Multilingual Semantic Sourcing using Product Images for Cross-lingual Alignment

Sourab Mangrulkar*
smangrul@amazon.com
Amazon
Bengaluru, Karnataka, India

Ankith M S*
ankiths@amazon.com
Amazon
Bengaluru, Karnataka, India

Vivek Sembium
viveksem@amazon.com
Amazon
Bengaluru, Karnataka, India

ABSTRACT

In online retail stores with ever-increasing catalog, product search is the primary means for customers to discover products of their interest. Surfacing irrelevant products can lead to poor customer experience and in extreme situations loss in engagement. With the recent advances in NLP, Deep Learning models are being used to represent queries and products in shared semantic space to enable semantic sourcing. These models require a lot of human annotated (query, product, relevance) tuples to give competitive results which is expensive to generate. The problem becomes more prominent in the emerging marketplaces/languages due to data paucity problem. When expanding to new marketplaces, it becomes imperative to support regional languages to reach a wider customer base and delighting them with good customer experience. Recently, in the NLP domain, approaches using parallel data corpus for training multilingual models have become prominent, but they are expensive to generate. In this work, we learn semantic alignment across languages using product images as an anchor between them. This overcomes the necessity of parallel data corpus. We use the human annotated data from established marketplace to transfer relevance classification knowledge to new/emerging marketplaces to solve the data paucity problem. Our experiments performed on datasets from Amazon reveal that we outperform state-of-the-art baselines with **2.4%-3.65%** ROC-AUC lifts on relevance classification task across non-English marketplaces, **34.69%-51.67%** Recall@k lifts on language-agnostic retrieval task and **6.25%-13.42%** Precision@k lifts on semantic neighborhood quality task, respectively. Our models demonstrate efficient transfer of relevance classification knowledge from data rich marketplaces to new marketplaces by achieving ROC-AUC lifts of **3.74%-6.25%** for the relevance classification task in the zero-shot setting where the human annotated relevance data of target marketplace is unavailable during training.

CCS CONCEPTS

• **Information systems** → **Multilingual and cross-lingual retrieval**; **Multimedia and multimodal retrieval**.

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

KEYWORDS

Deep Learning; Multilingual; Multimodal; Semantic Sourcing; E-Commerce

ACM Reference Format:
Sourab Mangrulkar, Ankith M S, and Vivek Sembium. 2022. Multilingual Semantic Sourcing using Product Images for Cross-lingual Alignment. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

E-commerce companies like Amazon, Walmart, Alibaba, eBay, etc. are gaining popularity across the globe and are becoming an integral part of lifestyles. As they expand to emerging and new marketplaces, they are increasingly being used by customers speaking different languages and dialects. To provide a smoother customer experience and to decrease customers' cognitive load, a good sourcing engine should parse queries expressed in any language as well as ill-formed queries to retrieve relevant products. For example, Amazon.in (https://amazon.in/) allows customers to shop in English, Hindi, Marathi, Telugu, Kannada, Tamil, Malayalam and Bengali.

A typical e-commerce product search involves two major components. (1) Sourcing system: Retrieving products that are relevant to the customer's intent (expressed as a query) and (2) Ranking system: prioritising the retrieved list of products in decreasing order of relevance for the customer. The focus of this work is the first component, i.e, sourcing relevant products for a given query. Traditionally, sourcing systems relied on syntactic matches between the query terms and the terms in the product title/description. This approach increases the cognitive load of the customer since they have to re-formulate queries many times to discover relevant products.

In order to incorporate semantics into these systems, sourcing systems can leverage historical anonymized user session logs containing queries, clicks and purchase data. For example, if many customers search for the term "payal" but ultimately in the same session end up purchasing "anklets", then the term "payal" can be added to the semantic index of that product. This method of enhancing the index, although useful, has a few shortcomings: (1) It does not generalise well due to a lack of understanding of the deep semantics between the query and the products; for example, products containing the term "anklet" but undiscovered by customers using the search query "payal" will not have "payal" in their semantic index; (2) It can lead to noisy matches between the query and the products, e.g., a customer searching for the "Colgate toothbrush" but purchasing "Colgate toothpaste" in that session can lead to the former being in the index of the latter. To alleviate this noise, large amounts of human-annotated relevance data are required,

which is expensive to generate. This problem is further amplified for emerging/new marketplaces due to the paucity of relevance audit data. Maintaining marketplace-specific models incurs computational and operational overhead. It also restricts multi-locale search, e.g., customers searching in Tamil in the English-predominant IN marketplace. Multilingual models provide a viable option to handle all these limitations.

Recently, there have been significant advances in language processing technologies with the introduction of transformer models such as [7] and these models have been extended to learn unified multilingual models that can be used to process multiple languages. Our approach to Semantic Sourcing is to independently represent queries and products into an n-dimensional semantic space such that the entities (query/product) that are relevant to each other are in the neighborhood of each other in the semantic space. Applying standard multilingual language models [7] to this approach by combining all language data does not enable the transfer of deep semantic knowledge between marketplaces and languages. Recent approaches for training multilingual semantic sourcing models such as [11] and [1] require human-annotated parallel-corpus and large datasets of cross-listed products for every language pair, respectively. [23] leverages customer behaviour data from various marketplaces to learn multilingual representation. All these datasets are either expensive to generate or limited/unavailable for emerging/new marketplaces. However, there is a large corpus of product catalog for each marketplace consisting of (product title, description, image) tuples. In this paper, we propose **M**ultilingual **M**ultimodal **S**emantic **S**ourcing model (**M2S2**) for cross-language product retrieval wherein product images act as an anchor between languages to learn multilingual semantic alignment.

The main contributions of our work are multi-fold. (1) **Multilingual alignment using visual modality**: To the best of our knowledge, we are the first to leverage product image which is language-agnostic in nature to learn alignment between languages for e-commerce product retrieval. (2) **Transfer Learning**: The Query-Product relevance knowledge is transferred from high resource marketplaces to low resource marketplaces. (3) **Model**: We enhance the Siamese architecture network to learn language agnostic semantic representations. (4) **Zero-Shot learning**: By leveraging product images for multilingual semantic alignment, we reduce the necessity for human-annotated relevance data and thereby improve the zero-shot relevance classification performance across marketplaces. We perform ablation studies to evaluate the effectiveness of major components of the proposed M2S2 framework.

2 RELATED WORK

Product Search: Traditionally, syntactic matching between query and document has been done to retrieve relevant documents for a given query. With the introduction of latent semantic models, the intent matching happens at the semantic level where syntactic matching often fails. With recent advances in NLP, deep learning models are being employed for semantic sourcing. These include embedding based models such as Deep Semantic Search Model (DSSM) [12, 25], Convolutional Deep Semantic Search Model (CDSSM) [33], Multi-Task DNN [20], Sentence-BERT [30] and ColBERT [13]. DSSM and CDSSM models are latent semantic models with a deep structure that project queries and documents into a common low-dimensional space where the relevance of a document given a query is computed as the cosine similarity between their embeddings.

