Enhancing Product Image Classification Performance with Swin Transformer: A Study on Small and Noisy Data with Minimal Preprocessing

(Gregoris Georgiou, Kyriakos Kyriakou, Alessandro Speggiorin)

Abstract

With the increasing popularity of E-commerce platforms, product image classification has become a prevalent approach for online retailers to enhance customer experience and reduce costs. However, product classification is particularly challenging due to the evolving nature of product categories, the visual similarity between products, and the varying lighting and camera angle conditions of user-uploaded images. Therefore, online E-commerce retailers often resort to an expensive and time-consuming manual annotation process to categorize product images. Thus, this study aims to address these challenges by investigating how pre-trained Shifted Window Transformers (Swin) can be leveraged to accurately predict product image categories when fine-tuned on limited data. More precisely, in this paper, we aim to improve Swin Transformers' generalization capabilities when fine-tuned on a small training subset of product images extracted from the Product-10k dataset and in conjunctions with various training approaches, including data augmentation, L2 regularization, label smoothing, and Sharpness-Aware Minimization (SAM) optimizer. Our findings highlight that our proposed approach improves the models' overall performance and achieves a total accuracy of 89.3%. Therefore, this study provides a resource-effective solution for product image classification and offers practical insights that can benefit online retailers seeking to streamline their product categorization processes.

1. Introduction

In the last recent years, the use of Deep Learning methodologies has become more and more prominent for computer vision tasks varying from object, and depth detection to image classification and semantic segmentation (Voulodimos et al., 2018). In this context, product image classification, defined as the categorization of products based on a set of predefined classes or taxonomy (Zahavy et al., 2016), has become a prevalent approach adopted by corporations such as Walmart and Amazon to reduce costs and enhance the customers' experience (Wei et al., 2020). More precisely, with the increased popularity of E-commerce platforms, product search has become an essential feature in order to allow customers to explore the products catalogue as well as search for related products by simply uploading a picture directly from their smartphone (Zahavy et al., 2016; Dagan et al., 2021). In light of this, several approaches involving the use of deep neural networks, such as VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016) and EfficientNet (Tan & Le, 2019), have been proposed over the years to tackle product image classification (Bai et al., 2020). More specifically, in a comparative study conducted by Mascarenhas & Agarwal, ResNet50 networks outperformed VGG16 and VGG19, achieving high classification accuracy on a dataset of 6000 product images (Mascarenhas & Agarwal, 2021). By contrast, Chen et al. proposed a salient-sensitive CNN model which leverages categorical and hierarchical features to improve classification on a large-scale dataset (Chen et al., 2019). Similarly, Zahavy et al. introduced a multi-modal fusion architecture (CNN-VGG based) to exploit textual and image features for product classification on a product dataset collected by Walmart (Zahavy et al., 2016). Furthermore, in recent years, the use of Vision Transformers (ViT) has become predominantly popular due to their versatility, and state-of-the-art results on image classification tasks (Dosovitskiy et al., 2021). In this context, ViT have successfully been adopted for retail product recognition, automatic retail checkout, and retrieval (Shihab et al., 2022; Wang et al., 2022).

Nonetheless, product classification is particularly challenging for several reasons (Bai et al., 2020). Firstly, the nature of the data is constantly evolving as new products are commercialized daily (Sinha et al., 2022). Secondly, the visual similarity between products belonging to different classes makes it difficult to distinguish between items (Wei et al., 2020). Furthermore, product pictures uploaded by users on E-commerce websites, usually taken with portable devices, are often subject to different lighting and camera angle conditions (Bai et al., 2020). To address these shortcomings, E-commerce websites often resort to hiring external annotators to manually categorize products, which can be a costly and time-consuming solution (Zahavy et al., 2016). Therefore, it can be safely stated that online retailers would greatly benefit from a streamlined automatic product image classification pipeline aimed to address the mentioned limitations (Zahavy et al., 2016).

1.1. Research Goal

This paper aims to investigate how product image classification can be conducted when operating with relatively limited and noisy data and when leveraging state-of-the-art transformers models, which are known to suffer when insufficient data is provided (Lee et al., 2021). More precisely, we aim to directly address the inherent limitations of training deep networks with reduced data, such as overfitting and poor generalization capabilities with the goal of proposing a resource-effective solution for product image classification (Chandrarathne et al., 2020; Han et al., 2021). Therefore, in the context of this paper, we attempt to answer the following research question:

Is it possible to exploit the inherent robustness of pre-trained vision transformers to accurately classify noisy product images when limited data is available?

