DOI: https://doi.org/10.15276/hait.09.2026.11
UDC 004:89

Cross-modal representation learning for accurate harmonized system code classification in e-commerce systems

Stepan M. Krupa[1)]
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
Yurii P. Kryvenchuk[1)]
ORCID: https://orcid.org/0000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
[1)] Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine

ABSTRACT

Accurate classification of goods according to the Harmonized System remains a critical challenge in international trade and e-commerce due to the complexity of product descriptions, ambiguity of textual data, and variability in product representation. The novelty of this study lies in the development of a cross-modal representation learning approach for automated Harmonized System code classification that integrates both textual and visual product information within a unified framework. By leveraging multimodal data, including product descriptions and images, the proposed system improves classification accuracy and robustness compared to traditional approaches that rely solely on textual information. In addition, the proposed framework enables more reliable identification of product characteristics by aligning semantic and visual representations in a shared feature space, which enhances the model's ability to handle incomplete or ambiguous product descriptions commonly encountered in e-commerce environments. **The methodology is based** on contrastive learning techniques that align semantic representations across modalities, enabling the model to capture deeper relationships between product attributes and Harmonized System codes. Transformer-based encoders are employed for textual feature extraction, while convolutional or vision transformer architectures are used for image representation. A joint embedding space is constructed to facilitate effective cross-modal interaction and classification. **Experimental evaluation** is conducted on a real-world e-commerce dataset, demonstrating that the proposed approach significantly outperforms baseline models in terms of accuracy, precision, and recall. The results highlight the effectiveness of multimodal learning in handling noisy, incomplete, and heterogeneous product data commonly encountered in customs and trade environments. **The proposed framework contributes** to the advancement of intelligent customs classification systems by enhancing automation, reducing human error, and improving compliance in international trade operations. Future work will focus on incorporating explainability mechanisms and extending the model to support multilingual and low-resource scenarios.

Keywords: Harmonized System classification; cross-modal learning; multimodal machine learning; contrastive learning; e-commerce; product classification; transformer models; computer vision; customs automation; trade compliance

***For citation*: Krupa S. M., Kryvenchuk Yu. P. "Cross-modal representation learning for accurate harmonized system code classification in e-commerce systems". *Herald of Advanced of Information Technology*. 2026; Vol.9 No.2: 158–167. DOI: https://doi.org/10.15276/hait.09.2026.11**

INTRODUCTION

In today's rapidly developing e-commerce environment, the volume of international trade transactions is growing significantly, which in turn increases the requirements for the accuracy and speed of product classification. The scientific problem addressed in this study concerns the accurate classification of goods according to the Harmonized System (HS), which has become increasingly important in e-commerce due to the complexity of product descriptions, ambiguity of textual data, and variability in product representation. One of the key elements of this process is the determination of the HS code, which is used for customs clearance, customs payments, and foreign trade statistics.

At the same time, the correct classification of goods remains a difficult task due to the large number of categories, the complex hierarchical structure of the system, and the ambiguity of product descriptions. Traditionally, the classification of HS codes is carried out manually or using semi-formalized rules based on the experience of experts. This approach is laborious, slow, and does not always provide stable quality, especially in conditions of a large flow of goods and constant updating of the assortment in e-commerce [1].

In this regard, machine learning methods are becoming increasingly popular, as they allow automating the classification process based on the analysis of textual descriptions of goods. However, most existing approaches focus exclusively on textual data, such as product names or descriptions. In practice, this data is often incomplete, contains errors, or does not reflect all product characteristics. At the same time, e-commerce systems usually also provide product images, which may contain additional important information for correct classification [2], [23].

Recent advances in multimodal machine learning open up opportunities for the simultaneous use of textual and visual data. In particular, cross-modal learning approaches allow the formation of a joint representation of information from different sources, which contributes to a better understanding of product characteristics. The use of contrastive learning makes it possible to effectively combine different modalities and identify semantic correspondences between them.

This article proposes an approach to automated HS code classification that is based on a cross-modal representation of data and combines textual and visual information about products. The main idea is to create a unified feature space in which heterogeneous data interact with each other, thereby improving classification accuracy even in cases of incomplete or ambiguous descriptions. The proposed approach is aimed at increasing the efficiency of customs processes, reducing classification errors, and improving the quality of automated classification systems used in international trade.

