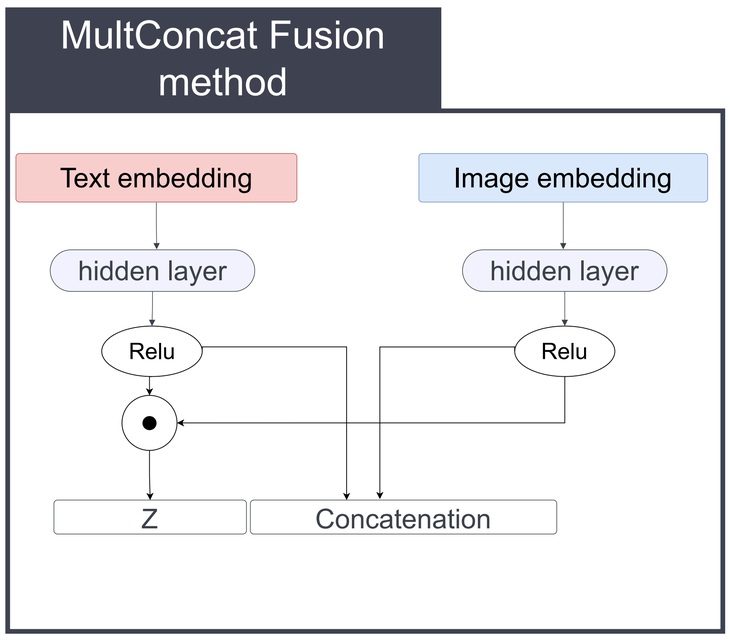
3 DATASET

- The dataset consists of 2144 customs declarations provided by our project partner E-Origin.
- It contains a total of 16 distinct HS code of 6 digits (HS6)
- Each declaration has an invoice description + marketplace metadata
- Metadata consists of product image (visual modality), title, and category

4 Contribution

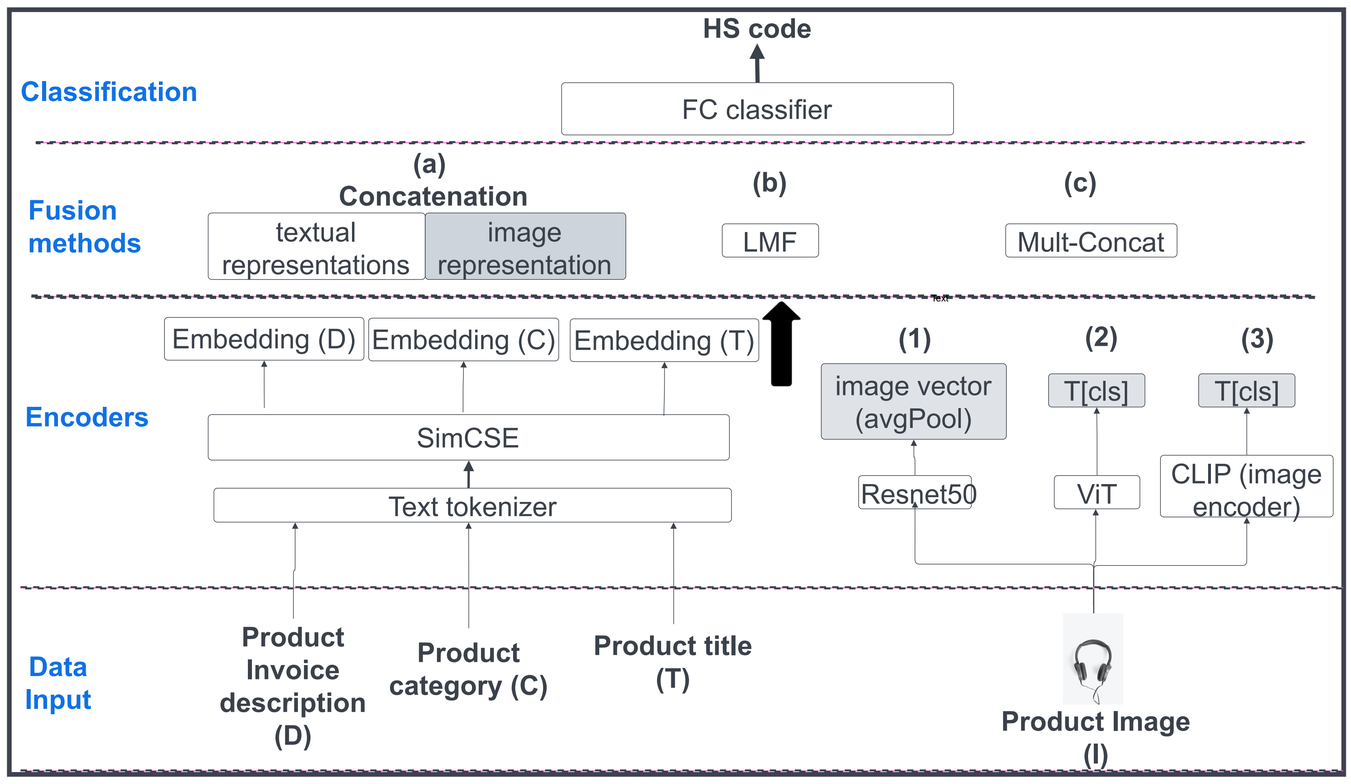
- We study the combination of image and multiple text modalities to enhance HS code prediction
- We conduct a comparative analysis of fusion methods at feature level
- We proposed an adapted fusion method **MultConcat**
- We assessed the visual modality impact

5 METHODOLOGY



MultConcat is obtained based on the concatenation of two terms: a concatenated representation of both modalities in a shared sub-space, and an element-wise multiplication of each of them.

The intuition behind this separation is to preserve modality-specific features while concurrently extracting cross-modal features.



The proposed Multimodal HS code prediction model follows a threefold data flow process: first, we employ encoders for feature extraction of textual and visual modalities, note that three encoders have been tested for image product(I); next, we test three fusion methods to integrate the multimodal features; and finally, we apply a classifier to make HS code predictions based on the combined features.

6 RESULTS

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.

7 CONCLUSION

- Improved HS code prediction using multimodal data
- Best results achieved with Resnet50 image encoder and MultConcat fusion method.
- Outperformed unimodal approaches by 8.2% in top-1 accuracy.
- Future work could focus on modality contributions quantification and handling missing data.

8 CONTACT US

otmane.amel@umons.ac.be **sedrick.stassin@umons.ac.be**
sidi.mahmoudi@umons.ac.be **xavier.siebert@umons.ac.be**

Article QR code