To answer the mentioned research question, we fine-tune a pre-trained (on ImageNet-1K (Version 1) (Deng et al., 2009)) Shifted Window Transformer (Swin) (Liu et al., 2021b) on a small training subset of 20000 product images extracted from the Product-10k dataset (Bai et al., 2020). We then explore how different training approaches, including data augmentation, L2 regularization, label smoothing and Sharpness-Aware Minimization (SAM) optimizer (Foret et al., 2021) affect the model's accuracy and generalization capabilities. Our findings highlight that pre-trained Swin networks can achieve improved performance when fine-tuned on limited and noisy product image data when data augmentation, regularization and optimization strategies are adopted in conjunction.

2. Data set and task

2.1. Product 10K Dataset

For this paper, we utilize a subset (28750 images) out of the Product-10k dataset containing 150000 images collected from JD.com (Bai et al., 2020). More precisely, the images in the dataset are hierarchically organized into 10 product macro-categories, spanning from food, healthcare and fashion, with an associated fine-grained label, for a total of 10000 labels, corresponding to frequently bought stock-keeping units (SKUs) (Bai et al., 2020). Moreover, the dataset, as shown in Table 1, contains both in-shop and customer photos. In-shop images are usually high-quality object-centric E-commerce photos, with a clear, standard background/lighting and no occlusions (Bai et al., 2020). By contrast, customer photos include pictures taken by users with personal devices. Therefore, the inclusion of customer photos, as shown in Table 1, results in a very challenging dataset due to the noise these photos introduce, such as complex backgrounds, colour distortion and various complicated settings of lighting and viewing angles. It is also worth noting that the images of the products were manually labelled by an expert team of JD.com production and passed through further examination by at least three independent experts, resulting in a low noise rate of 0.5% (Bai et al., 2020).

For the purpose of our task, we consider only the 10 macro-classes, namely Top, Bottom, Accessories (Acc), Toy, Footwear, Case & Bag, Digital, Food & Drink, Home Decoration and Appliances (Home) and Pharmaceuticals and Personal Care Products (PPCPs). Furthermore, as the aim of this paper is to explore product image classification when limited noisy data is available, we reduce the dataset to contain approximately 2875 photos per class for a total of 28750 images.

2.2. Data Preprocessing

Image classification models such as Convolutional Neural Networks and Vision transformers require the size of each image to be the same for all input data. As Talebi & Milanfar (Talebi & Milanfar, 2021) suggest, in order for the models to be efficient, the input images can be resized in a small spatial resolution of 224x224x3, where 224 is the height and width of each image in pixels, and 3 is the number of channels of the image (RGB). Thus the first preprocessing step applied to the dataset was to resize every image in those dimensions. Furthermore, the pixel values were scaled from range [0,225] to range [0,1] and standardized to have a mean of 0 and standard deviation of 1. The standardization was done by calculating the mean and standard deviation of the training data pixels for each channel, and then subtracting the corresponding mean and dividing by the corresponding standard deviation for all data pixel values for each channel. It has to be mentioned that the calculation of the mean and standard deviation was done only using the training data, ensuring a consistent pixel distribution and enabling a fair comparison of image features while preventing any leak of validation or testing data during training.

2.3. Product Classification Task

As introduced in Section 1.1, in this paper, we aim to leverage Swin vision transformers with the goal of improving product image classification on a small and noisy dataset comprising 20000 training in-shop and customer product images. As aforementioned, the task of product classification is challenging for several reasons, including the overlap and similarity of items belonging to different categories (i.e. top and bottom in Table 1) (Wei et al., 2020), as well as the constantly evolving nature of product due to seasonal trends and customer preferences (Sinha et al., 2022). Furthermore, the lack of annotated data highlights the need for an automatic product image classification process (Zahavy et al., 2016). However, due to the limited size of our dataset, the performance of large deep neural networks might be degraded due to their tendency to overfit (Liu & Deng, 2015). In this context, Figure 4 presents the training and validation accuracy of a pre-trained Swin-T (on ImageNet-1K (Version 1)) transformer model fine-tuned on the subset of the Product 10k dataset presented in Section 2.1. To be more specific, it is clear from Figure 4 (a) that the model overfits after the third epoch. More precisely, while the