ANALYSIS OF LITERARY DATA

The problem of automated Harmonized System (HS) code classification has attracted increasing attention in recent years, driven by the growing complexity of international trade and the digitalization of customs procedures. Existing research can be broadly categorized into three principal directions: (i) rule-based and knowledge-driven systems, (ii) text-centric machine learning approaches, and (iii) multimodal and cross-modal learning frameworks. Early studies predominantly relied on deterministic, rule-based systems grounded in expert knowledge and predefined taxonomies. These approaches typically utilize keyword matching, decision trees, and manually constructed ontologies to map product descriptions to HS codes. While such systems offer interpretability and domain transparency, their performance is inherently constrained by limited adaptability to evolving product vocabularies and the high maintenance cost associated with rule updates [3], [4], [5]. Moreover, their inability to generalize beyond predefined patterns renders them unsuitable for large-scale, dynamic e-commerce environments. Subsequent advancements have shifted focus toward data-driven methodologies, particularly text-based machine learning models. Traditional approaches employ feature engineering techniques such as term frequency–inverse document frequency (TF-IDF) combined with classical classifiers, including Support Vector Machines and ensemble methods. Although these models demonstrate moderate performance improvements, they remain sensitive to feature sparsity and lexical variability [6], [7]. The emergence of deep learning, and in particular transformer-based architectures such as BERT, has significantly enhanced the ability to capture contextual semantics within product descriptions. These models leverage contextual embeddings to model linguistic dependencies and have been successfully applied to various classification tasks, including trade-related applications. Despite these improvements, text-centric approaches exhibit fundamental limitations when applied to real-world e-commerce data. Product descriptions are often incomplete, noisy, or intentionally simplified, leading to semantic ambiguity and reduced classification reliability. Furthermore, critical product attributes – such as shape, material, or design – may not be explicitly encoded in textual form, thereby constraining the representational capacity of text-only models. In response to these challenges, recent research has explored the integration of multimodal data, particularly the combination of textual and visual information. Multimodal learning frameworks typically employ convolutional neural networks or vision transformers for image feature extraction, alongside transformer-based encoders for text processing. Fusion strategies - ranging from early concatenation to attention-based interaction mechanisms - have been proposed to combine modality-specific representations. Empirical evidence suggests that multimodal approaches can significantly improve robustness and classification accuracy, especially in scenarios characterized by incomplete or noisy textual inputs. More recently, cross-modal representation learning has emerged as a promising paradigm for multimodal integration. In contrast to conventional fusion techniques, cross-modal methods aim to learn a shared embedding space in which semantically related entities from different modalities are closely aligned. Contrastive learning has proven particularly effective in this context, enabling models to maximize agreement between corresponding text–image pairs while minimizing similarity across unrelated samples. A prominent example is CLIP, which demonstrates strong generalization capabilities through large-scale pretraining on aligned image–text data. Nevertheless, the application of cross-modal and contrastive learning techniques to HS code classification remains limited. Existing studies in this domain often focus on general product categorization or recommendation tasks, without addressing the specific challenges associated with the HS taxonomy, such as

hierarchical structure, fine-grained distinctions between categories, and strict regulatory requirements. Furthermore, many approaches do not adequately consider the heterogeneity and noise inherent in real-world trade datasets. A critical analysis of the literature reveals several unresolved issues. First, there is a lack of unified frameworks that effectively integrate multimodal data within the context of HS classification. Second, existing models often fail to generalize across diverse product domains due to insufficient representation learning strategies [8]. Third, the majority of studies do not explicitly address robustness to data imperfections, which are pervasive in practical applications. In light of these limitations, there is a clear need for advanced methodologies that combine the strengths of multimodal learning and cross-modal representation alignment. The present study addresses this gap by proposing a contrastive cross-modal framework tailored specifically to the requirements of HS code classification in e-commerce systems [9], [10].

