Catalog Phrase Grounding (CPG): Grounding of Product Textual Attributes in Product Images for e-Commerce Vision-Language Applications

Wenyi Wu
Amazon
Seattle, USA
wenyiwu@amazon.com

Karim Bouyarmane
Amazon
Seattle, USA
bouykari@amazon.com

Ismail Tutar
Amazon
Seattle, USA
ismailt@amazon.com

ABSTRACT

We present Catalog Phrase Grounding (CPG), a model that can associate product textual data (title, brands) into corresponding regions of product images (isolated product region, brand logo region) for e-commerce vision-language applications. We use a state-of-the-art modulated multimodal transformer encoder-decoder architecture unifying object detection and phrase-grounding. We train the model in self-supervised fashion with 2.3 million image-text pairs synthesized from an e-commerce site. The self-supervision data is annotated with high-confidence pseudo-labels generated with a combination of teacher models: a pre-trained general domain phrase grounding model (e.g. MDETR) and a specialized logo detection model. This allows CPG, as a student model, to benefit from transfer knowledge from these base models combining general-domain knowledge and specialized knowledge. Beyond immediate catalog phrase grounding tasks, we can benefit from CPG representations by incorporating them as ML features into downstream catalog applications that require deep semantic understanding of products. Our experiments on product-brand matching, a challenging e-commerce application, show that incorporating CPG representations into the existing production ensemble system leads to on average 5% recall improvement across all countries globally (with the largest lift of 11% in a single country) at fixed 95% precision, outperforming other alternatives including a logo detection teacher model and ResNet50.

KEYWORDS

Phrase Grounding, Object Detection, Transformers, MultiModal Model, Natural Language Understanding

ACM Reference Format:
Wenyi Wu, Karim Bouyarmane, and Ismail Tutar. 2022. Catalog Phrase Grounding (CPG): Grounding of Product Textual Attributes in Product Images for e-Commerce Vision-Language Applications. In *Proceedings of KDD '22: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Incorporating multimodal understanding of image and textual data of products is essential for many e-commerce applications. A typical product page on an e-commerce website consists of a product title in the form of short textual description, and an image representation of the product. The image can display an isolated product image or the image of the product in the context of use in an environment. Additional fields can be found in the product page, such as brand, dimensions, etc. An e-commerce catalog constitutes, therefore, a very rich corpus that can be used to design self-supervised vision-language tasks for the pre-training of deep learning product-vision-language understanding models.

One such vision-language self-supervision task that can be crafted from the e-commerce catalog and that takes particular advantage of its specificities is phrase grounding[4]. Phrase grounding consists in *associating* (or *grounding*) a textual phrase or part of it to a specific region of an image. The nature of e-commerce product data allows for expressive and domain-specific multi-task phrase grounding pre-training: grounding product title noun to isolated product region in the image, grounding product brand field to brand logo region in the product image, etc. We call the multi-task phrase grounding of e-commerce specific entities such as product-brand-to-logo and product-noun-to-object *Catalog Phrase Grounding* (CPG). CPG outputs semantic rich representations that are particularly suited for e-commerce domain-specific downstream tasks.

Pre-training of the CPG model is done in a self-supervised way with two teacher models: a pre-trained general-domain phrase grounding model, and a specialized logo-detection model. The student model learns to combine the knowledge distilled from both teacher models in multi-teacher multi-task learning setting. The tasks and phrases are crafted and self-generated from the catalog corpus, allowing to benefit from very large amount of unlabeled data. Figure 1 illustrates self-generated annotations for two products.

The learned CPG embeddings are powerful, semantic-rich, fine-grained representations of e-commerce product images. We demonstrate it with a challenging task: product-brand matching. A brand is a complex e-commerce entity that is represented by fields such as the brand name, brand logo (rarely), as well as a set of representative sample products. Brand name alone is not sufficient to characterize the brand entity, due to the sheer number of brands and homonyms. Matching a product to a brand entity thus consists in using the information from the brand and the representative sample products and inferring whether the query product *belongs to* that brand. We

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.
KDD '22, August 14–18, 2021, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

show that GPG representations significantly improve the previous SOTA brand-matching system.