BERT [7] which is based on encoder part of Transformer [39] has achieved state-of-the-art performance in various NLP applications [7]. Sentence-BERT (S-BERT) [30] is a modification of BERT based on Siamese network which can be used for tasks with higher computational requirement where BERT is infeasible. ColBERT [13] improves the performance over S-BERT by performing delayed-interaction on query and document token embeddings. RoBERTa [22] leverage various pre-training tricks and achieves significant improvements over BERT. There has been an explosion in BERT based models exploring architectural innovation, improvements in training methodologies and leveraging more data along with entities from knowledge base [4, 10, 17, 29, 31, 44]. All these models can be used in Siamese fashion to get embedding based models for semantic sourcing.

Multilingual Search: M-BERT [7] is the multilingual variant of BERT which is trained on 104 languages and has demonstrated competitive results on Multilingual tasks [27]. XLM [15] leverages bilingual parallel-corpora and a combination of masked language modelling (MLM) and translation language modelling (TLM) for cross-lingual representations. XLM-RoBERTa [5] combines XLM approach with RoBERTa tricks to pre-train a model on more than 100 languages on MLM task. Other multilingual BERT based variants include [21, 41]. [42] introduces 2 cross-lingual retrieval oriented pre-training tasks, namely, query language modelling (QLM) and relevance ranking (RR) to enhance generalizability of multilingual BERT based models on cross-lingual information retrieval (CLIR) tasks.

Deep learning based model for semantic similarity between a pair of sentences from different languages has been proposed in [2]. [32] leverages bilingual corpora and represents query & document in shared semantic space for enabling CLIR but doesn't explicitly align queries/documents in language-agnostic semantic space. In the e-commerce domain, multilingual models based on Transformer architecture are gaining prominence. Query translation based approach is presented in [11] which requires parallel-corpus for each language pair. Model proposed in [1] uses cross-listed products across each language pair to learn multilingual alignment. [23] proposes graph based multilingual model leveraging customer behaviour data from all marketplaces. However, the data used for these approaches is limited/unavailable for emerging/new marketplace. We overcome these limitations by using abundant catalog data consisting of product and images for multilingual alignment.

Multimodal Search: With widespread adoption of Transformer based models, they are being actively explored in conjunction with Convolutional Neural Networks (CNNs) for jointly learning visual and textual representations [3, 18, 19, 24, 35, 37]. VirTex [6] approach use Image Captioning task to jointly train CNN and Transformer. Approaches outlined in ConVIRT [43] and CLIP [28] use bi-directional contrastive loss between the visual and textual modalities for representation learning. All these models leverage abundantly available images along with the related raw text instead of restricting to labelled data consisting of specific number of predefined classes. This enables wider adoption and generalizability.

However, all these models focus on leveraging raw text to learn better visual representation and are mono-lingual.

Multilingual Search using Visual Modality: Multilingual models using image-text data is still an active research area. [34] approach uses 3 Billion (image, caption, relevance) tuples from different marketplaces for learning multilingual embeddings. Globetrotter model proposed in [36] uses visual modality to learn multilingual alignment. This approach serves as an inspiration upon which we build. Due to its unsupervised setting, it uses contrastive learning for training image model. We instead used abundantly available customer behaviour data from established marketplaces to pre-train image model in a supervised setting.

3 OUR APPROACH

In this section, we describe our proposed M2S2 framework- a Deep Learning based model that learns to map query and product into a language-agnostic space using image as an anchor for cross-lingual retrieval. This consists of two challenges: (1) learning multilingual alignment. (2) Learning query-product relevance classification.

Product images are language-agnostic in nature, e.g., from Figure 1 we can observe that images are similar but their product titles are in different languages. Hence, we can use image as an implicit signal for cross-lingual alignment. This does not require any bilingual data or common products across the marketplaces which is expensive to generate. Inspired by the Globetrotter framework [36], we learn an aligned embedding space without bilingual corpora by mapping similar products to language-agnostic semantic space using product image as an anchor. Globetrotter framework uses contrastive loss for learning visual alignment between images in absence of explicit supervision. Instead, we exploit customer behaviour data which is abundantly available across marketplaces to generate (Image1, Image2, label) tuples for learning visual alignment in supervised setting. We demonstrate that it will improve the performance in comparison to contrastive learning, which is empirically found to align dissimilar products with similar images closer and similar products with dissimilar images further away. We then incorporate the pre-trained image model in globetrotter framework to learn multilingual semantic alignment.

Our proposed model M2S2 differs from globetrotter model in several aspects. Globetrotter framework focuses on machine translation by modelling it as retrieval based task. This is completely different from our problem statement which is to source relevant products for a given query across marketplaces/languages/locales. Globetrotter framework which only leverages catalog data comprising (image, product) pairs performs poorly on end task of relevance classification as it doesn't leverage any query information which is part of relevance data. Hence, globetrotter framework helps only in addressing challenge 1 (learning multilingual alignment). We leverage supervised data using anonymized and aggregated customer behaviour data from established marketplaces along with taxonomy based hard negatives to pre-train the image encoder for robust visual alignment capability. As visual alignment is critical component for multilingual alignment, this enhances language-agnostic retrieval performance considerably (Section 4.4). We experiment with Vision Transformer (ViT) [8, 28] which has shown state-of-the-art results on computer vision tasks in addition to ResNet-50 [9]. We also experiment with XLM-RoBERTa (XLM-R) [5] which has demonstrated better results when compared to M-BERT [27]. Experimental results demonstrate that ViT as image encoder and XLM-RoBERTa as text encoder achieves the best performance on all tasks (Section 4.4). Details of pre-training image model and cross-lingual alignment are described in Section 3.3 and 3.4, respectively.

In order to solve query-product relevance classification (challenge 2), we leverage human-annotated relevance data from different marketplaces, and we describe our approach in Section 3.5. To further improve the neighborhood quality in semantic space, we leverage catalog taxonomy data, and we describe this in Section 3.5.