THE PURPOSE AND OBJECTIVES OF THE RESEARCH

The rapid proliferation of e-commerce platforms and the increasing complexity of international trade operations necessitate the development of advanced, scalable, and accurate methods for automated HS code classification. Despite notable progress in machine learning, existing approaches remain constrained by their reliance on unimodal data representations, predominantly textual, which limits their effectiveness in real-world scenarios characterized by noisy, incomplete, and heterogeneous product information. In this context, the primary purpose of this research is to develop a robust and generalizable framework for HS code classification that leverages cross-modal representation learning to integrate complementary information from multiple data sources. Specifically, this study aims to address the fundamental limitations of traditional and text-centric approaches by incorporating both textual and visual modalities into a unified learning paradigm [11]. The research is motivated by the hypothesis that the joint modeling of semantic and visual features within a shared representation space can significantly enhance classification accuracy and robustness, particularly in cases where individual modalities are insufficient or ambiguous. Furthermore, the study seeks to explore the potential of contrastive learning as an effective mechanism for aligning heterogeneous data representations and improving discriminative feature learning. To achieve this purpose, several research objectives are defined. The first objective is to design a multimodal architecture capable of extracting and encoding informative features from both textual descriptions and product images. This involves the selection and adaptation of appropriate deep learning models, including transformer-based encoders for textual data and visual feature extractors for image data, as well as the development of mechanisms for their effective integration. The second objective is to implement a cross-modal representation learning strategy based on contrastive learning principles. This includes the construction of a joint embedding space in which semantically related text–image pairs are closely aligned, while unrelated pairs are effectively separated. The objective further encompasses the formulation of an appropriate loss function and training procedure that ensures stable convergence and meaningful representation learning. The third objective is to develop a classification mechanism that utilizes the learned multimodal representations for accurate HS code prediction. Particular attention is given to the hierarchical nature of the HS system and the need for fine-grained classification across a large number of categories. The fourth objective is to conduct a comprehensive experimental evaluation of the proposed framework using real-world e-commerce data. This involves benchmarking the model against existing baseline approaches, analyzing performance across multiple metrics (such as accuracy, precision, and recall), and assessing robustness under conditions of noisy or incomplete input data. Finally, the study aims to evaluate the practical implications of the proposed approach for customs automation and trade compliance. This includes assessing its potential to reduce manual effort, minimize classification errors, and improve operational efficiency in international trade systems. Collectively, these objectives are intended to provide a systematic and empirically validated contribution to the field of automated trade classification, advancing the application of cross-modal machine learning techniques in real-world e-commerce and customs environments.

RESEARCH METHODS

In this study, the task of automated classification of goods according to the codes of the HS is considered as a task of multiclass classification using multimodal data. Unlike traditional approaches based solely on textual information, the proposed formulation takes into account both textual descriptions of goods and their visual representations [12], [13].

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Formally, a data set is defined as:

$$D = \{(x_i^t, x_i^v, y_i)\}_{i=1}^{N}, \quad (1)$$

where x_i^t is text description of the product (name, specification), x_i^v is corresponding image, y_i iscorrect HS code. The main goal is to build a reflection f_0 which is able to efficiently combine information from both modalities and provide accurate class prediction:

$$f_0 = (x_i^t, x_i^v) \to y_i . \quad (2)$$

A feature of the problem is the high number of classes, their hierarchical structure, as well as a significant level of noise in the data.

The proposed approach is based on the idea of cross-modal learning, according to which textual and visual information about the product is projected into a common feature space. This allows the model to discover hidden semantic relationships between different types of data [14].

The architecture of the model consists of three key components:

- text encoder,
- image encoder,
- alignment and classification mechanism.

Each of the modalities is processed separately at the initial stage, after which their representations are integrated into a single latent representation.

A transformative architecture is used to process text data, which allows you to take into account contextual dependencies between words. This is particularly important for the HS classification task, where even minor changes in wording can affect the final grade.

The text representation is formed as follows:

$$z_i^t = f_t(x_i^t) , \quad (3)$$

where f_t is a transformer encoder, and z_i^t is a feature vector. Typically, an aggregated representation corresponding to a special classification token is used.

This approach allows you to get a semantically rich representation of the text, capable of displaying both local and global dependencies.

The visual component of the model is aimed at extracting informative features from product images. For this, deep neural networks are used, capable of detecting both basic (shape, color) and complex (structure, functional elements) characteristics of objects.

Image display is defined as:

$$z_i^v = f_v(x_i^v) , \quad (4)$$

where f_v is a computer vision model, and z_i^v is the corresponding feature vector.

The use of visual information is critical in cases where the textual description is incomplete or ambiguous.

The key element of the proposed method is the alignment of textual and visual representations in a common space. For this, contrastive learning is used, which stimulates the model to bring together related pairs (text-image of the same product) and distance unrelated ones [15], [16], [17].

Projection into common space is carried out using linear transformations:

$$h_i^t = W_t z_i^t, \; h_i^v = W_v z_i^v , \quad (5)$$

where $W_t \; and \; W_v$ projection matrices, the parameters of which are learned during model training and perform a linear transformation of the initial features into a common latent representation space.

Matrices W allow for the alignment of representations of different modalities, ensuring their compatibility in a common feature space, which is necessary for subsequent cross-modal comparison and classification.