2 RELATED WORK AND CONTRIBUTIONS

A standard joint visual and textual understanding model [5, 10] is typically trained with a fixed set of visual concepts (classes) on vision-language tasks, such as visual question answering [1], object detection [16] and phrase grounding. MDETR[9] extended transformer based object detection model, i.e. DETR [3], to a modulated multimodal model trained with two tasks: object detection and phrase grounding. Therefore, MDETR could be pre-trained with 1.3M text-images pairs having explicit alignment between phrases in text and objects in the image from combined pre-existing datasets, e.g. MS COCO[12], Flickr30k [14] and etc.

In order to further expand the visual concepts of image regions beyond vocabularies of pre-existing datasets, a recent line of work [8, 11, 15] considers using web-scale raw image-text pairs. CLIP [15] demonstrates that an image-level representations can be learned effectively through alignment between raw image-text pairs collected from the internet. Following the same idea, GLIP [11] pre-trains a phrase grounding model with 24 million web-crawled image-text pairs. The regions of interest in images were detected by a pre-trained teacher model. It has been shown that the pre-trained GLIP learned effectively from the broader set of raw text to generate semantic enriched region-level visual representations. However, these models are not trained for e-commerce specific vocabulary and entities in synergy with downstream e-commerce applications.

Our proposal, CPG, is a transformer encoder-decoder architecture based model learning semantic rich visual representations for e-commerce specific entities through multiple domain-specific tasks. Following the trend of using free-form text, we train the CPG model with 2.3M product entities synthesized from an e-commerce site in a self-supervised fashion. The bounding boxes for product-brand-to-logo grounding task are generated from a YOLO-based logo detection model [6] while the product brands exist in product page already. The bounding boxes for product-noun-to-object task are generated by a pre-trained general domain modulated detection model[9] conditional on noun phrases parsed from free-form product title using a general NLP parser[2].

We present usage of the CPG representations as ML features to address one of major challenges of downstream product-brand matching application, which is differentiating homonym brands, especially when logos are absent, as shown in Figure 2. The GPG representations contain comprehensive visual-language understanding of logos, brand strings, product details for the query product entity and for all brand representative product entities. Therefore, the similarity between them shed light on identifying the correct brand of an input product from homonym ones, either through straightforward logo comparison or through product region comparison in the absence of logos.

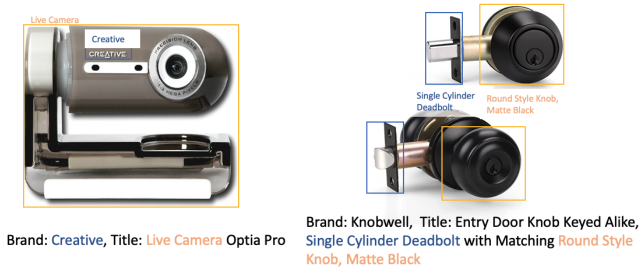
We summarize the contributions of our work as follows:

- We propose an efficient and scalable method to learn semantic rich visual representations for e-commerce products in a self-supervised fashion. We leverage massive raw product text data and images and apply teacher models to obtain region-phrase alignment annotations. In this way, we extend the limited general vocabulary to substantial visual concepts expressed in the e-commerce catalog.
- We transfer knowledge from two teacher models: a logo detection model and a general domain phrase grounding model by leveraging high-confidence predictions as pseudo labels for catalog specific tasks so that CPG benefits from both general-domain knowledge and specialized catalog knowledge.
- We leverage CPG representations to address challenging product-brand matching task and show improved performance.

3 MODAL ARCHITECTURE AND TRAINING

3.1 CPG Model Architecture

CPG in this paper grounds logos and isolated product details in image to the brand attribute and noun phrases in free-form title simultaneously. We manually craft the caption of product image by concatenating seller provided brand string and the product title together and inserting comma in between. To facilitate understanding of the logo concept and distinguishing brand tokens from general



Brand: Creative, Title: Live Camera Optia Pro

Brand: Knobwell, Title: Entry Door Knob Keyed Alike, Single Cylinder Deadbolt with Matching Round Style Knob, Matte Black

Figure 1: Left: logo localized by the logo detection teacher model and product localized by the phrase-grounding teacher models; Right: only product regions localized by the phrase-grounding teacher model. Teacher models can locate rare entities expressed in product titles.

