Semi-supervised Learning and Visual Transformers for Product Attribute Extraction from E-commerce Images

Brayan Impata* biimpata@amazon.com
Manuel Lagunas* mlgns@amazon.com
Victor Martinez vicmg@amazon.com
Virginia Fernandez virfer@amazon.com
Christos Georgakis georgak@amazon.com
Sofia Braun brasofia@amazon.com
Felipe Bertrand felipb@amazon.com

Abstract

Extracting product attributes from images involves classifying subtle differences between similar objects. Visual Transformers (ViT) are powerful but usually supervised, while e-commerce labels are noisy or expensive to clean. In this work, we demonstrate the benefit for e-commerce of semi-supervised techniques like Semi-ViT, a ViT model fine-tuned with unlabeled data. We demonstrate that when Semi-ViT is compared to fully supervised ResNets and ViTs, it improves precision by at least +107bps and coverage by +311bps even when trained with 25% less labeled data.

1. Introduction

E-commerce companies have to deal with the quality of their product's attributes in their catalogues, a problem known as Item Data Quality (IDQ). These data or attributes can be anything from the neck style of a vest to the pattern of a cellular phone case. Human annotators usually work to provide product catalog attributes with clean values, but this is costly and complex as it requires classifying images into fine-grained classes.

Alternatively, supervised deep neural networks like ResNet [6] or Visual Transformer (ViT) [4] can be trained to learn to extract visual attribute values from catalog images [5]. However, these models require labeled images to learn. To alleviate labeled data requirements, transfer learning can be applied. Nevertheless, this approach only exploits labeled data, whereas unlabeled samples are not used for learning, despite they being readily available in e-commerce catalogues. Semi-supervised learning (SSL) [30] has emerged as a powerful technique for leveraging unlabeled data to improve the performance of deep neural networks. CNN methods have significantly advanced the field [1, 2, 12, 20, 23], while ViT architectures have only recently demonstrated promising results with methods like SVFormer [24] or Semi-ViT [3]. Hence, the use of SSL in the context of e-commerce presents a unique opportunity.

We hypothesise that techniques like Semi-ViT can improve over supervised methods like ResNet and ViT on the task of extracting fine-grained product attributes from e-commerce images. To evaluate our hypothesis, we collect five datasets from a known e-commerce website for various attribute extraction problems related to IDQ. The datasets contain unlabeled images as well as labeled images which have been annotated using Amazon Mechanical Turk. Our main contribution is, therefore, an analysis of the performance of Semi-ViT on attribute extraction for e-commerce compared to typically employed supervised models.

2. Related Work

Visual Transformers (ViT) have recently achieved state-of-the-art performance in many computer vision tasks [4, 14, 22]. A comprehensive review on ViTs can be found in the work of Khan *et al*. [8]. In this work, we apply a model based on ViT to the task of extracting visual attributes based on e-commerce images.

Transfer Learning leverages pre-trained models and adapts them to new domains [10, 17, 19, 31]. Yosinski *et al*. [26, 27] investigate the transferability of features learned by neural networks on different tasks, demonstrating their effectiveness. We perform transfer learning on each evaluated architecture using models pre-trained on ImageNet.

Semi-Supervised Learning (SSL) uses labeled and unlabeled data to improve model performance when labeled data are scarce [11, 25, 28]. SSL leverages intelligent data augmentation techniques paired with consistency regularization to improve performance [1, 15, 23, 29]. Other approaches rely on pseudo-labelling [13], teacher-student models [18, 21], ensembles [12], or adversarial training [9, 16]. In this work, we demonstrate the applicability of SSL with Semi-ViT [3] in the context of extracting attribute values for e-commerce products.

* Joint first authors.

Related works in the literature have showcased advances of ViT and SSL in image classification benchmarks like ImageNet. We demonstrate in this paper their applicability to a more challenging industrial scenario: classifying e-commerce products by fine-grained attribute values to deal with IDQ. To this end, we showcase that Semi-ViT is efficient at reducing the need for labeled data while increasing performance when compared to supervised models.

3. Data Collection

To create our IDQ benchmarks, we sampled product images from an e-commerce catalog from 21 marketplaces worldwide to create 5 datasets related to fine-grained attribute extraction: 3 basic sets that target one attribute for 1 or 2 types of products, and 2 complex sets that target one attribute on 9+ product types and nearly 30 classes each.