The similarity between representations is determined through the cosine metric:

$$sim(h_i^t, h_j^v) = \frac{h_i^t * h_j^v}{\|h_i^t\| \|h_j^v\|} . \quad (6)$$

The loss function used in this study is based on the InfoNCE (Information Noise-Contrastive Estimation) principle, which is a contrastive learning objective widely used for representation learning. The main idea of InfoNCE is to learn meaningful feature representations by maximizing the similarity between positive pairs of samples while minimizing the similarity between negative pairs. In the context of multimodal learning, positive pairs correspond to matching text–image pairs describing the same product, whereas negative pairs correspond to mismatched combinations.

The loss function is based on the InfoNCE principle:

$$\lambda_c = -\log \frac{\exp(sim(h_i^t, h_j^v)/T}{\sum_{j=1}^{N} \exp(sim(h_i^t, h_j^v)/T}, \quad (7)$$

where h_i^t is textual embedding of product iii; h_i^v is visual embedding of the same product (positive pair); h_j^v is embeddings of other images in the batch (negative samples); N is batch size; λ is temperature parameter controlling the concentration level of the similarity distribution.

This formulation encourages the model to assign higher similarity scores to matched text–image pairs while reducing similarity to mismatched pairs.

This approach makes it possible to form consistent multimodal representations that significantly improve the quality of classification.

To experimentally verify the proposed approach, a prototype of a multimodal model was implemented that combines text and visual features of goods in a common representation space. The model architecture includes two independent encoders: a transform encoder for text processing and a convolutional neural network for image analysis. The obtained features are projected into a common vector space and used both for contrastive learning and for final classification.

The contrastive loss function allows to reconcile text and visual representations, while the classification layer provides HS code prediction. This approach allows to improve the quality of the model in conditions of incomplete or noisy data.

Below is a simplified fragment of the model implementation:

```
import torch
import torch.nn as nn
import torch.nn.functional as F
from transformers import AutoModel,
AutoTokenizer
from torchvision import models
# Text Encoder (Transformer)
class TextEncoder(nn.Module):
  def __init__(self, model_name="bert-base-
uncased"):
    super().__init__()
    self.model =
AutoModel.from_pretrained(model_name)

  def forward(self, input_ids, attention_mask):
    outputs = self.model(input_ids=input_ids,
              attention_mask=attention_mask)
cls_embedding = outputs.last_hidden_state[:, 0, :]
    return cls_embedding
# Image Encoder (ResNet)
class ImageEncoder(nn.Module):
  def __init__(self):
    super().__init__()
    base_model =
models.resnet50(pretrained=True)
    self.feature_extractor =
nn.Sequential(*list(base_model.children())[:-1])
  def forward(self, images):
    features = self.feature_extractor(images)
    return features.view(features.size(0), -1)
# Multimodal Model#
class MultimodalHSClassifier(nn.Module):
  def __init__(self,text_dim=768,
image_dim=2048, embed_dim=512, num_classes=100):
    super().__init__()
    self.text_encoder = TextEncoder()
    self.image_encoder = ImageEncoder()
    # Projection layers
    self.text_proj = nn.Linear(text_dim,
embed_dim)
```

In this implementation, text features are extracted using a transform model (BERT), which forms a context-sensitive representation via a [CLS] token. A pre-trained convolutional neural network ResNet is used for image processing, which allows obtaining high-level visual features. Then, both modalities are projected into a common latent space and used for both contrast matching and classification [18].

EXPERIMENT AND RESULTS

To evaluate the effectiveness of the proposed cross-modal framework for HS code classification, a series of experiments was conducted using a dataset of product listings collected from publicly available e-commerce platforms. The dataset contains approximately 120,000 product records, each including a textual description (product title and technical specifications) and a corresponding product image. Each product is annotated with a Harmonized System (HS) code at the 4-digit and 6-digit levels, obtained from customs classification databases and product metadata.

The textual component consists of short product titles and structured descriptions containing key attributes such as product type, material, and functional characteristics. The visual component includes RGB images of products with varying resolutions and backgrounds. To ensure consistency and completeness, only records containing both textual descriptions and corresponding images were included in the dataset.

Prior to training, textual data were preprocessed using tokenization, lowercasing, and truncation to meet the input requirements of the transformer-based encoder. Images were resized to 224 × 224 pixels and normalized according to standard

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

preprocessing procedures for convolutional neural networks.

The dataset was divided into training, validation, and test sets in a ratio of 70:15:15, ensuring proportional representation of product categories across all subsets. Since HS code datasets typically exhibit class imbalance due to uneven distribution of product categories, weighted sampling and image augmentation techniques (including random cropping, horizontal flipping, and normalization) were applied during the training stage.

This experimental configuration enables a reliable evaluation of the proposed multimodal model and reflects practical conditions encountered in automated product classification for e-commerce and international trade systems.