3.1 Problem Formulation and Notation

We formulate the cross-lingual retrieval problem as follows. Let $L = \{l_1, l_2 ....l_z\}$ be the set of languages, $P = \{p_1^{g_1}, p_2^{g_2}, ..., p_n^{g_n}\}$ be the set of products and $Q = \{q_1^{h_1}, q_2^{h_2}, ..., q_m^{h_m}\}$ be the set of queries where $g_i \in L$ and $h_i \in L$. Let $D_{cls} = \{(q_1^{g_1}, p_1^{g_1}, y_1), ...(q_z^{g_z}, p_z^{g_z}, y_z)\}$ be human-annotated relevance data where y_i is the relevance label. Let $D_{align} = \{(p_1^{g_1}, image_1), ...(p_n^{g_n}, image_n)\}$ be the data used for semantic alignment across languages. Let $D_{cat} = \{(p_1^{g_1}, image_1, cn_1), ...(p_n^{g_1}, image_n, cn_n)\}$ and $D_{qq} = \{(q_1^{g_1}, cn_1), ...(q_n^{g_n}, cn_n)\}$ represent the product and query catalog data, respectively, where cn_i represents corresponding category in the product taxonomy. Based on the historical customer purchases, let D_{pp+} and D_{qq+} be product-product and query-query positive pairs, respectively. First, we filter $(q_i^{g_i}, p_i^{g_i}, c_i^{g_i})$ tuples where $c_i^{g_i} > \theta_{co-occur}^{threshold}$, where $c_i^{g_i}$ represents number of in-session purchases of $p_i^{g_i}$ for query $q_i^{g_i}$. Let this data be D_{filter} which can be represented as bipartite graph between Q and P with edges being weighted by $c_i^{g_i}$s. Next, we apply the topic co-occurrence methodology based on Normalized Point-wise Mutual Information (NPMI) scores from [16] to this bipartite graph to generateD_{pp+} and D_{qq+} datasets.

Let $B(x_i; \Theta_{txt})$ be the text encoder that maps $x_i \in Q \cup P$ to embedding space $LA \in \mathbb{R}^d$. Let $I(image_i; \Theta_{img})$ be the image encoder that maps $image_i$ to embedding space $S \in \mathbb{R}^d$. Given an arbitrary query $q_i^{h_i} \in Q$, our objective is to find K nearest products $P_i = \{p_1^{g_1}, p_2^{g_2} ...., p_k^{g_k}\}$ in LA.

3.2 Model Architecture:

Figure 2 illustrates our proposed M2S2 retrieval model optimized for the relevance classification task. Figure 1 provides an overview of our cross-lingual alignment model using images as an anchor. The details of these components are explained below.

1) Text Encoder : We use BERT based model in our text encoder to independently represent query and product in intermediate contextual embedding space. Text Encoder $B(x_i, \Theta_{txt})$, first computes fixed size contextual representation for an entity x_i by using the 'cls' token output of the BERT model. We pass this to prediction head comprising of hidden layer with non-linear activation followed by output layer. This provides embedding $e_i^{txt} \in \mathbb{R}^d$. We use the same text encoder for representing both query and product to enable the transfer of language semantics between them.

2) Image Encoder : Image encoder I leverages ResNet-50/Vision Transformer (ViT) in conjunction with prediction head (as described above) to generate image embedding $e_i^{img} \in \mathbb{R}^d$.

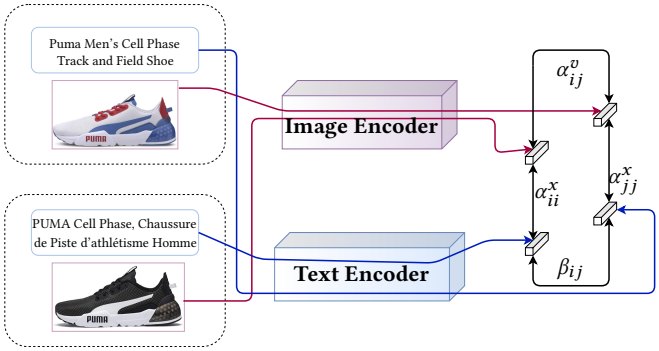

Figure 1: Semantic alignment across languages using product images as a bridge between them

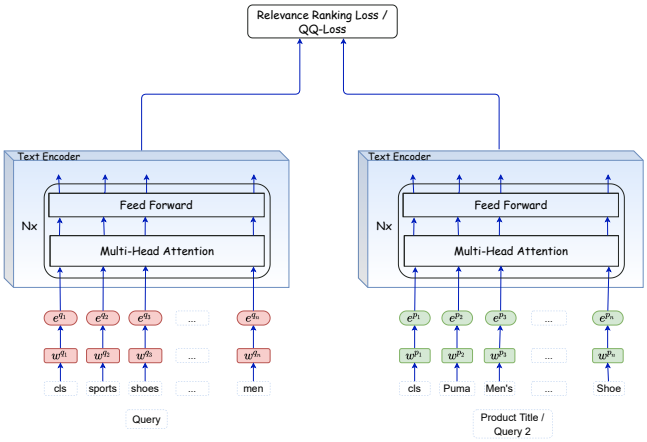

Figure 2: M2S2 Retrieval Model

3.3 Pre-Training Image Model

To pre-train our image model I, we leverage D_{pp+} and product to category mapping information available in product catalog $cn_i \in D_{cat}$ to generate hard negative samples D_{pp-}. These hard negative product pairs have images which are closer in embedding space but have significant cn_i (product category) mismatch. We pre-train our image model for several iterations; in each iteration, hard negatives are mined as explained above and loss ℓ_{img} is optimized (equation 1). θ_{pos} and θ_{neg} are hyper-parameters.

$$\ell_{img} = \sum_{(image_i^1, image_i^2) \in \mathcal{D}_{pp+}} \min(0, \hat{y}_i^{img} - \theta_{pos})^2 + \sum_{(image_i^1, image_i^2) \in \mathcal{D}_{pp-}} \max(0, \hat{y}_i^{img} - \theta_{neg})^2 \quad (1)$$

3.4 Multilingual Semantic Alignment using Images

Figure 1 illustrates multilingual semantic alignment model. We learn multilingual semantic alignment across languages by optimizing the following text-to-text contrastive learning problem where τ is hyper-parameter:

$$\ell_{xl} = -\sum_i \sum_{j \neq i} \alpha_{ij} \log \frac{\exp(\beta_{ij}/\tau)}{\sum_{k \neq i} \exp(\beta_{ik}/\tau)} \text{ where } \beta_{ij} = \text{cosine}(e_i^{txt}, e_j^{txt}) \quad (2)$$

We don't have parallel corpus data. In order to get the positive and negative pairs for contrastive learning, we use the scalar $\alpha_{ij} \in [0, 1]$ as soft labels. These scalars are generated using visual modality from product images. Here the core idea is to use transitive relations via visual modality, wherein sentences from different languages are semantically similar if they appear in similar visual contexts. Let α^x_{ii} be cross-modal similarity of product $p_i^{g_i}$ and corresponding image $image_i$. Let α^v_{ij} be the image-image similarity between images $image_i$ and $image_j$. The computation for α_{ij}:

$$\alpha_{ij} = f(\alpha^x_{ii} \cdot \alpha^v_{ij} \cdot \alpha^x_{jj}) \quad \text{where } f(x) = \frac{max(0, x - m)}{1 - m} \text{ and } m \text{ being margin} \quad (3)$$