1. *Vest Neck Style* is a dataset with annotated 'neck style' on 'vest' products.
2. *Cellular Phone Case Pattern* is a dataset with annotated 'pattern' on 'cellular phone case' products.
3. *Apron Food Bib Pattern* is a dataset with annotated 'pattern' on 'apron' and 'food bib' products.
4. *Fashion Pattern* is a dataset with annotated 'pattern' on 9 product types such as 'hat' and 'scarf'.
5. *Home Linen Pattern* is a dataset with annotated 'pattern' on 12 product types such as 'rug' and 'curtain'.

The five datasets were initially considered as unlabeled. We consulted with e-commerce experts to define a list of valid values for each attribute and product type. We used then these values as classes to label a subset of the samples using Amazon Mechanical Turk, creating as a result labeled datasets for fine-tuning models. Table 1 shows a summary of the datasets statistics.

Note that the collected datasets include a proportion of unrelated products, like skirts in the *Vest Neck Style* dataset. These products were included because models working on IDQ need to be robust to misclassified products in the catalogue. We created a class named *other* to label images that do not belong to the aforementioned product type. Therefore, models need to learn to classify the attribute values and implicitly distinguish the relevant product type. We show in Figure 1 an example of images collected for each task.

DATASET SUMMARY

Basic Sets	Labeled	Unlabeled	Classes
Vest Neck Style	34K	227K	13
Cellular Phone Case Pattern	37K	287K	21
Apron Food Bib Pattern	39K	284K	26
Complex Sets	**Labeled**	**Unlabeled**	**Classes**
Fashion Pattern	157K	939K	29
Home Linen Pattern	106K	1.11M	27

Table 1. Number of images that are labeled, unlabeled, and number of classes for each of the datasets that we collected.

Figure 1. Examples of the images collected for our datasets: (top row) basic sets, (bottom row) complex sets.

The labeled data sets are split into *train*, *validation*, and *test* sets. The underlying class distribution is unknown, so the *train* and *validation* sets are not balanced. However, business use cases give equal importance to each class. Therefore, we balanced as much as possible the *test* set to have a similar number of images for all classes. As a result, the distribution of the labeled data is roughly 75%, 15%, and 10% for *train*, *validation*, and *test* respectively.

4. Methodology

The first supervised model tested is ResNet [6]. We use ResNet152 pre-trained on ImageNet and then fine-tuned on our benchmark datasets. We chose this version because it is the largest available (58M parameters), so it is comparable in size with the rest of the models tested. The second model is ViT [4]. For the architecture, we have relied on Masked Autoencoders (MAE) [7] ViT-Base model pre-trained on ImageNet, which is the most similar in size to ResNet152 (86M parameters). We do not compare ResNet and ViT models in depth since that has been already done in the literature [3].

The semi-supervised model is Semi-ViT [3]. This is the same ViT model described in the previous paragraph, but it is further fine-tuned using unlabeled samples. In this semi-supervised stage, an exponential moving average (EMA)-Teacher framework is adopted together with a probabilistic pseudo mixup method [29]. We chose this SSL method because it was the state-of-the-art SSL-based image classifier on ImageNet at the time of our work.

We use the following metrics to compare models:

- Precision: it is the percentage of images that have been correctly classified according to their attribute over the total number of images that we processed – the images in the test set. The higher, the better.

- Loss: it is a cross entropy loss which takes into account the predicted confidence of a class. The lower, the better. Ideally, we want models with lower loss because these are models that have higher confidence when they might be correct and lower confidence when they might be wrong.
- Coverage: in our real business problem, models automatically contributing to improve IDQ need to guarantee a minimum precision per predicted class (90%). To do so, we compute a confidence threshold per class so that predictions over this threshold are 90% correct. As a result, coverage measures the percentage of images in the test set whose prediction confidence is above the confidence threshold. The higher the coverage, the better because it means the model can automatically fix more products.

5. Results

In this section we show the results that ResNet, ViT and Semi-ViT yield using a different amount of the available training data, and using different amounts of unlabeled data. We report models performance on the test splits.

5.1. Baseline with 100% of labeled data

We first compare ResNet, ViT and Semi-ViT on the basic sets when trained with 100% of labeled data. For Semi-ViT, we used 100% of the unlabeled data. Results are reported at Table 2. Semi-ViT consistently obtains the best performance when compared to both supervised models (ResNet and ViT): it improves precision and coverage by at least +129bps (basis points, 0.01% difference), peaking at +405bps in precision and +511bps in coverage. It is remarkable that on *Apron Food Bib Pattern*, ResNet yields better metrics than ViT. However, Semi-ViT beats ResNet. This shows that fine-tuning the ViT model with unlabeled samples significantly boosts its performance.