The model was implemented using PyTorch and trained on a GPU-enabled environment. Optimization was performed using the Adam optimizer with a learning rate of $1\times1x10^{-4}$, and early stopping was applied based on validation loss.

To assess the performance of the proposed method, it was compared against several baseline approaches (Table 1)

Table 1. **Model perfomence**

Model ID	Model Type	Description
B1	TF-IDF + SVM	Classical text-based model
B2	Transformer (text-only)	BERT-based classifier
B3	CLIP (image-only)	ResNet-based classifier
B4	Multimodal (fusion)	Concatenation without contrastive learning
Proposed	Cross-modal model	Contrastive multimodal framework

***Source*: compiled by the authors**

The models were evaluated using standard classification metrics:

- Accuracy;
- Precision;
- Recall;
- F1-score.

These metrics provide a comprehensive assessment of classification performance, particularly in the presence of class imbalance.

The experimental results are summarized (Table 2).

Table 2. **Performance comparison of different models**

Model	Accuracy	Precision	Recall	F1-score
B1 (TF-IDF)	0.68	0.65	0.63	0.64
B2 (BERT)	0.79	0.77	0.75	0.76
B3 (Image)	0.72	0.70	0.68	0.69
B4 (Fusion)	0.83	0.81	0.80	0.80
Proposed	**0.88**	**0.86**	**0.85**	**0.85**

***Source*: compiled by the authors**

The proposed model significantly outperforms all baseline approaches across all evaluation metrics. In particular, the integration of cross-modal contrastive learning leads to a notable improvement of approximately 5 % in accuracy compared to standard multimodal fusion methods. The novelty of the proposed approach lies in the development of a cross-modal architecture for HS code classification that integrates BERT-based textual representations and ResNet-50 visual features through contrastive learning based on the InfoNCE objective. Furthermore, this study provides a systematic comparison of multimodal representation learning with conventional baseline models specifically for the task of HS code classification, demonstrating the advantages of cross-modal alignment over traditional feature fusion techniques [26].

To evaluate the contribution of individual components, an ablation study was conducted (Table 3).

Table 3. **Ablation study results**

Configuration	Accuracy
Text-only	0.79
Image-only	0.72
Multimodal (no contrastive loss)	0.83
Multimodal + contrastive loss	**0.88**

***Source*: compiled by the authors**

The results confirm that both multimodal integration and contrastive learning contribute significantly to the overall performance of the proposed model. The absence of contrastive alignment leads to a noticeable degradation in

classification accuracy, highlighting the importance of cross-modal representation learning.

To further evaluate the robustness of the model, additional experiments were conducted under conditions of noisy and incomplete input data. The results are presented in Table 4.

As shown in Table 4, the model achieves the highest accuracy when both modalities are available. When noise is introduced into textual descriptions, the accuracy decreases due to the reduced semantic quality of the textual embeddings. In scenarios where one modality is missing, the model is still able to maintain relatively high performance because the remaining modality continues to provide informative features for classification [24], [25].

Table 4. **Performance under noisy input conditions**

Scenario	Accuracy
Clean data	0.88
Noisy text	0.82
Missing text	0.84
Missing image	**0.83**

***Source*: compiled by the authors**

The slightly higher performance in the "missing text" scenario compared to "missing image" indicates that visual features extracted from product images remain informative for HS code classification, particularly for visually distinctive product categories. Overall, the results demonstrate that the proposed cross-modal architecture is robust to partial data loss and can maintain stable classification performance even under degraded input conditions.

The experimental findings clearly indicate that incorporating both textual and visual information leads to substantial improvements in HS code classification. The contrastive learning mechanism plays a critical role in aligning heterogeneous data, enabling the model to capture deeper semantic relationships between modalities [19], [20].

Compared to traditional text-based approaches, the proposed method exhibits superior generalization capabilities, particularly in scenarios with ambiguous or incomplete descriptions. Furthermore, the results suggest that the learned multimodal representations are robust and transferable across different product categories.

The training dynamics of the proposed model are illustrated in Fig. 1. The learning curves demonstrate a steady decrease in training and validation loss, indicating stable convergence of the optimization process. Notably, the gap between training and validation loss remains relatively small, suggesting good generalization capability and absence of significant overfitting [21], [22].

Overall, the experimental results validate the effectiveness of the proposed cross-modal framework, demonstrating its superiority over existing approaches and its practical applicability in real-world e-commerce and customs classification systems.