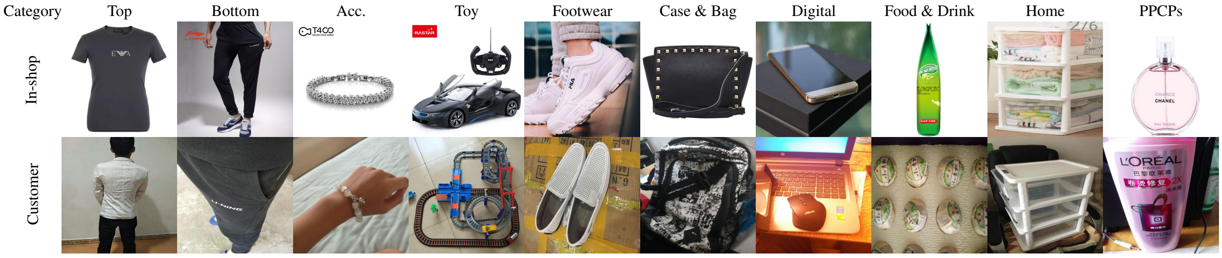
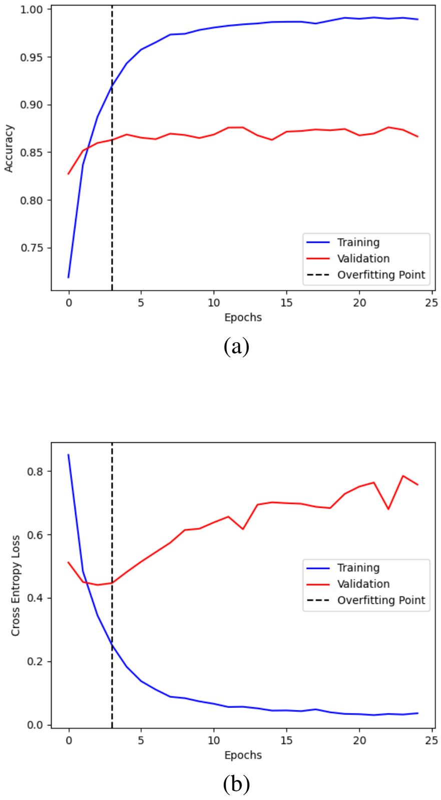


Table 1. Product-10k Products Macro-classes showing examples of *In-Shop* and *Customer* pictures belonging to each class.



(a)

(b)

Figure 1. Swin-T Vision Transformer Training and Validation accuracy (Figure a) and loss (Figure b) over 25 epochs.

training accuracy steadily increases until convergence, the validation accuracy rapidly stabilizes after the third epoch. Similarly, as shown in Figure 4 (b), the validation loss drastically increases after the third epoch. This suggests that the Swin-T model is overfitting the training data, failing to learn any meaningful pattern, and ultimately struggling to generalize well on unseen data. In this context, the model's behaviour and tendency to overfit could be explained due to the limited training data and the network's high parameter count (29 million).

Therefore, our goal in this paper is to investigate the extent to which the robustness of pre-trained Swin transformers can be exploited to improve their performance and generalisation capabilities. More precisely, our task is to explore how various strategies such as data augmentation, regularization and optimization algorithms help address overfitting and the inherent challenges associated with product image classification (more details in Sections 3 and 4).

2.4. Evaluation Metrics

We utilise standard classification metrics to evaluate models for the product image classification task addressed in this paper, including Accuracy, Precision, Recall and F1 Score. We briefly describe each metric below.

Accuracy (Acc): Accuracy is one of the most popular evaluation metrics adopted for classification tasks (Hossin & Sulaiman, 2015). However, as accuracy is known to produce a less discriminative and representative value (Huang & Ling, 2007), we opted to consider additional metrics for a more comprehensive evaluation. In light of this, accuracy can be formally expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (true positives) and TN (true negatives) define the number of samples belonging to the positive/negative class being classified correctly while FP (false positives) and FN (false negatives) represent the number of samples of the positive/negative being misclassified (Huang & Ling, 2007). It is worth noting that the same notation will be applied also to the following metrics.

Precision: Precision provides the proportion of positive samples being classified correctly over the total number of positive samples (Hossin & Sulaiman, 2015).

$$Precision = \frac{TP}{TP + FP}$$

Recall: Similarly, recall provides the actual proportion of positive samples being classified correctly (Hossin & Sulaiman, 2015).

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: The F1 Score combines both precision and recall in a unique metric which provides a more compact evaluation of a model's performance. (Hossin & Sulaiman, 2015). The F1 score is formally expressed as follows:

$$F1\,Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Lastly, we aim to investigate the models' performance by exploring their generalisation gap as well as their per-class classification accuracy with confusion matrices.