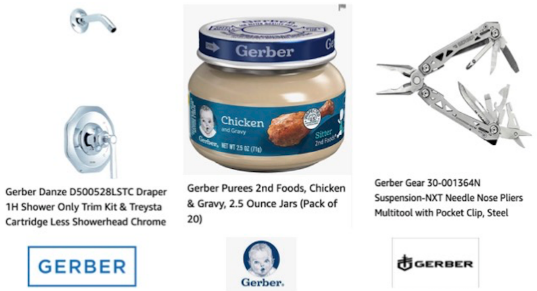

Gerber
Chicken
and Gravy
Gerber Danze D500528LSTC Draper 1H Shower Only Trim Kit & Treysta Cartridge Less Showerhead Chrome
Gerber Purees 2nd Foods, Chicken & Gravy, 2.5 Ounce Jars (Pack of 20)
Gerber Gear 30-001364N Suspension-NXT Needle Nose Pliers Multitool with Pocket Clip, Steel
GERBER
Gerber.
GERBER

Figure 2: Homonym brands are independent brands associated with the same name (e.g. Gerber) selling different products. The *Gerber* baby food has the logos in image while *Gerber* tools and *Gerber* plumbing fixtures don't.

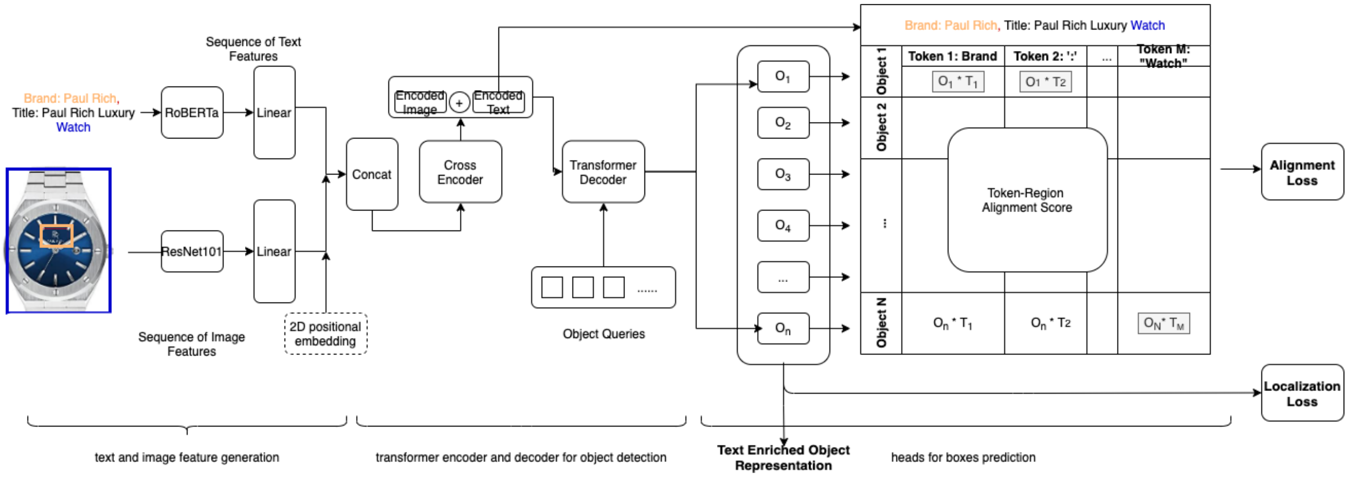

Sequence of Text Features
Brand: Paul Rich, Title: Paul Rich Luxury Watch
RoBERTa
Linear
ResNet101
Linear
Sequence of Image Features
2D positional embedding
Concat
Encoded Image
+
Encoded Text
Cross Encoder
Transformer Decoder
Object Queries
......
O1
O2
O3
O4
...
On
Brand: Paul Rich, Title: Paul Rich Luxury Watch
Token 1: Brand
Token 2: ':'
...
Token M: "Watch"
Object 1
Object 2
⋮
Object N
O1 * T1
O1 * T2
Token-Region Alignment Score
On * T1
On * T2
ON * TM
Alignment Loss
Localization Loss
text and image feature generation
transformer encoder and decoder for object detection
Text Enriched Object Representation
heads for boxes prediction