We can observe α_{ij} is high only when cross-modal alignments are high between image-text pairs and visual-modal alignment is high between the image-image pair. Visual similarity and cross-modal similarities are learnt contrastively by optimizing following losses:

$$\ell_v = -\sum_{ij} \log \frac{\exp(\alpha^v_{ij}/\tau)}{\sum_{k \neq i} \exp(\alpha^v_{ik}/\tau)} \quad \text{where } \alpha^v_{ij} = \text{cosine}(e_i^{img}, e_j^{img}) \quad (4)$$

$$\ell_{xm} = -\sum_i \left(\log \frac{\exp(\alpha^x_{ii}/\tau)}{\sum_j \exp(\alpha^x_{ij}/\tau)} + \log \frac{\exp(\alpha^x_{ii}/\tau)}{\sum_j \exp(\alpha^x_{ji}/\tau)} \right) \quad (5)$$
$$\text{where } \alpha^x_{ij} = \text{cosine}(e_i^{txt}, e_j^{img})$$

Full objective is the combination of all above losses where λ are hyperparameters :

$$\ell_{align} = \lambda_{xl}\ell_{xl} + \lambda_v\ell_v + \lambda_{xm}\ell_{xm} \quad (6)$$

The image encoder I is a crucial component as it guides the model to align the product titles expressed in various languages in a language-agnostic space. The quality of the cross-lingual alignment will depend on the robustness of the image encoder. Globetrotter framework [36] learns visual similarity using contrastive learning as specified in equation 4. We instead use the pre-trained image model I from Section 3.3 which has already learnt visual alignment using supervision from customer behaviour data. We initialize the parameters of image model I while optimizing equation 6. Experimental results reveal that our image model outperforms the globetrotter image model [36] (trained only using the contrastive loss) in the cross-lingual retrieval task (Section 4.4).

3.5 Query-Product Relevance

Our objective is to learn text encoder B, such that given a query $q_i^{h_i}$, it should retrieve all the relevant products $P = \{p_1^{g_1}, p_2^{g_2} .. p_j^{g_j}\}$ in a language-agnostic manner from semantic space LA. In this step, we fine-tune parameters of model B by leveraging the D_{cls} dataset. The following loss function ℓ_{cls} is optimized where θ^{cls}_{pos}, θ^{cls}_{neg}, and λ_{neg} are hyper-parameters.

$$l_{cls} = \mathbb{1}_{y_i=1} min(0, \hat{y}_i - \theta_{pos}^{cls})^2 + \lambda_{neg} * \mathbb{1}_{y_i=0} max(0, \hat{y}_i - \theta_{neg}^{cls})^2 \quad (7)$$

Train with hard negatives: Similar to training image encoder I, we fine-tune the model B with product taxonomy as a heuristic to generate negative samples to improve neighborhood quality. Product taxonomy encodes relevance among products and can be exploited to infer various relations among them. We found this information particularly useful for recovering user intent with search queries containing ambiguous tokens. For example: Puma shoes vs Puma backpack. To sample negatives at any given point, we find queries that are close in the current embedding space LA but have a significant cn_i mismatch and add them to D_{neg} as hard negatives and optimize the loss ℓ_{cls}.

3.6 M2S2 Model Training

We first pre-train the image encoder using supervised image data mined using anonymized and aggregated customer behaviour data along with taxonomy based hard negatives. Next, in each epoch of alternate training procedure, we have 2 steps. First step is to train the model on the multilingual alignment task leveraging the pre-trained image encoder. Second step is to train on the relevance classification task leveraging the text encoder in Siamese fashion. Second step utilizes relevance data and taxonomy based hard negatives. Training procedure is outlined in Algorithm 1.

We train the M2S2 model in an alternative training procedure for 2 main reasons. (1) We demonstrate that simple approach of training globetrotter model followed by fine-tuning on relevance data (Finetuned-Globetrotter model) leads to forgetting problem wherein model forgets and performs poorly on the multilingual alignment task. Alternate training procedure alleviates this problem (Section 4.4). (2) Multi-task training incorporating relevance task along with multilingual alignment task is inapt because we want to first learn language-agnostic representations and leverage those for the main task of relevance classification.

Algorithm 1: Training Multilingual Multimodal Semantic Sourcing (M2S2) model

Require: D_{cls}, D_{align}, D_{qq}, D_{cat}, $N_{epochs}^{pre-train}$, N_{epochs} and model hyper-parameters

Initialize M2S2 model parameters θ_{txt} and θ_{img};
Train θ_{img} using D_{pp+} and D_{cat} for $N_{epochs}^{pre-train}$. **for** $epoch = 1$ *to* N_{epochs} **do**
  Train θ_{txt} and θ_{img} on the multilingual alignment task using D_{align};
  Generate D_{qq-} using embedding from θ_{txt} and taxonomy data from D_{qq} ;
  $D_{qq_{epoch}} \leftarrow D_{qq+} \cup D_{qq-}$;
  Train θ_{txt} on the relevance classification task using D_{cls} and $D_{qq_{epoch}}$;
end

4 EXPERIMENTS AND RESULTS

We explore the following research problems in this paper:
RP1: Do Product Images from catalog data improve semantic alignment across languages?
RP2: Is it necessary to use human-annotated relevance data in addition to catalog data?
RP3: Do we need to use product taxonomy based hard negatives?
RP4: Does the performance improve upon using supervision in addition to contrastive learning for training Image model?
RP5: How our model performs in zero-shot experimental setup when we don't have human-annotated relevance data for a given marketplace?
RP6: Does the performance improve upon using Vision Transformer as image encoder and XLM-RoBERTa as text Encoder? Which component has the most impact?