3. Methodology

In this section, we present the approach we adopted to tackle the task introduced in Section 1. As the goal of this paper is to improve deep neural networks capabilities when operating with small noisy datasets, we explore how several techniques such as data augmentation, different optimizers, and regularization techniques, affect the model generalization capabilities.

3.1. Models

We used several baseline pre-trained models in our study, which include Swin Transformer Tiny (Swin-T) model, ResNet50, EfficientNetB5, and Base Visual Transformer. We chose these models as they have been widely used in the literature and can serve as strong baselines for our experiments. The decision to use EfficientNetB5 as one of the baseline models in our study was based on the evaluation of the Product-10k dataset by the authors in paper (Bai et al., 2020), who employed the EfficientNetB3 model for this purpose. Our rationale for selecting the EfficientNetB5 model was to achieve better comparability with the other baseline models in our study, as it is a larger and more complex model than EfficientNetB3. Furthermore, we chose ResNet50 as another baseline model in our study because it outperformed VGG-16 and VGG-19 in terms of accuracy for the problem of product classification in the study of Mascarenhas et al. (Mascarenhas & Agarwal, 2021), making it a better competitor to the other models. Finally, we selected Base ViT as it has shown excellent performance compared to state-of-the-art CNNs in tasks of image classification (Dosovitskiy et al., 2021) despite its huge size, as we wanted to highlight its weakness in generalizing well on small datasets (Lee et al., 2021). Table 2 displays the size of each model.

We chose Swin-T as the model to explore with new techniques to improve performance, as it has shown promising results in the literature, and is more computationally efficient (Liu et al., 2021b; 2022).

In more detail, Swin-T Tiny model introduced by the authors in (Liu et al., 2021b), is a small version of the Swin Transformer model, which has been shown to achieve state-of-the-art performance on various image classification tasks (Liu et al., 2021b). Swin-T is specifically designed for efficient computation and memory usage, making it suitable for deployment on resource-constrained devices (Liu et al., 2021b). Its architecture consists of a hierarchical grouping of patches, followed by multi-scale windows and non-local attention mechanisms (Liu et al., 2021b). The Swin Transformer architecture, as proposed by Liu et al. (Liu et al., 2021b), effectively captures both local contexts by allowing each pixel to attend to its neighboring pixels and global contexts by enabling information exchange across different parts of the image. This approach leads to a strong performance on image classification tasks. Additionally, these methods within the model also help to reduce overfitting.

3.2. Data Augmentation (DA)

In order to mitigate the problem of over-fitting when models are trained on small datasets, the model has to improve its generalization ability. By using a small dataset as a source, new transformed images can be generated and added to the dataset, improving the dataset quality in terms of diversity and the number of training samples. This technique is called Data Augmentation (DA). A variety of DA techniques can be used, including rotation, translation, adjustment of pixel distributions, colour jittering and others. Let et al. did a preliminary study on DA for deep learning for image classification, and they concluded that simple techniques such as translation, rotation and shearing can be much more beneficial than more complicated techniques (Lei et al., 2019). Thus it was decided to utilize only simple techniques, namely random horizontal flip, random rotation, random translation and random shearing. Random horizontal flip was used with probability 0.5 while random rotation was applied in the range [-30°,30°]. Moreover, random translation was allowed in both directions up to 67 pixels (0.3*224=image dimension) both horizontally and vertically and at last random shearing was applied in the range [-0.3,0.3] both horizontally and vertically.

3.3. Optimizers

To enhance the ability of the models to generalize well on small noisy datasets in the context of product classification, we incorporated two optimization techniques: ADAM optimizer and SAM optimizer.

The ADAM optimizer is a popular stochastic gradient descent optimization algorithm that calculates adaptive learning rates for each parameter of the model based on the first and second moments of the gradients (Kingma & Ba, 2017). The algorithm has been shown to converge faster (computationally efficient) and more reliably than other optimization techniques, such as stochastic gradient descent (SGD) (Kingma & Ba, 2017).