Figure 3: CPG has a unified framework for logo detection and product phrase grounding. CPG jointly train a text encoder, a image encoder, and a transformer-based region detection model (DETR) to predict the correct pairings of the mixture of (logo, brand) and (product, phrases) training examples and region locations simultaneously.

vocabulary, e.g. brand *Apple* versus fruit *apple*, we prefix the brand string with "Brand :" and prefix the free-form title with "Title:". Therefore, given the product with brand "Paul Rich" and title "Paul Rich Luxury Watch", the crafted caption is "Brand: Paul Rich, Title: Paul Rich Luxury Watch". Given the product image and the augmented caption, CPG is trained to ground logo regions to "Brand: Paul Rich" and to ground the watch region to "watch". We illustrate CPG model architecture in Figure 3.

The proposed CPG model architecture consists of 3 parts. First, the manually crafted image caption is tokenized and encoded using a pre-trained text encoder: RoBERTa[13]. The text encoder has 12 transformer encoder layers, each with hidden dimension of 768 and 12 heads in the multihead attention. The product image is encoded using a pre-trained image encoder: ResNet101[7]. Additionally, we concatenated encoded image vector with 2-dimensional positional embedding to conserve the spatial information. The encoded text and image are projected into a shared latent space by two independent linear projection functions.

Second, the image and textual features are concatenated as a multimodal vector and fed to a joint transformer encoder with cross attention between image and textual features. Following DETR [3], we feed the model with a fixed small set of random initialized object queries (100), each of which is used to represent an object, such as a logo or a product region, in the image. CPG applies a transformer decoder to object queries while cross attends to the final hidden state of the joint encoder. We use 6 encoder layers, 6 decoder layers and 8 attention heads in the attention layers. The learned object queries are textual-attribute-semantics enriched visual representations for logos and products, which are leveraged for downstream applications.

Finally, CPG reasons the object locations and relations between phrases and objects simultaneously without differentiating logo objects versus product objects. Each object representation is fed into two prediction heads: a token-region alignment head and a bounding box regression head, which are trained with contrastive alignment loss $\mathcal{L}_{align}$ and localization loss $\mathcal{L}_{loc}$, respectively. CPG aims to optimize the following combined loss function.

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{loc} \tag{1}$$

For contrastive alignment loss between noun phrases spanned multiple tokens and object queries, we use the same fine-grained contrastive loss proposed in MDETR[9]. We consider the maximum number of tokens to be M, the number of the fixed set of object queries to be N, the set of tokens representing object o_i to be T_i^+ and the set of objects associating with token t_i to be O_i^+. The contrastive loss for objects is normalized by the number of positive tokens:

$$l_o = \sum_{i=0}^{N-1} \frac{1}{|T_i^+|} \sum_{j \in T_i^+} -log\left(\frac{exp(o_i^T t_j/\tau)}{\sum_{k=0}^{M-1} \exp(o_i^T t_k/\tau)}\right) \tag{2}$$

By symmetry, the contrastive loss for all tokens is given by:

$$l_t = \sum_{i=0}^{M-1} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -log\left(\frac{exp(t_i^T o_j/\tau)}{\sum_{k=0}^{N-1} \exp(t_i^T o_k/\tau)}\right) \tag{3}$$

where τ is a temperature parameter set to 0.07. The average of l_t and l_o is used as contrastive alignment loss $\mathcal{L}_{align}$ to be minimized during model training.

3.2 Scalable Training with Pseudo Labels

In order to extend general vocabulary to noun phrases in product titles in e-commerce platform, the traditional way is to manually label phrases in corpus. Furthermore, to train a unified model for logo detection and product phrase grounding, it requires manual labels of bounding boxes which is expensive. Therefore, to scale the CPG model training with abundant textual concepts in e-commerce, we query 2.3M raw product images and textual attributes from English countries of an e-commerce site and use them in a self-supervised framework. In order to generate product-noun-to-object grounding pseudo labels, we start with noun phrases extraction. Given augmented product image caption combining product brand