4.1 Datasets

For all experiments, we collect datasets from Amazon marketplaces. The data used for experimentation is collected for 5 different languages: English (EN), German (DE), French (FR), Italian (IT) and Spanish (ES). It spans 8 marketplaces: India (IN), United States of America (US), United Kingdom (UK), Canada (CA), German (DE), French (FR), Italian (IT) and Spanish (ES). There are 4 datasets which are sampled down to a small subset for training as well as testing:

(1) D_{align}: Catalog data from different Amazon marketplaces without explicit supervision comprising of (Product Title, Product Image) pairs used for semantic alignment across languages. Test data is used for evaluating language-agnostic retrieval performance of the models. Training data consists of 1M samples per marketplace. Test data consists of 10K products for each marketplace which are common across all the marketplaces and aren't part of training data. Training and test data spans across 60 different categories of products.
(2) D_{cls}: Human-annotated monolingual relevance data from different Amazon marketplaces comprising of (Query, Product Title, Relevance label) tuples used for learning relevance classification. The average length of the queries and product titles is 2.82 and 18.31, respectively. We sample 60K instances per marketplace (50k + | 10k -) as test data. Train and test datasets are based on random splits and no samples are common between train and test splits.
(3) D_{qq}: Catalog data from Amazon IN marketplace comprising of (Query, Query taxonomy ladder) pairs used for improving the neighborhood quality in semantic space. They are used for generating (Query, Query) hard negatives (QQ Hard Negatives). This makes sure model has distant embeddings when query and product are similar but belong to different taxonomical categories. It also includes catalog data of (Query, Query) positive pairs (D_{qq+}) based on the anonymized and aggregated historical customer purchases. D_{qq+} consists of 1.2M samples and D_{qq} consists of 400k samples.
(4) D_{cat}: Catalog data from various Amazon marketplaces comprising of (Product, Image, Product taxonomy ladder) tuples used for pre-training and improving the performance of image encoder I. They are used for generating (Image, Image)

hard negatives. It includes catalog data of (Product, Product) positive pairs (D_{pp+}) based on the the anonymized and aggregated historical customer purchases. D_{pp+} consists of 4.2M samples and D_{cat} consists of 800k samples.

4.2 Evaluation Metrics based on task

As mentioned in Section 1, the main focus of this work is related to sourcing relevant products for a given query. Hence, we limit our discussion to only sourcing relevant products and evaluate models on relevance classification (ROC-AUC) and decision support metrics (Precision and Recall) outlined below instead of rank-aware metrics.

(1) **Relevance Classification**: ROC-AUC score for the binary relevance classification task.
(2) **Multilingual Semantic Alignment**: For every product in test data of D_{align}, we retrieve 200 nearest neighbours in the semantic space. We measure the multilingual semantic alignment based on whether the nearest neighbours have the corresponding product from other marketplace. We use Recall@k for $k \in \{1, 10, 50, 100\}$ to evaluate the fraction of common products retrieved for each product at a given k.
(3) **Neighborhood Quality in the semantic space**: Test data comprises of 10K queries and 2.4M products from one day of anonymized search data logs from Amazon IN marketplace. For each query we get 200 nearest products in the semantic space resulting in dataset of 2M (Query, Product) pairs. We get relevance labels using a BERT model trained on relevance classification task using English only Dataset. It has classification head on top of CLS token embedding where input is "[CLS] query [SEP] product_title" concatenated sequence. We treat this model as our oracle because it has the best performance on relevance task in the EN (English) language. We use Precision@k for $k \in \{1, 10, 50, 100\}$ to evaluate the fraction of relevant (query, product) pairs at a given k.

4.3 Baseline Models

Monolingual Siamese models fine-tuned on Multilingual-Bert (Monolingual Text-Only S-BERT): We take pre-trained M-BERT model [7] and fine-tune it using D_{cls} relevance data for a given marketplace to get monolingual model corresponding to that marketplace. It relies only on textual data.

Multilingual Siamese model fine-tuned on Multilingual-Bert (Multilingual Text-Only S-BERT): We take pre-trained M-BERT model [7] and fine-tune it using D_{cls} relevance data from all the available marketplaces to get one multilingual model. This model relies only on textual data.

Globetrotter model: We use the pre-trained M-BERT [7] and pre-trained ResNet-50 [9] as backbone and train globetrotter model [36] using D_{align} catalog data from all the available marketplaces.

Finetuned-Globetrotter model: Firstly, we use the pre-trained M-BERT [7] and pre-trained ResNet-50 [9] as backbone and train globetrotter model [36] using D_{align} catalog data from all the available marketplaces. Post that, we fine-tune it using D_{cls} relevance data from all the available marketplaces.

4.4 Results

Based on experiments performed, we will present the quantitative results of our model. To evaluate the ability of sourcing relevant products for a given query, we leverage the metrics outlined in Section 4.2. The M2S2 model trained with Supervised Image model on catalog data as backbone is called **M2S2-SIM** (**M**ultilingual **M**ultimodal **S**emantic **S**earch model using **S**upervised **I**mage **M**odel as backbone). The **M2S2-FSIM** (**M**ultilingual **M**ultimodal **S**emantic **S**earch model using **F**rozen **S**upervised **I**mage **M**odel as backbone) model is trained using Supervised Image model on catalog data as backbone, with the image model's weights frozen during training.

M2S2-ViT model leverages ViT [8] from the pre-trained CLIP model [28] instead of ResNet-50 as image encoder backbone. M2S2-XLMR model leverages pre-trained XLM-RoBERTa [5] instead of M-BERT [27] model as text encoder backbone. M2S2-ViT-XLMR/M2S2-SIM-ViT-XLMR leverages pre-trained ViT from CLIP model and pre-trained XLM-RoBERTa as image and text encoders, respectively. Table 1 shows the ROC-AUC scores of various baseline models and M2S2 model variants. Table 2 shows the Recall@k scores on test dataset of D_{align}. Table 2 also shows the Precision@k scores on the test data of D_{qq}. The best scores are highlighted in **bold**, second best scores are underlined and all the gains reported below are in relative terms.

Relevance classification task analysis: From Table 1, we observe that Globetrotter model which wasn't explicitly fine-tuned/trained on relevance classification task performs the worst. This reveals the importance of having human-annotated relevance data comprising of (Query, Product, Human Judgement) tuples. The catalog data used to train the Globetrotter model doesn't have query data and as such misses many nuances of queries like shorter lengths, broad intents in general, vernacular nature, code-switching and various levels of linguistic sophistication etc. This answers our **RP2**: experiments performed reveal that human-annotated relevance data is necessary in addition to the catalog data .

M2S2 model variants outperform all other baselines with great margins on non-EN speaking marketplaces where we have order of magnitude less human annotated data in comparison to EN marketplace. When compared to Monolingual Text-Only models, we observe ROC-AUC gains of **5.39%-9.18%** across non-EN speaking marketplaces and 1.31% in EN speaking marketplace. We observe that Multilingual Text-Only model improves over monolingual models with gains of **2.92%-5.87%** across non-EN speaking marketplaces demonstrating transfer learning happening to an extent. This further improves by incorporating product images from catalog data of different marketplaces, with the best M2S2 variant outperforming Multilingual Text-Only model with ROC-AUC gains of **2.4%-3.65%** across non-EN speaking marketplaces and 1.38% in EN speaking marketplace.