The Sharpness-Aware Minimization (SAM) optimizer was proposed in the study by Foret et al. (Foret et al., 2021). SAM is an optimization algorithm that focuses on improving generalization performance by minimizing a sharpness-aware objective function (Foret et al., 2021). This optimizer modifies the gradient update rule of the optimizer (computes two times the gradients) to consider the sharpness of the loss landscape, which can help to avoid local minima and improve the generalization of the model (Andriushchenko & Flammarion, 2022; Foret et al., 2021). The SAM optimizer has shown to achieve state-of-the-art performance on several image classification tasks, particularly on small datasets where overfitting and bad generalization are common problems (Foret et al., 2021). According to

Foret et al., incorporating sharpness-aware updates, leads to a broader minimum as shown in Figure 2, resulting in improved generalization properties (Foret et al., 2021).
While SAM is a powerful optimizer, it is designed to work in conjunction with a base optimizer such as SGD (Foret et al., 2021). The base optimizer is responsible for computing the gradients and updating the parameters of the model, while the SAM modifies the update rule to improve generalization. Moreover, while SAM has shown state-of-the-art performance on several image classification tasks, AdaSAM optimizer (Sun et al., 2023), which integrates SAM with adaptive learning rate and momentum acceleration, has also been explored to train large-scale deep neural networks. In our paper, we propose SAM optimizer that uses as a base ADAM optimizer. By using ADAM as the base optimizer, we can benefit from its fast convergence and adaptive learning rates (Kingma & Ba, 2017), while also improving the generalization ability of the model with SAM.
This combination of optimization techniques is expected to enhance the ability of the models to generalize well on relatively small noisy datasets in the context of product classification, and reduce the problem of overfitting. To the best of our knowledge, the exploration of Swin Transformers with SAM optimizer, and ADAM as a base optimizer in the area of of Product Classification is a novel study.

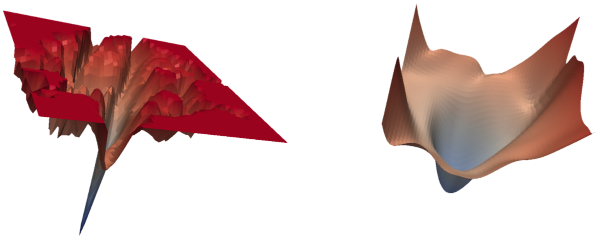


Figure 2. Comparison of loss landscapes obtained after training a model with and without Sharpness-Aware Minimization (SAM). The left plot shows a sharp minimum that the model trained with SGD optimizer converged to, while the right plot displays a wide minimum that the same model trained with SAM optimizer reached. Figure adapted from (Foret et al., 2021)

3.4. Regularization Techniques

In order to prevent overfitting on our limited noisy dataset (Product-10k), we employed two commonly used regularization techniques, namely label smoothing (LS) and weight decay L2 regularization. Label smoothing, introduced by the authors in (Szegedy et al., 2016), is a technique used to prevent the model from becoming overconfident in its predictions, which can lead to overfitting. It involves replacing the hard targets (image labels) with soft targets that are a smoothed version of the true labels (Szegedy et al., 2016). This encourages the model to be more robust and perform better on unseen data. Mathematically, label smoothing can be represented as:

$$\hat{y}_i = (1-\epsilon)y_i + \frac{\epsilon}{K} \tag{1}$$

where $\hat{y}_i$ is the soft target for the i^{th} class, y_i is the true label for the i^{th} class, ϵ is the smoothing parameter, and K is the number of classes.

Weight decay L2 regularization is another commonly used technique that helps tackle overfitting, by adding a penalty term to the loss function that encourages the weights to stay small. L2 regularization showed effective results in the study of (He et al., 2016) for deep complex networks, and in the study of (Krogh & Hertz, 1991) for small noisy datasets. Mathematically, weight decay L2 regularization can be represented as:

$$E_w = E_0(w) + \frac{\lambda}{2}\sum_i \|w_i\|^2 \tag{2}$$

where E_w is the regularization loss, $E_0(w)$ is an error measure (e.g. sum of squared erros), λ is the regularization strength, w is a vector containing all the networks' parameters (Krogh & Hertz, 1991). By using label smoothing and weight decay L2 regularization, we can help to regularize the model and prevent it from memorizing the training data, leading to better generalization performance on unseen data.
We used these regularization techniques in combination with the Swin-T transformer model to improve its generalization performance and prevent overfitting on our small dataset.

We analyze the different experiments in the section 4, and evaluate the effectiveness of the mentioned techniques in the 4.2 Results section of our study.