and title attributes, as described in Section 3.1, we apply general Natural Language Processing parser [2] to identify all noun words in title and include all leading adjective words prior to noun words to obtain noun phrases. We then apply a unified object detection and phrase grounding teacher model, MDETR[9] to detects product regions conditioned on the extracted noun phrases. We obtain 3.2M unique noun phrases from titles with 6.1M associated bounding-boxes. To collect product-brand-to-logo grounding pseudo labels, we simply apply a logo detection model [6] that is trained with e-commerce product images to localize logo regions in the image if exist. Then we associate logo regions to the brand section in augmented image captions to obtain 92k logo region grounding labels in total. We illustrate two sample annotations in Figure1, the MDETR can localize rare language concepts expressed in the free-form product title, like *camera*, *single cylinder deadbolt*, and *round style knob*. Finally, we train the CPG model on these two tasks simultaneously with two types of grounding labels together.

Our self-supervised data augmentation is inspired by GLIP[11] which applied the pre-trained GLIP model to obtain pseudo-labels of raw web-crawled image-text pairs to finetune the same GLIP model. We first collect samples from an e-commerce site to scale grounding data and especially to enrich e-commerce related semantic concepts which is different from general domain. We step further to obtain pseudo-labels from two pre-trained models so that as a student model, CPG model benefits from both general knowledge transferred from MDETR and specialized brand knowledge transferred from the logo detection model. Because of cross attention between logos and product details, CPG model enables product understanding in both the general context and the brand specific context. Furthermore, CPG model enables comprehensive brand understanding by unifying logo understanding and representative sample products understanding. Therefore, CPG model outperforms the logo detection teacher model when applying to the product-brand matching task, which will be described later in Section 4.

4 PRODUCT-BRAND MATCHING APPLICATION

We show the effectiveness of CPG representations pre-trained with pseudo labels by applying them to a downstream e-commerce application: product-brand matching. Brand is a key attribute impacting customers' shopping decisions, which however is nontrivial to infer because of homonyms. Differentiating homonym brands is challenging and can only be done if different logos and/or different products sold by brands are provided. For example, we should map a*Gerber* knife to the right *Gerber* in Figure 2 instead of the other two, because we learn from the representative product that right *Gerber* sells gears while the middle one sells baby food and the left one sells plumbing fixtures. For the middle *Gerber*, we can easily map products correctly by identifying the logo. Therefore, in order to map products to brands correctly, we have to comprehensively understand the brand entity consisting of multiple products and logos (may not exist). The number of brand representative products varies per brand because of remarkable brand scope difference. Some brands, like *3M*, sell products from dozens of different categories while others focus on one specific category. We formulate the product to brand matching problem as an asymmetric entity matching problem, which makes binary predictions regarding whether a product belongs to a brand. The existing state-of-the-art product-brand matching system ensembles a wide variety of features based on product text data (details in Section 5.2.1). It fails to capture image features and hence is incapable of distinguishing homonym brands by logos or other fine-grained regions in product images.

We use the semantics rich CPG representations extracted from the model with high confidence (> 0.5) as additional ML features to the existing system. We can not simply concatenate CPG representations and feed to downstream task because the number of representative products per brand has a long tailed distribution. The number of object representations with high confidence also varies from one image to another. As a result, concatenating them and padding to the longest makes the feature vector extraordinary long, which increases the matching system complexity and therefore breaks the latency requirement for real-time use cases. Another way is to truncate CPG representations to a fix number. However, it will prevent model from making informative predictions. For example, if the *Apple* brand has only one representative product, e.g. a computer, model can not tell if *Apple* phone cases should be mapped to it. Therefore, in order to utilize all information but still have a light enough system to satisfy latency requirement, we first compute distance-based similarity scores, i.e. Euclidean distance and Cosine distance, between CPG representations of the input product and CPG representations of each representative product. Then the summary statistics, i.e. minimum, maximum, medium and variance, of similarity scores are fed to ensemble system. In addition, we feed two boolean values to indicate whether the number of learned CPG representations is 0 for the product and for the brand, respectively. We refer these summary statistics derived from CPG representations as CPG features.

5 EXPERIMENTATION AND RESULTS

We first train the CPG model using augmented product image text pairs with pseudo labels. Then we evaluate the textual-attribute-semantic rich object representations learned by CPG by supplementing them to existing product-brand matching model. We report the relative gain of recall at 95% precision in 9 countries as we need to meet a high precision threshold for deployment to ensure customer shopping experience. We further compare the visual representations learned by CPG with representations learned from two types of vision model: the logo detection model and image-level understanding model, i.e. ResNet50.