Multilingual semantic alignment task analysis: From Table 2, we observe that Globetrotter model is having better Recall@k scores in comparison to Multilingual Text-Only model with gains of **4.13%-22.9%** across different thresholds. This demonstrates that Globetrotter model is better at multilingual alignment by using product images as a bridge across languages. This reveals that using product images from catalog data of different marketplaces we can improve the multilingual alignment and thereby answers **RP1**.

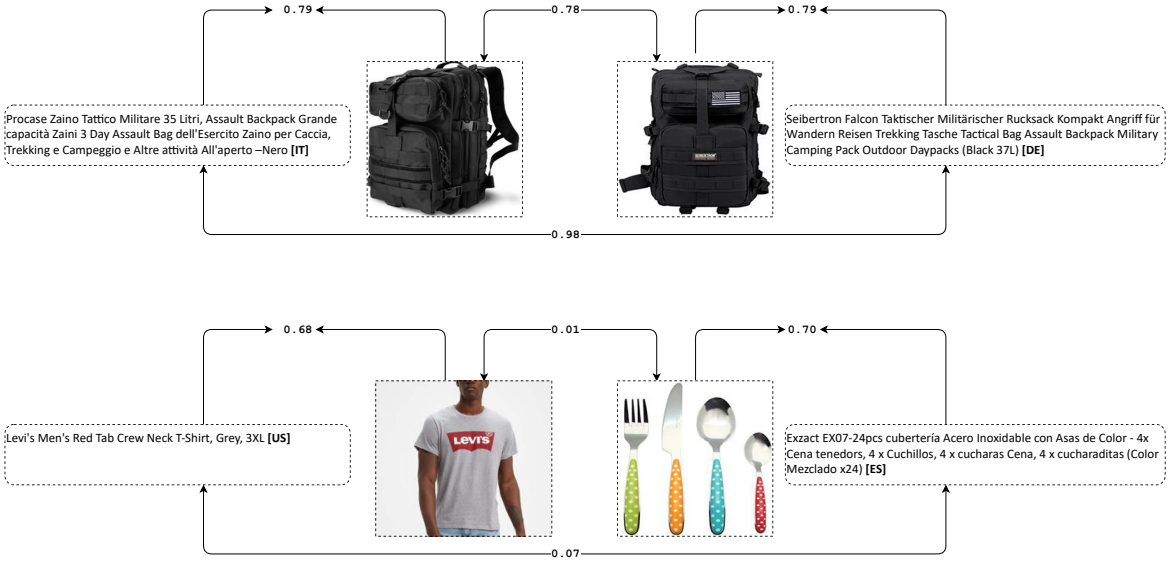

Figure 3: We show an example of positive match (top) and an example of negative match (bottom) demonstrating the multilingual semantic alignment capability of M2S2 models.

Table 1: ROC-AUC scores of various baseline models and M2S2 model variants.

Model	IN	DE	FR	IT	ES
Monolingual Text-Only S-Bert	0.92597	0.85393	0.86083	0.82487	0.8283
Multilingual Text-Only S-Bert	0.92535	0.88348	0.88602	0.87332	0.86252
Globetrotter	0.57532	0.58386	0.53272	0.54022	0.55486
Finetuned-Globetrotter	0.91818	0.88471	0.88766	0.87761	0.86751
M2S2	0.93088	0.8961	0.89741	0.88935	0.87954
M2S2-SIM	0.93421	0.8981	0.9015	0.89297	0.88399
M2S2-SIM w/o QQ Hard Negatives	0.92351	0.886	0.88981	0.88097	0.86858
M2S2-FSIM	0.93329	0.89745	0.90064	0.89401	0.88507
M2S2-ViT	0.93452	0.89762	0.90137	0.89216	0.88217
M2S2-XLMR	0.93788	0.90783	0.9048	0.89859	0.8914
M2S2-ViT-XLMR	0.93733	0.90599	**0.90861**	**0.90078**	**0.89457**
M2S2-SIM-ViT-XLMR	**0.93813**	**0.90793**	0.90725	0.90056	0.894

M2S2 models outperform all the baselines by great margins across all k thresholds, with **34.69%-51.67%** improvements over Multilingual Text-Only model and **18.96%-34.37%** over Globetrotter model. Finetuned-Globetrotter model is having the lowest performance indicating the importance of alternate training approach carried out in M2S2 models.

Analysis of neighborhood quality in semantic space: From Table 2, we observe that Globetrotter model which wasn't trained using D_{cls} and D_{qq} had the lowest Precision@k scores. This again reinforces the importance of having query data which can be seen in case of Finetuned-Globetrotter model with gains of 39.64%-53.77% over Globetrotter model. M2S2 models which incorporates D_{qq} data in the alternate training phase outperform all the baseline models, with improvements of **6.25%-13.42%** & **43.46%-46.46%** over Multilingual Text-Only model & Finetuned-Globetrotter, respectively. This demonstrates the importance of incorporating taxonomy related catalog data to further improve the embedding quality of semantic space and thereby answers **RP3**.

Zero-shot performance analysis: Table 3 shows the zero-shot ROC-AUC scores of various models when the human-annotated relevance data for the target marketplace is unavailable during training. We can observe that M2S2-FSIM model and Finetuned-Globetrotter model are having zero-shot ROC-AUC scores above 0.8. Here, the relevance data D_{cls} for target marketplace is absent during training and only D_{align} catalog data of target marketplace is present which is used for multilingual alignment steps. This demonstrates we can efficiently transfer the learnings of relevance classification from data-rich marketplace to data-deficit marketplace in a language agnostic manner. When compared to Multilingual Text-Only model, M2S2 model has gains of **3.74%-6.25%** across marketplaces,

Table 2: Recall@k|Precision@k scores on test dataset of $D_{align}|D_{qq}$

Model	k=1	k=10	k=50	k=100
Multilingual Text-Only S-Bert	0.09444 | 0.8675	0.35327 | 0.84026	0.48765 | 0.80987	0.54765 | 0.79036
Globetrotter	0.10692 | 0.4601	0.39112 | 0.43158	0.58932 | 0.40867	0.67307 | 0.39691
Finetuned-Globetrotter	0.0755 | 0.6425	0.29762 | 0.63289	0.48606 | 0.61719	0.57219 | 0.60657
M2S2	0.11066 | 0.895	0.44159 | 0.87736	0.65272 | 0.86097	0.73364 | 0.84956
M2S2-SIM	0.11742 | 0.894	0.47758 | 0.88157	0.69144 | 0.86693	0.76925 | 0.85699
M2S2-SIM w/o QQ Hard Negatives	0.07811 | 0.6241	0.30473 | 0.6162	0.47956 | 0.60446	0.56079 | 0.59645
M2S2-FSIM	0.11968 | 0.9057	0.49009 | 0.89114	0.70426 | 0.87439	0.78147 | 0.86271
M2S2-ViT	0.11296 | 0.9006	0.45887 | 0.88401	0.67262 | 0.86509	0.75297 | 0.85323
M2S2-XLMR	0.12379 | 0.9186	0.5069 | 0.90067	0.72439 | 0.88116	0.7995 | 0.86794
M2S2-ViT-XLMR	0.12496 | 0.9217	0.51062 | **0.91207**	0.72681 | **0.89764**	0.80136 | **0.88622**
M2S2-SIM-ViT-XLMR	**0.1272** | **0.9217**	**0.5191** | 0.91042	**0.73678** | 0.895544	**0.80953** | 0.883988