4. Experiments

4.1. Experimental Setup

In this study, **Swin-t** vision transformer served as the baseline model for subsequent experiments. Two additional baseline models, **ResNet-50** and **EfficientNet-B5**, were also introduced. These convolutional neural networks (CNNs) were selected as baselines because they are commonly used for image classification tasks and have comparable sizes to Swin-t. EfficientNet-B3 is the model of choice used by the authors of the Product 10k dataset (Bai et al., 2020), but EfficientNet-B5 was used as a baseline as it has a more comparable number of parameters with Swin-t. Finally, **ViT b/16**, a larger model with approximately 86 million parameters, was included as a baseline model despite its size as this study focuses on vision transformers, resulting in 4 different baseline models. All models were pre-trained using the ImageNet-1K (Version 1) dataset (Deng et al., 2009) and fine-tuned with our training dataset using the models' default settings. In this context, we

Method	#Params	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Resnent-50	23M	83.7	83.9	83.7	83.7
Efficientnet-b5	30M	82.0	82.0	82.0	82.0
ViT b/16	86M	81.7	82.7	81.7	81.8
Swin-t (Baseline)	29M	85.8	86.4	85.8	86.0
Swin-t + DA	29M	87.1 (+1.3%)	87.4 (+1.0%)	87.1 (+1.3%)	87.2 (+1.2%)
Swin-t + DA + L2	29M	86.6 (+0.8%)	86.9 (+0.5%)	86.6 (+0.8%)	86.7 (+0.7%)
Swin-t + DA + LS	29M	87.2 (+1.4%)	87.4 (+1.0%)	87.2 (+1.4%)	87.3 (+1.3%)
Swin-t + DA + L2 + SAM	29M	88.3 (+2.5%)	88.7 (+2.3%)	88.3 (+2.5%)	88.4 (+2.4%)
Swin-t + DA + LS + SAM	29M	88.7 (+2.9%)	88.9 (+2.5%)	88.7 (+2.9%)	88.7 (+2.7%)
Swin-t + DA + L2 + LS + SAM	29M	**89.3** (+3.5%)	**89.5** (+3.1%)	**89.3** (+3.5%)	**89.4** (+3.4%)

Table 2. Classification results for the models and conditions considered in the experiment and computed on the test set. Percentage increments are reported with respect to the Swin-T baseline.

decided to adopt pre-trained models as that is a standard approach when operating with small datasets and to minimize GPU usage (Liu et al., 2021a;b).

Furthermore, as a measure of mitigating over-fitting, all the experiments were run only for 10 epochs and with a batch size of 32 as large batch sizes might reduce models' generalization capabilities (Keskar et al., 2016). Additionally, ADAM optimizer was used for the experiments without SAM, with learning rate $1e^{-4}$, while the loss function used for all experiments was cross-entropy. For the application of L2 regularization, the weight decay coefficient was set to $1e^{-4}$. Note that we chose the aforementioned values for learning rate and weight decay coefficient as they were also used in the Product 10k paper (Bai et al., 2020) and are commonly used in general. In addition, ϵ value of 0.2 was used for label smoothing regularization, as it has been shown to be more effective in balancing the model (Szegedy et al., 2016). At last, ADAM optimizer was used as base optimizer for SAM optimizer, utilizing the aforementioned learning rate, while all other parameters were the default.

For the experiments conducted, the dataset was split in a stratified manner to ensure a balanced number of images for each category in each type of data. The dataset was split into training (70%), validation (15%) and test data (15%), resulting in about 20125 images for training and about 4313 images for each of validation and test data.

Lastly, when Data Augmentation is applied in an experiment, the techniques used are described in 3.2. DA was applied only to training data, and the new data generated were augmented back to the training data, resulting in a more extensive and more diverse training dataset of 40250 images. However, we kept the same validation and testing test for a meaningful comparison across models.

4.2. Results & Discussion

In this Section, we present and discuss the results obtained from our experiments by evaluating the performance of Swin-T in conjunction with the various configurations considered, as well as comparing its performance against the baselines. In light of this, it is clear from Table 2 that ResNet50, EfficientNet-B5, ViT B/16 and Swin-T provide strong baselines scoring more than 80.0 across all the metrics considered. However, ViT B/16 exhibits the lowest performance among the four baselines proposed. The reason for this could be attributed to the fact that large vision transformers are known to perform poorly when limited data is available (Dosovitskiy et al., 2021). Nonetheless, Swin-T provides the most robust baseline resulting in an accuracy of 85.6, highlighting the benefits of the shifted-window self-attention paradigm for image classification tasks.