5.1 Product-Brand Matching Dataset

All samples for the product-brand matching task were in (product, brand) pair format. Input product was a structured entities with a fixed set of attributes, such as title, brand and image. The textual attributes could be inaccurate or even missing. Input brand was also a structured entities with varying number of representative products that shared the same data structure as the input product.

We collected 50, 000 (product, brand) training pairs per country from 6 countries (we will refer them as A-F). The textual attributes in these countries were in English and multiple Romance languages, e.g. French, Spanish and etc. The training set were randomly split into 80% training set and 20% validation set. For test, we collected 20, 000 (product, brand) pairs per country from 9 countries globally consisting of A-F and 3 new English countries (we would refer to them as G, H and I). We used different sampling strategies to collect training and test samples. For training, in order to utilize auditing resources efficiently, we excluded trivial negative pairs, the products in which were obvious generic products irrelevant to any known brand entity. To collect test pairs, we randomly sampled from the catalog of an e-commerce site. The catalogs in different countries showed different patterns. The proportion of products with homonym brands in country I was the highest, two times higher than lowest proportion in country F, and therefore, it was expected to show performance lift once we included image signals. The ground truth of a pair was manually labeled as positive if the product belonged to the brand and labeled as negative if not.

5.2 Baseline Models

5.2.1 Existing ensemble model with Text Features. The existing product-brand matching model ensembles in a total of 139 features containing manually crafted syntactic similarity features and ML features learned from base models, e.g. brand extraction model, textual attributes understanding model and etc. The brand extraction model tackles the challenge of missing brand attribute when brand strings are mentioned in other textual attributes. The textual attributes understanding model captures semantic similarity between the product and the brand despite the brand string variation, e.g. similarity between *James Bond 007 Fragrances* and *007 Frangrances*. This existing ensemble model doesn't contain any signal from product image, which serves as the first baseline to the ensemble model with features derived from object representations learned from CPG. The existing model is a country-aware model trained with all training samples collected from 6 countries. The performance of the ensemble model is evaluated in each of 9 countries.

5.2.2 Ensemble model with logo features. One typical way to distinguish homonym brands is comparing logos. We applied the YOLO-based logo detection teacher model to both input product image and brand representative product images and used detected logos as features. Because the number of representative products per brand varied, directly using detected logos led to various input length. Furthermore, the detected logo region similarity could be impacted by the original product image quality, size, and angle. To address these challenges, we first leveraged encoded vector from the last hidden layer of the logo detection model to replace raw detected logo regions as input features. Then, to address various input length challenge and to have fair comparison with CPG representations, we supplemented the existing ensemble model with the same set of summary statistics based on similarity between detected logos, as described in 4. We refer this set of features as logo features.

5.2.3 Ensemble model with ResNet50 features. An alternative way to leverage image information is through image-level understanding. We fine-tuned a ResNet50 model using product images synthesized from English e-commerce catalogs. The data collection process was independent of CPG self-supervised training data collection process. The fine-tune task was to detect whether images are from duplicate products. The training samples contained similar yet different products with similar images as well as duplicate products with images in different angles, lights and size. Therefore, the fine-tuned ResNet50 model learned e-commerce specialized image patterns and learned to pay attention to both local regions as well as the whole image. We used vectors from last hidden layer as an image-level representations for product image. The same set of summary statistics of Euclidean and Cosine distances between image representations of the input product and of the brand representative products were supplemented to existing ensemble model for re-training and evaluation. We refer this set of features as ResNet50 features

5.3 Results

We note that all results reported in this paper are in absolute terms. We evaluate model performance using recall at high precision because we need to prevent incorrect brand mappings from impairing customers' shopping experience. Therefore, we report relative gains in terms of recall at precision 90% and recall at precision 95% and denote lift as $\Delta R@P90$ and $\Delta R@P95$ in tables.