5.2. Influence of the amount of labeled data

In this experiment we evaluate the performance difference of models when labeled data are scarce. To this end, we trained ViT and Semi-ViT using 75%, 50% and 25% of the available labeled data. For Semi-ViT, we used 100% of the unlabeled data as in the previous experiment. Results are shown in Table 2.

- *Vest Neck Style*: Semi-ViT trained with 75% of labeled samples increases precision by +150bps when compared to ResNet, and coverage increases by +96bps.
- *Cellular Phone Case Pattern*: Semi-ViT trained with 75% of labeled samples increases precision by +70bps when compared to ResNet, and coverage increases by +344bps.
- *Apron Food Bib*: Semi-ViT trained with 75% of labeled samples increases precision by +102bps when compared to ResNet, and coverage increases by +182bps.

Semi-ViT trained with 75% of the labeled samples outperforms the baseline ResNet models trained with 100% of labeled samples. According to these results, we can reduce labeled data requirements by 25%. This has a positive impact in an industrial scenario as it reduces the development costs. For example, if we had planned to label 40K samples, we can label 30K samples and develop a Semi-ViT model instead. Those 10K less samples could reduce expenses by 900$ on Amazon Mechanical Turk (0.09$/image).

VEST NECK STYLE				
Model	Labeled	Precision	Loss	Cov@90P
ResNet152	100%	81.62	0.56	71.44
ViT-Base	25%	76.34	0.81	60.88
ViT-Base	50%	80.58	0.70	65.22
ViT-Base	75%	81.05	0.66	67.10
ViT-Base	100%	**81.24**	**0.63**	**71.25**
Semi-ViT	25%	81.24	0.67	63.33
Semi-ViT	50%	81.43	0.63	70.87
Semi-ViT	75%	83.12	0.57	71.53
Semi-ViT	100%	**85.29**	**0.54**	**74.64**

CELLULAR PHONE CASE PATTERN				
Model	Labeled	Precision	Loss	Cov@90P
ResNet152	100%	79.24	0.66	61.59
ViT-Base	25%	74.17	0.87	51.93
ViT-Base	50%	77.57	0.75	61.40
ViT-Base	75%	77.81	0.73	60.25
ViT-Base	100%	**79.00**	**0.68**	**63.22**
Semi-ViT	25%	76.27	0.80	53.08
Semi-ViT	50%	79.14	0.70	64.61
Semi-ViT	75%	80.01	0.66	65.04
Semi-ViT	100%	**81.54**	**0.61**	**68.34**

APRON FOOD BIB PATTERN				
Model	Labeled	Precision	Loss	Cov@90P
ResNet152	100%	80.52	0.71	69.23
ViT-Base	25%	69.89	1.07	48.06
ViT-Base	50%	75.05	0.86	56.82
ViT-Base	75%	**78.30**	0.75	**65.05**
ViT-Base	100%	78.07	**0.73**	64.02
Semi-ViT	25%	73.49	0.98	56.06
Semi-ViT	50%	77.23	0.81	61.71
Semi-ViT	75%	81.54	0.68	71.05
Semi-ViT	100%	**81.81**	**0.66**	**71.40**

Table 2. Results using the ResNet, ViT, and Semi-ViT models with different data regimes on the three basic sets. Labeled is the percentage of labeled training data. Cov@90P is the coverage obtained when the model predictions are filtered based on per-class confidence thresholds to guarantee 90% precision per class.

APRON FOOD BIB PATTERN - 39K LABELED				
Model	Unlabel	Precision	Loss	Cov@90P
ViT-Base	N/A	78.07	0.73	64.02
ResNet152	N/A	80.52	0.71	69.23
Semi-ViT	284K	81.81	0.66	71.40
Semi-ViT	570K	82.08	0.65	**72.87**
Semi-ViT	855K	81.59	0.65	71.89
Semi-ViT	1.14M	**82.92**	**0.65**	72.20

Table 3. Results using Semi-ViT with different unlabeled data regimes on the *Apron Food Bib Pattern* benchmark. ViT and ResNet152 results are reported for reference. Cov@90P is the coverage when the model predictions are filtered based on per-class confidence thresholds to guarantee 90% precision per class.

FASHION PATTERN				
Model	L. Ratio	Precision	Loss	Cov@90P
ResNet152	100%	74.01	0.80	52.19
ViT-Base	100%	73.66	0.78	50.78
Semi-ViT	14.32%	**77.68**	**0.73**	**60.55**

HOME LINEN PATTERN				
Model	L. Ratio	Precision	Loss	Cov@90P
ResNet152	100%	80.08	0.63	71.95
ViT-Base	100%	80.03	0.60	73.10
Semi-ViT	8.66%	**83.47**	**0.57**	**79.56**

Table 4. Performance of the ResNet, ViT, and Semi-ViT models on the two complex sets. L. Ratio is the label/unlabel sample ratio. Cov@90P is the coverage at 90% precision.