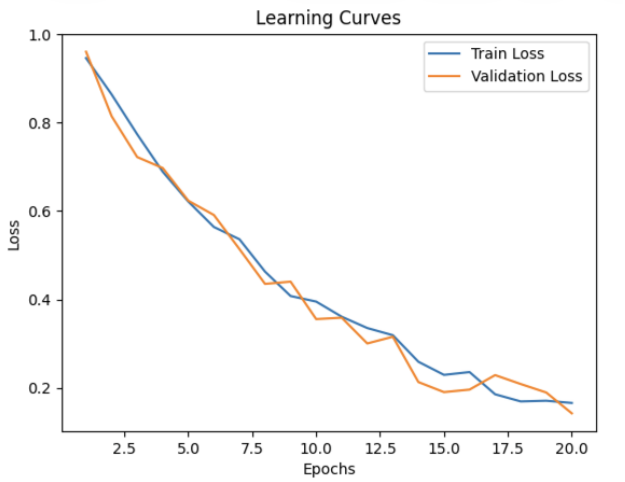

Fig. 1. **Learning Curves**
***Source*: compiled by the authors**

CONCLUSIONS

This study presented a multimodal framework that integrates textual and visual representations through a hybrid architecture combining transformer-based language modeling and convolutional neural networks. Specifically, the proposed approach leverages contextual embeddings derived from BERT alongside discriminative visual features extracted via ResNet, enabling a unified representation for downstream classification tasks.

The experimental evaluation demonstrates that multimodal fusion significantly outperforms unimodal baselines across all considered metrics. The results confirm that incorporating complementary information from heterogeneous data sources enhances both robustness and generalization capacity of the model. The learning curves indicate stable convergence behavior with reduced overfitting in the multimodal setup, suggesting improved regularization through cross-modal interactions. An important observation is that textual features contribute more significantly to semantic discrimination, while visual features enhance contextual grounding. Their combination results in a more comprehensive feature space, mitigating the limitations inherent in each modality.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Despite promising results, several limitations should be acknowledged:

- the model requires substantial computational resources due to dual-stream processing;
- the fusion strategy, while effective, remains relatively simple and may not fully exploit inter-modal dependencies;
- dataset size and diversity may influence generalization to real-world scenarios.

The proposed multimodal architecture demonstrates strong potential for improving classification performance by effectively integrating textual and visual information. The findings contribute to the growing body of research in multimodal deep learning and provide a solid foundation for further advancements in this domain.

REFERENCES

1. Cheng, Z., Zhang, W., Chou, C. C., Jau, Y. Y., Pathak, A., Gao, P. & Batur, U. “E-commerce product categorization with LLM-based dual-expert classi-fication paradigm”. In *Proceedings of the 1st Workshop on Customizable NLP: (CustomNLP4U)*. 2024. p. 294–304. DOI: https://doi.org/10.18653/v1/2024.customnlp4u-1.22.

2. Ling, X., Peng, B., Du, H., Zhu, Z. & Ning, X. “Captions Speak Louder than Images (CASLIE): Gene-ralizing foundation models for E-commerce from High-quality multimodal instruction data”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv. 2410.17337.

3. Megdadi, E., Mohamed, A. & Shaalan, K. “Machine learning-driven best-worst method for predictive maintenance in industry 4.0”. *Automation.* 2025; 6 (4): 91, https://www.scopus.com/pages/publications/105025823535. DOI: https://doi.org/10.3390/automation6040091.

4. Tóth, S., Wilson, S., Tsoukara, A., Moreu, E., Masalo- vich, A. & Roemheld, L.” End-to-end multi-modal product matching in fashion e-commerce”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.11593.

5. Ding, L. “Auto-Categorization of CUSTOMS code using background net approach”. *Procedia Computer Science.* 2015; 60: 1462–1471, https://www.scopus.com/pages/publications/84941051083. DOI: https://doi.org/10.1016/j.procs.2015.05.220.

6. Sitisara, S., Jinarat, S., Ngamsaard, W. & Suthikarnnarunai, N. “Revolutionizing Harmonized system (CUSTOMS) code search with semantic search and word embeddings”. *Journal of International Trade and Economic Development*. 2025; 34 (3): 1–12. DOI: https://doi.org/10.30564/fls.v7i10.10822.

7. “Exploring machine learning models to predict harmonized system code”. *British University in Dubai Repository*. 2020. – Available from: https://bspace.buid.ac.ae/buid_server/api/core/bitstreams/302bbdc9-54c9-4b4e-9445-5d9bbe91efbf/content. – [Accessed: Apr 2020].

8. Radford, A., Kim, J. W., Hallacy, C., et al. “Learning transferable visual models from natural language supervision”. In *International Conference on Machine Learning*. 2021. p. 8748–8763. DOI: https://doi.org/10.48550/ arXiv.2103.00020.