Table 3: Zero-shot ROC-AUC scores of various models

Model	IN	DE	FR	IT	ES
Multilingual Text-Only S-Bert	0.80388	0.79755	0.78201	0.81194	0.79773
Finetuned-Globetrotter	0.82589	0.82391	0.81501	0.8465	0.82945
M2S2-FSIM	**0.83391**	**0.83651**	**0.82057**	**0.85659**	**0.84132**

demonstrating the importance of multilingual semantic alignment training using catalog data comprising of product images.

With respect to **RP5**, our experiments reveal that our models showcase competitive performance in zero-shot settings with ROC-AUC scores being above 0.8 across marketplaces. This is very crucial for E-Commerce services to expand across marketplaces and to launch new language search experience in existing marketplaces; M2S2 models can be deployed to new marketplaces in zero-shot setting providing a great customer experience from the get-go.

Based on all of the above quantitative analysis, we can observe that the best performing M2S2 variant is the one using supervised ViT based image encoder and XLM-R based text encoder. It has **0.46%-1.24%** ROC-AUC gains, **5.24%-9.22%** Recall@k gains and **3.06%-3.38%** Precision@k gains over M2S2-SIM. This answers first part of **RP6**.

4.5 Ablation Studies

Based on Tables 1 and 2, We observe that M2S2-SIM/M2S2-FSIM models outperforms M2S2 model on all the tasks with great Recall@k gains of **4.85%-11.50%** on multilingual semantic alignment task. We also observe that M2S2-SIM-ViT-XLMR has good Recall@k gains of **1.02%-1.78%** on multilingual semantic alignment task while achieving similar performance on other tasks. This answers **RP4**. This demonstrates the benefit of using catalog data with supervision based on customer behaviour and taxonomy based hard negatives for pre-training the image encoder. Removing QQ Hard Negatives hurts the performance on semantic neighborhood quality task with Precision@k decrease of **30.19%-30.44%** (Table 2) along with decrease in other metrics. This reinforces the importance of incorporating taxonomy related catalog data to further improve the embedding quality of semantic space (**RP3**).

Based on Tables 1 and 2, we observe that replacing ResNet-50 model with ViT model leads to ROC-AUC gains of 0.17%-0.44%, Recall@k gains of 2.08%-4.31% and Precision@k gains of 0.21%-0.79%. Replacing M-BERT model with XLM-RoBERTa model leads to ROC-AUC gains of 0.75%-1.35%, Recall@k gains of 8.98%-15.4% and Precision@k gains of 1.84%-2.78%. Therefore, XLM-RoBERTa (text encoder) is adding more value when compared to ViT (image encoder). This answers second part of **RP6**. The best performance is achieved by replacing both ResNet-50 and M-BERT with ViT and XLM-RoBERTa, respectively.

In Figure 3, M2S2 model scores of various embedding heads are depicted for positive and negative pairs wherein text pairs are from different languages. We observe that for positive match, image to image and image to text cosine scores are high and thereby leading to high text to text cosine scores. For negative match, we see that image to text scores are high but image to image score is very low, thereby leading to very low text to text cosine score. This showcases how images act as anchor across different languages. Implementation and hyper-parameters details are provided in Appendix A.1. Qualitative results are provided in Appendix A.2.

5 CONCLUSION

In this paper, we presented a novel deep learning based Multilingual Multimodal Semantic Sourcing (M2S2) model for sourcing relevant products for any given query in a language-agnostic manner. The model utilizes catalog data from various marketplaces consisting of product images for learning multilingual semantic alignment. It uses human-annotated relevance data for learning relevance classification and transfers relevance knowledge to low resource marketplaces. Experiments performed reveal that our model outperformed all baselines by great margins across various tasks.

REFERENCES

[1] Ahuja, A., Rao, N., Katariya, S., Subbian, K., and Reddy, C. K. Language-agnostic representation learning for product search on e-commerce platforms. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (New York, NY, USA, 2020), WSDM '20, Association for Computing Machinery, p. 7–15.

[2] Bjerva, J., and Östling, R. Cross-lingual learning of semantic textual similarity with multilingual word representations. In *NODALIDA* (2017).

[3] Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *ECCV* (2020).

[4] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv abs/2003.10555* (2020).

[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *ACL* (2020).

[6] Desai, K., and Johnson, J. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR* (2021).

[7] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).

[9] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778.

[10] He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv abs/2006.03654* (2021).

[11] Hu, Q., Yu, H.-F., Narayanan, V., Davchev, I., Bhagat, R., and Dhillon, I. S. Query transformation for multi-lingual product search. In *The 2020 SIGIR Workshop on eCommerce* (San Diego, USA, Aug. 2020), ACM.

[12] Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (New York, NY, USA, 2013), CIKM '13, Association for Computing Machinery, p. 2333–2338.

[13] Khattab, O., and Zaharia, M. A. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).

[14] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Y. Bengio and Y. LeCun, Eds.

[15] Lample, G., and Conneau, A. Cross-lingual language model pretraining. In *NeurIPS* (2019).

[16] Lau, J. H., Newman, D., and Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg, Sweden, Apr. 2014), Association for Computational Linguistics, pp. 530–539.

[17] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[18] Li, G., Duan, N., Fang, Y., Jiang, D., and Zhou, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI* (2020).

[19] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *ArXiv abs/1908.03557* (2019).

[20] Liu, X., Gao, J., He, X., Deng, L., Duh, K., and Wang, Y.-y. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, Colorado, May–June 2015), Association for Computational Linguistics, pp. 912–921.

[21] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation.

[22] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] Lu, H., Hu, Y., Zhao, T., Wu, T., Song, Y., and Yin, B. Graph-based multilingual product retrieval in E-commerce search. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers* (Online, June 2021), Association for Computational Linguistics, pp. 146–153.

[24] Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS* (2019).

[25] Nigam, P., Song, Y., Mohan, V., Lakshman, V., Ding, W. A., Shingavi, A., Teo, C. H., Gu, H., and Yin, B. Semantic product search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2019), KDD '19, Association for Computing Machinery, p. 2876–2885.