In light of this, and as shown in Table 2, data augmentation provides consistent improvements across all the conditions considered. More specifically, Swint-T + DA leads to a gain of more than 1% across all the metrics considered and with respect to the Swin-T baseline. This suggests that applying standard data augmentation strategies, such as random rotation, translation, and flipping, can enrich the dataset leading models to generalise better. However, while data augmentation proves to be effective, it comes with the overhead of requiring longer training times which might not be suitable depending on the resources available. Similarly, it is clear that the addition of regularization techniques such as label smoothing (Swin-T + DA + LS) and L2 regularization (Swin-T + DA + L2) help the model to improve even further. This suggests that constraining a model's weights (L2 regularization) as well as adding perturbations to the target labels (label smoothing) reduces the model's confidence and tendency to overfit leading to better generalization capabilities. In this context, the introduction of SAM optimizer in our experiments drastically affected the model's performance. More precisely, while SAM in disjointed combination with L2 and LS led to improvements across all metrics, the best-performing model is found when all the configurations are considered simultaneously (Swin-T + DA + L2 + LS + SAM). As shown in Table 2 Swin-T + DA + L2 + LS + SAM leads to an improvement of more than 3% for all metrics and with respect to the baseline, achieving a final accuracy of 89.3. This suggests that affecting the smoothness of the loss landscape with SAM (Foret et al., 2021) as well as the path taken by the optimizer in such landscape with L2 regularization (Fort & Jastrzebski, 2019), help the model's learning process, leading to better

results and generalization.
Nonetheless, it is clear from Figure 3, that the model's performance is far from optimal. More precisely, it is clear that the model still struggles to distinguish between certain classes, such as Food & Drinks with PPCPs and Top with Bottom (and vice-versa). This can be explained due to the similarity between products belonging to those classes. For example, as shown in Table 1, a beverage bottle could be easily misclassified as a shampoo container due to their feature similarity in terms of shape and colour.

Lastly, it is clearly indicated by figure 4(b) that the use of the SAM optimizer, compared to the Swin-T baseline, reduces the rate of increase of the generalization gap along the training epochs where over-fitting occurs. On the other hand, 4(a) shows that Swin-T combined with only data augmentation and regularization techniques such as L2 and LS, fail this task, and their generalization gap lines fluctuate at similar levels as the Swin-T baseline. As a conclusion, the combination of SAM optimizer with the model, effectively oppress over-fitting, letting the model to generalize better, make improved predictions for unseen data and thus elevating the models overall performance.

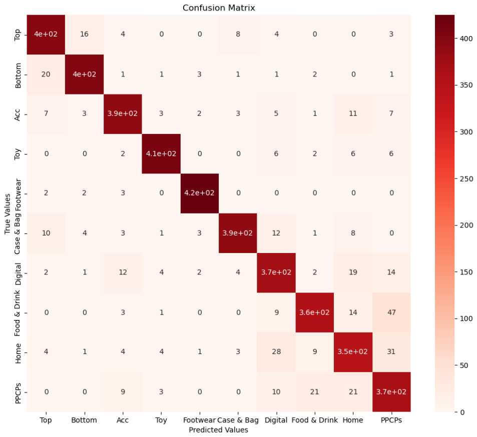
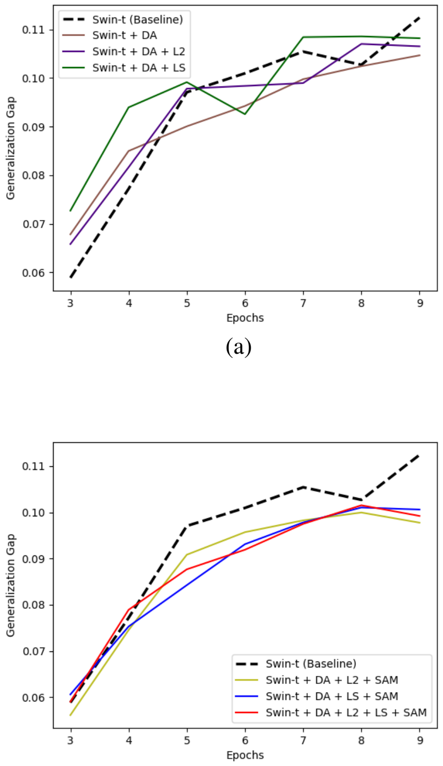


Figure 3. Confusion matrix computed for the best performing model (Swin-T + DA + L2 + LS + SAM) on the Test set.