5.3. Influence on the amount of unlabeled data

In this experiment, we investigate the impact of having additional unlabeled data. We added more unlabeled samples to *Apron Food Bib Pattern* dataset, growing from the original 284K unlabeled samples to a set of 1.14M samples. This benchmark dataset was chosen as it is the most complex of the basic sets (2 product types and 26 fine-grained classes). Results in Table 3 show the performance of Semi-ViT when trained on this dataset with different ratios of unlabeled data (25% - 284K, 50% - 570K, 75% - 855K, and 100% - 1.14M). We fixed the number of labeled samples to 39K as that was the amount of labeled data that yields the best performance in previous experiments.

There is little difference in metrics when we add more unlabeled samples to the training. The largest performance boost comes from the Semi-ViT trained with 284K unlabeled samples (+129bps in precision compared to ResNet152). From that point, precision increases by +111bps when we use 1.14M unlabeled samples. We conclude that increasing the amount of unlabeled data reaches a point in which it is no longer beneficial. In our basic sets, the ratio of labeled/unlabeled samples was above 1%.

5.4. Performance on complex datasets

We learnt from previous experiments on basic datasets that to get the most performance we need to: i) use as many labeled data we have available, ii) keep the ratio of label/unlabeled samples for Semi-ViT above 1%. In this section, we evaluated if these insights still apply on more complex datasets. For this purpose, we used the *Fashion Pattern* and *Home Linen Pattern* datasets: they target almost 30 fine-grained classes each, and they contain data for 9-12 different products.

From Table 4 we observe that Semi-ViT still yields the best performance on complex sets. On the *Fashion Pattern* dataset, Semi-ViT improves precision and coverage by +367bps and +816bps respectively when compared to ResNet. On the *Home Linen Pattern*, the improvement on metrics is +339bps and +761bps. The improvement of Semi-ViT on these complex datasets is larger than that of the basic datasets in Section 5.1: on those, the average improvement in precision and coverage was +242bps and +350bps. This shows Semi-ViT is well suited to learn these complex tasks. We argue this is due to the capacity of Semi-ViT to learn from unlabeled samples, which in our business use case are largely available (*e.g.* we collected 1M unlabeled samples for each complex dataset).

6. Conclusion

In this work we demonstrate that Semi-ViT, a ViT model fine-tuned on unlabeled data with SSL, yields high performance and outperforms supervised models like ViT and ResNet. In detail, we compared these models on datasets related to the classification of product images into fine-grained attribute classes, like the neck style of a vest product. We have proved that Semi-ViT improves precision at least by +107bps and coverage by +311bps even when trained with 25% less labeled data than a ResNet model.

We have observed that the labeled data should represent at least 1% of the whole dataset so that the model can learn the task and take advantage of the unlabeled data. Our next steps are to evaluate the throughput of Semi-ViT to check whether the increment on model complexity leads to a system that processes significantly less images per second than the baseline ResNet models. However, cloud services like Amazon Web Services have tools to automatically scale models, so we expect that this potential limitation can be mitigated using such tools.

We are also interested to adapt Semi-ViT to learn the multi-class problem, so we can train a single model for multiple products and attributes.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 1

[2] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021. 1

[3] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *arXiv preprint arXiv:2208.05688*, 2022. 1, 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[5] Patricia Gutierrez, Pierre-Antoine Sondag, Petar Butkovic, Mauro Lacy, Jordi Berges, Felipe Bertrand, and Arne Knudson. Deep learning for automated tagging of fashion images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[8] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 1

[9] Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017. 1

[10] Manuel Lagunas and Elena Garces. Transfer learning for illustration classification. *arXiv preprint arXiv:1806.02682*, 2018. 1

[11] Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. A similarity measure for material appearance. *ACM Transactions on Graphics (TOG, Proc. SIGGRAPH)*, 2019. 1

[12] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1

[13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 1

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[15] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 1

[16] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[17] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 1

[18] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 1

[19] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. Deep transfer learning for art classification problems. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[20] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1

[21] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[22] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021. 1

[23] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 1

[24] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. *arXiv preprint arXiv:2211.13222*, 2022. 1

[25] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[26] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 1

[27] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 1

[28] Brayan S Zapata-Impata and Pablo Gil. Prediction of tactile perception from vision on deformable objects. In *Workshop on Robotic Manipulation of Deformable Objects (ROMADO) in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2020)*, 2020. 1

[29] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1, 2

[30] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005. 1

[31] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 1