[26] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[27] Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 4996–5001.

[28] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML* (2021).

[29] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21*, 140 (2020), 1–67.

[30] Reimers, N., and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP* (2019).

[31] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv abs/1910.01108* (2019).

[32] Sasaki, S., Sun, S., Schamoni, S., Duh, K., and Inui, K. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 458–463.

[33] Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. A latent semantic model with convolutional-pooling structure for information retrieval. In *CIKM* (November 2014).

[34] Singhal, K., Raman, K., and ten Cate, B. Learning multilingual word embeddings using image-text data. *CoRR abs/1905.12260* (2019).

[35] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations* (2020).

[36] Surís, D., Epstein, D., and Vondrick, C. Globetrotter: Unsupervised multilingual translation from visual alignment. *CoRR abs/2012.04631* (2020).

[37] Tan, H. H., and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP* (2019).

[38] van der Maaten, L., and Hinton, G. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (2017), I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., pp. 5998–6008.

[40] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *CoRR abs/1910.03771* (2019).

[41] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, June 2021), Association for Computational Linguistics, pp. 483–498.

[42] Yu, P., Fei, H., and Li, P. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021* (New York, NY, USA, 2021), WWW '21, Association for Computing Machinery, p. 1029–1039.

[43] Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. Contrastive learning of medical visual representations from paired images and text. *ArXiv abs/2010.00747* (2020).

[44] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. Ernie: Enhanced language representation with informative entities. In *ACL* (2019).

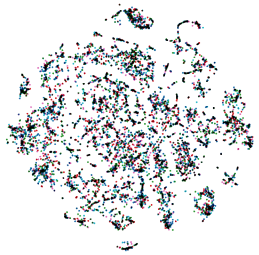

Figure 4: T-SNE plot of test data of D_{align} with different colour dots for marketplaces based on M2S2-FSIM model.

Source Query: jumkalu earrings traditional	Source Query: gas chulha accessories	Source Query: big helicopter toy
A K Jewellers triangle jhumka	Rajeev Heavy Body Based LPG Gas Stove Jali Parts (Steel Round Shape Cross Point 7.5 inch) Support Gas Stoves -Pack of (1)	Velocity Helicopter Remote Control
10010 Golden Stylish Long Jhumka White	Ninki Fresh LPG Stainless Steel Gas Stove Parts Pan Support (Black 19 x 19 x 2.5 cm) 2-Piece	divsheen limited edition kids flying exceed helicopter with remote 3d lights and usb charger (Multi colors)
CEYLONMINE yellow sapphire earrings natural & original stone pukhraj stud earrings for women & girls	Freshee Square Aluminium Foil Stove Disposables Burner Guard (Silver) - 8 Pieces	Vezol® RC Car for Kids Off-Road RC Car with Camera 1/16 30 Minutes Operation Time 20 km/h Speed 2.4 GHz WiFi FPV Real Time Remote Control Car for Kids
Contemporary Indo western Jhumki Earring	Freshee Square Aluminium Foil Stove Disposables Burner Guard Pack of 2 x 8pcs Keeps The Burner Area Clean with No Scratches of Cleaning Helps to Conserve Water	new kids velocity mini outdoor helicopter rechargeable remote control with infrared sensors (Multi color)
PRANAAM Traditional Jhumkas for Women	Ninki Fresh LPG Gas Stove Parts 2-Piece Black Pan support gas stove (Black)	kids velocity outdoor/indoor helicopter rechargeable remote control with infrared sensors and unbreakable blades (Multi color) with data cable included for charging

Figure 5: Five nearest products for a given query from test dataset of D_{qq} based on M2S2-FSIM model.

A APPENDIX

A.1 Implementation Details

We implemented all the experiments using PyTorch [26] and HuggingFace [40]. The backbone for text models was pre-trained multilingual-bert-base-cased (M-BERT) [7] and for image models it was pre-trained ResNet-50 [9] unless otherwise specified. Output embedding dimension for all experiments was fixed to be 512. Monolingual Text-Only S-Bert and Multilingual Text-Only S-Bert were trained using weighted random sampler for 3 epochs using Adam optimizer with a learning rate of 5e-5. Globetrotter model [36] was trained using Adam optimizer with learning rate of 1e-3 for 10 epochs with weights for all different loss components set to 1. Finetuning of Globetrotter model was done using Adam optimizer [14] with learning rate of 5e-5 for 3 epochs using weighted random sampler. M2S2 model was trained using Adam optimizer with a learning rate of 5e-5 for 10 epochs with weights for all loss components of multilingual alignment task set to 1 and weight for negative samples λ_{neg} in relevance classification task set to 5. For all experiments, θ^{cls}_{pos} and θ^{cls}_{neg} were set to 1.0 and 0.0, respectively. For all experiments having loss ℓ_{align} from equation 6, τ was set to 0.1 for loss components ℓ_v and ℓ_{xm}; it was set to 1.0 for loss component ℓ_{xl}. For all experiments having loss ℓ_{align} from equation 6, margin m was set to 0.4. Pre-training of image model I was carried out for 12 epochs using Adam optimizer with a learning rate of 5e-5. θ_{pos} and θ_{neg} were set to 0.8 and 0.0, respectively. All experiments were performed on 4 Nvidia Tesla V-100 GPUs on p3.8xlarge EC2 instance on AWS . All hyperparameters were chosen empirically based on experiments performed.

A.2 Qualitative Results

Figure 4 depicts the embedding of product titles from test data of D_{align} using t-distributed stochastic neighbor embedding (TSNE) method [38]. We see no language specific cluster signifying the language-agnostic nature of M2S2 models. Figure 6 depicts the 5 nearest products for a given product from test dataset of D_{align}. We can observe that model is able to retrieve relevant products in a language-agnostic manner wherein it is either retrieving same product from different languages or other relevant products across languages. This shows the multilingual semantic alignment capability of M2S2 models. Figure 5 depicts the importance of using taxonomy data for mining query-query hard negatives. We can observe that model is able to retrieve relevant products to the query while handling nuances like vernacular nature of phrases like "jumkalu", transliteration phrase like "chulha" and broad intents for product discovery like "big helicopter toy". This demonstrates the value of using D_{qq} for improving the neighborhood quality in the semantic space. Figure 7 displays few examples of image-image hard negatives generated using taxonomy information from D_{cat} data. It is used for pre-training Image model as specified in section 3.3.

Figure 6: Five nearest products for a given product from test dataset of D_{align} based on M2S2-FSIM model. Marketplace is mentioned at the end of corresponding titles.

Figure 7: Few examples of Image-Image hard negatives generated using D_{cat} data.