5.3.1 Comparison with existing ensemble model with text features. Table 1 shows the performance lift caused by adding CPG features into the existing ensemble model with text features only as described in 5.2.1 across 9 countries. From the table, we see that leveraging similarity between object representations learned from CPG leads to significant performance jump on top of existing ensemble model in all countries globally except one country C. In country C, the model with CPG features performs comparably with the baseline model because textual attributes quality is high in country C where we observe less inaccurate or missing attributes. Therefore, the model with text features already contain enough product information to distinguish homonym brands by understanding the textual attributes. As expected, in the country with largest homonym proportion, I, we observe the largest performance lift. We conclude that CPG representations provide semantic rich visual signals to differentiate homonym brands.

5.3.2 Comparison with ensemble model with logo features. Table 2 shows the performance lift of the ensemble model with CPG features over the ensemble model with logo features as described in Section 5.2.2 across 9 countries. We can see that, leveraging object representations learned from CPG outperforms detected logo representations learned from the teacher model in 7 countries while these two models perform comparably in C and D. Despite the performance lift in country I in table 2 is lower than in table 1, we still observe positive performance lift. This indicates that leveraging logo features can only mitigate the homonym challenge partially. It doesn't solve all problems because not all product images contain logos. In fact, only 40% of product images have logo detected with high confidence (> 0.5). For cases where the input product image doesn't contain logos, the object representations learned from CPG can provide fine-grained product image understanding to guide correct brand mappings.

5.3.3 Comparison with ensemble model with ResNet50 features. Table 3 shows the performance lift of the ensemble model with CPG features over the ensemble model with image-level understanding

	A	B	C	D	E	F	G	H	I
Δ R@P90	1.8%	0.1%	0.0%	0.0%	1.5%	0.0%	2.4%	8.2%	9.1%
Δ R@P95	3.9%	5.0%	0.0%	4.0%	4.5%	5.6%	7.6%	7.1%	**11.4%**

Table 1: Performance gains of supplementing CPG features to the existing ensemble model with textual features only in 9 countries

	A	B	C	D	E	F	G	H	I
Δ R@P90	1.7%	0.1%	0.0%	0.0%	2.9%	1.4%	0.9%	9.0%	8.0%
Δ R@P95	1.6%	4.0%	0.0%	0.0%	7.3%	6.9%	4.6%	**13.3%**	3.8%

Table 2: Performance gains of leveraging CPG features over leveraging features derived from the logo detection teacher model in 9 countries

	A	B	C	D	E	F	G	H	I
Δ R@P90	0.6%	0.0%	0.0%	0.0%	2.9%	0.0%	-0.4%	5.7%	8.7%
Δ R@P95	4.1%	3.8%	0.0%	2.0%	2.8%	2.8%	5.2%	4.8%	**10.5%**

Table 3: Performance gains of leveraging CPG features over leveraging features derived from the ResNet50 model fine-tuned on catalog from English countries for image matching in 9 countries.

features learned from the ResNet50 model as described in 5.2.3 across 9 countries. We see that CPG features lead to higher performance gains in all countries except country C and G. In country G, both models perform comparably at precision 90% and our model performs better at precision 95%. All ensemble models perform similarly in country C. The lift of leveraging CPG features on top of leveraging ResNet50 features are mainly from two sources: more fine-grained image understanding and logo specialized knowledge. In country I, where logo features lead to performance lift, we see that CPG features outperform ResNet50 features significantly. This demonstrates that CPG model effectively transfers knowledge from the logo detection teacher model while ResNet50 features are lack of logo specialized knowledge.

6 CONCLUSION

In this paper, we present CPG which learns fine-grained semantically-rich visual representations of products. We investigate how to train the model with catalog-scale raw product attributes in a self-supervised fashion by transferring knowledge from two teacher models. Therefore, the learned representations of logos and isolated product details from the same latent space provide an integrated understanding of brands and products. CPG representations learned from pre-training show promising results on a crucial and challenging e-commerce application: product-brand matching. We further show that integrated understanding of products and brands makes CPG competitive with the logo detection teacher model. It's worth future study to apply CPG representations to other e-commerce applications such as duplicate detection.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[4] Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*. 824–832.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[6] István Fehérvári and Srikar Appalaraju. 2019. Scalable logo recognition using proxies. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 715–725.

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

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[9] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1780–1790.

[10] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[11] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2021. Grounded Language-Image Pre-training. *arXiv preprint arXiv:2112.03857* (2021).

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[14] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

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

[16] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and*

learning systems 30, 11 (2019), 3212–3232.