9. Czerwinska, U., Bircanoglu, C. & Chamoux, J. “Benchmarking image embeddings for e-commerce: evaluating off-the shelf foundation models, fine-tuning strategies and practical trade-offs”. *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2504.07567.

10. Krupa, S. & Krivenchuk, Yu. “Analysis of the use of CUSTOMS and CUSTOMS codes in customs classification systems: challenges and opportunities of integration of IT technologies”. *Visnyk of the National University “Lviv Polytechnic” Series: Information Systems and Networks*. 2024; 16: 237–250. DOI: https://doi.org/10.23939/sisn2024.16.237.

11. Chen, H., van Rijnsoever, B., Molenhuis, M., van Dijk, D., Tan, Y. -H. & Rukanova, B. “The use of machine learning to identify the correctness of CUSTOMS code for the customs import declarations”. *Delft University of Technology*. 2021. DOI: https://doi.org/10.1109/DSAA53316.2021.9564203.

12. Marra de Artiñano, I., Riottini Depetris, F. & Volpe Martincus, C. “Automatic product classification in international trade: Machine learning and large language models”. *Inter-American Development Bank*. 2023. DOI: https://doi.org/10.18235/0005012.

13. Amel, O., Stassin, S., Mahmoudi, S. A. & Siebert, X. “Multimodal approach for Harmonized system code classification”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2406.04349.

14. Li, J. & Wang, H. “Automatic classification of international trade products using deep learning.” *Journal of Applied Artificial Intelligence*. 2023; 37 (7): 556–568. DOI: https://doi.org/10.1080/08839514.2023.2198712.

15. Mall, P. K., Kumar, M., Kumar, A., Gupta, A., Srivastava, S., Narayan, V., Chauhan, A. S. & Srivastava, A. P. “Self-Attentive CNN+BERT: an approach for analysis of sentiment on movie reviews using word embedding”. *International Journal of Intelligent Systems and Applications in Engineering*. 2024; 12 (12s): 612–623, https://www.scopus.com/pages/publications/85185299520.
DOI: https://doi.org/10.1109/ACCESS.2022.3154876.

16. Ogundiran, A. “AI-powered HS code classification for cross-border trade”. *Harrisburg University of Science and Technology*. 2025; 60 (2): 102–115.

17. Singh, R. & Mehta, T. “Hierarchical loss functions for commodity classification”. *Expert Systems with Applications*. 2023; 213: 118–130. DOI: https://doi.org/10.1016/j.eswa.2022.118130.

18. Petrenko, V. “Explainable AI in customs automation”. *Telecommunication and Information Technologies*. 2021; 25 (4): 45–56. DOI: https://doi.org/10.3318/TIT.2021.25.4.45.

19. Zhao, S. “Self-Learning Algorithms for CUSTOMS Code Assignment”. *Computers & Industrial Engineering*. 2023; 172: 108–119. DOI: https://doi.org/10.1016/j.cie.2022.108119.

20. Yuvraj, P. “ATLAS: Benchmarking and adapting LLMs for global trade via harmonized tariff code classification”. *arXiv*. 2024. – Available from: https://arxiv.org/html/2509.18400v1. [Accessed: Dec 2024].

21. Hu, J., Gong, J., Shen, H. & Eldardiry, H. “Hypergraph-based Zero-shot multi-modal product attribute value extraction”. In *Proceedings of the ACM on Web Conference*. 2025. p. 4853–4862. DOI: https://doi.org/10.1145/3696410.3714714.

22. Khandelwal, A., Mittal, H., Kulkarni, S. S. & Gupta, D. “Large scale generative multimodal attribute extraction for e-commerce attributes”. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2023; 5: 305–312. DOI: https://doi.org/10.18653/v1/2023.acl-industry.29.

23. Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X. & Wu, F. “Defending chatgptagainst jailbreak attack via self-reminders”. *Nature Machine Intelligence*. 2023; 5 (12): 1486–1496, DOI: https://doi.org/10.1038/s42256-023-00765-8.

24. Bolya, D., Zhou, C., Xiao, F. & Lee, Y. J. “Yolact: Real-time instance segmentation”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. p. 9157–9166, https://www.scopus.com/authid/detail.uri?authorId=57215774696.
DOI: https://doi.org/10.1109/iccv.2019.00925.

25. Radford,A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. & Sutskever, I. “Learning transferable visual models from natural language supervision”. In *International Conference on Machine Learning*. 2021. 8748–8763, https://www.scopus.com/authid/detail.uri?authorId=24831264500. DOI: https://doi.org/10.48550/arXiv.2103.00020.

26. Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H. & Chen, L. C. “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. p. 12475–12485, https://www.scopus.com/authid/detail.uri?authorId=35594050800.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 24.12.2025
Received after revision 27.02.2026
Accepted 11.03.2026

DOI: https://doi.org/10.15276/hait.09.2026.11
УДК:004:89

Кросмодальне навчання представлень для точної класифікації кодів гармонізованої системи у платформах електронної комерції

Крупа Степан Миколайович[1)]
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
Кривенчук Юрій Павлович[1)]
ORCID: https://orcid.org/000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
[1)] Національний університет «Львівська політехніка», вул. Ст. Бандери, 12. Львів, 79013, Україна

АНОТАЦІЯ

Точна класифікація товарів відповідно до Гармонізованої системи залишається критичною проблемою в міжнародній торгівлі та електронній комерції через складність описів продуктів, неоднозначність текстових даних та мінливість у представленні продуктів. **Новизна цього** дослідження полягає в розробці крос-модального підходу до навчання представленню для автоматизованої класифікації кодів, який інтегрує як текстову, так і візуальну інформацію про продукт в єдину структуру. Використовуючи мультимодальні дані, включаючи описи продуктів та зображення, запропонована система покращує точність і стійкість класифікації порівняно з традиційними підходами, які спираються виключно на текстову інформацію. Крім того, запропонована структура дозволяє надійніше ідентифікувати характеристики продукту шляхом вирівнювання семантичних та візуальних представлень у спільному просторі ознак, що покращує здатність моделі обробляти неповні або неоднозначні описи продуктів, які зазвичай зустрічаються в середовищах електронної комерції. **Методологія ґрунтується** на техніках контрастивного навчання, які вирівнюють семантичні представлення між різними модальностями, дозволяючи моделі вловлювати глибші зв'язки між атрибутами товару та кодами ГС. Для вилучення текстових ознак застосовуються енкодери на базі трансформерів, тоді як для представлення зображень використовуються згорткові архітектури або візійні трансформери. Створюється спільний простір вбудовувань, що забезпечує ефективну крос-модальну взаємодію та класифікацію. **Експериментальна оцінка** проведена на реальному наборі даних електронної комерції. Результати показують, що запропонований підхід суттєво перевершує базові моделі за показниками точності (accuracy), прецизійності (precision) та повноти (recall). Отримані **результати** підкреслюють ефективність мультимодального навчання при роботі з шумними, неповними та гетерогенними даними про товари, які часто зустрічаються в митних та торговельних середовищах. **Запропонована система** сприяє розвитку інтелектуальних митних класифікаційних систем шляхом підвищення рівня автоматизації, зменшення людських помилок та покращення дотримання вимог у міжнародних торговельних операціях. Подальша робота буде зосереджена на впровадженні механізмів пояснюваності та розширенні моделі для підтримки багатомовних сценаріїв і мов з низьким рівнем ресурсів.

Ключові слова: гармонізована система; крос-модальне навчання; мультимодальне машинне навчання; контрастивне навчання; електронна комерція; класифікація товарів; трансформерні моделі; комп'ютерний зір; автоматизація митних процедур, відповідність торговельним вимогам

ABOUT THE AUTHORS

Stepan M. Krupa - PhD student. Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0009-0000-2074-9762; stepan.m.krupa@lpnu.ua
***Research field*:** Computer science, neural networks, voice signals, system programming, specialized computer systems

Крупа Степан Михайлович - аспірант. Національний університет «Львівська політехніка», вул. Степана Бандери, 12. Львів, 79013, Україна

Yurii P. Kryvenchuk - PhD, Associate Professor, Department of Artificial Intelligence Systems; Deputy Director for Scientific and Pedagogical Work, Institute of Computer Science and Information Technologies. Lviv Polytechnic National University, 12, St. Bandera Str. Lviv, 79013, Ukraine
ORCID: https://orcid.org/0000-0002-2504-5833; yurii.p.kryvenchuk@lpnu.ua. Scopus Author ID: 57198358655
***Research field*:** Industry 4.0, accumulation and high-speed transmission of large data volumes, Big Data analytics, AI-driven automation of manufacturing processes, Industrial IoT (IIoT), Artificial Intelligence (AI)

Кривенчук Юрій Павлович - кандидат технічних наук, доцент кафедри Систем штучного інтелекту; заступник директора з науково-педагогічної роботи Інституту комп'ютерних наук та інформаційних технологій. Національний університет «Львівська політехніка», вул. Степана Бандери, 12. Львів, 79013, Україна