5. Future work

The results of this study demonstrate that pre-trained vision transformers can be successfully fine-tuned on limited and noisy produce image datasets. However, there are still several directions for future research that can be studies to improve upon the current approach and address some of the limitations of the study.
One possible avenue for future work is to focus on fine-grained classification tasks, such as more specific product classification using the whole 10k classes of the Product 10k dataset (Bai et al., 2020) or other similar datasets such as RPC dataset that contains 200 different products (Wei et al., 2019).
Additionally, another interesting direction would be to explore multi-modal learning, which involves using data from different areas such as video, or text, in addition to images. In the paper by Zahavy et al., the authors demonstrate generalization improvement by fusing image and text on a large-scale classification dataset (Zahavy et al., 2016). Therefore, applying multi-modal technique could be a worth exploring.
Another potential area of future work is to investigate the effectiveness of transfer learning across different product domains. In this study, we focused on the 10 macro categories of the Product 10k dataset (Bai et al., 2020). However, it would be interesting to investigate whether the model can be fine-tuned on even smaller subsets of data from different product domains and still achieve good generalization. This could be particularly useful for smaller businesses that may not have large product datasets.
Finally, it might be beneficial to explore the use of additional data augmentation and preprocessing techniques to further improve the model's generalization capabilities. In this study, we used basic data augmentation techniques such as random horizontal flip, random rotation, translation and shearing. However, more advanced techniques such as CutMix (Yun et al., 2019), background complication processing for image denoising (Zhu, 2022), or MixUp (Zhang



(a)

(b)

Figure 4. Generalization gap comparison across all the Swin-T transformer conditions considered. More precisely, Figure (a) shows the generalization gap of Swin-T in combination with data augmentation, L2 regularization and label smoothing. Similarly, Figure b present results for the same conditions with the addition of SAM.

et al., 2018) may be worth investigating.
Overall, there are a number of intriguing directions for future study in the field of using Swin-T to classify product images.

6. Conclusions

In conclusion, this paper investigated how to conduct product image classification with limited and noisy data using state-of-the-art transformer models. The study aimed to propose a resource-effective solution for product image classification by addressing the limitations of training deep networks with reduced data, such as overfitting and poor generalization capabilities. The study fined-tuned a pre-trained Shifted Window Tiny Transformer (Liu et al., 2021b) on a small training subset of 20k product images extracted from the Product-10k dataset (Bai et al., 2020) and explored the impact of different training approaches on the model's accuracy and generalization capabilities. The finding suggests that pre-trained Swin networks can achieve improved performance when fine-tuned on limited and noisy product image data when adopting regularization and optimization strategies. To be more precise, the Swin-T model that utilizes data augmentation, label smoothing and L2 weight decay regularization, and the sharpness-aware minimization optimizer (Foret et al., 2021), with Adam as its base optimizer, demonstrated an accuracy increase of 3.5%, resulting in a total accuracy of 89.3%. Finally, these efforts can ultimately lead to more accurate and efficient product categorization processes, benefiting online retailers and customers alike. While this study provides promising results, further research in the area of product image classification with Swin transformers is necessary to continue advancing the accuracy and robustness of these models.

References

Andriushchenko, Maksym and Flammarion, Nicolas. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668. PMLR, 2022.

Bai, Yalong, Chen, Yuxiang, Yu, Wei, Wang, Linfang, and Zhang, Wei. Products-10k: A large-scale product recognition dataset. *CoRR*, abs/2008.10545, 2020. URL https://arxiv.org/abs/2008.10545.

Chandrarathne, Gayani, Thanikasalam, Kokul, and Pinidiyaarachchi, Amalka. A comprehensive study on deep image classification with small datasets. In *Advances in Electronics Engineering: Proceedings of the ICCEE 2019, Kuala Lumpur, Malaysia*, pp. 93–106. Springer, 2020.

Chen, Zhineng, Ai, Shanshan, and Jia, Caiyan. Structure-aware deep learning for product image classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–20, 2019.

Dagan, Arnon, Guy, Ido, and Novgorodov, Slava. An image is worth a thousand terms? analysis of visual e-commerce search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 102–112, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462950. URL https://doi.org/10.1145/3404835.3462950.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, and Houlsby, Neil. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Foret, Pierre, Kleiner, Ariel, Mobahi, Hossein, and Neyshabur, Behnam. Sharpness-aware minimization for efficiently improving generalization, 2021.

Fort, Stanislav and Jastrzebski, Stanislaw. Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32, 2019.

Han, Xu, Zhang, Zhengyan, Ding, Ning, Gu, Yuxian, Liu, Xiao, Huo, Yuqi, Qiu, Jiezhong, Yao, Yuan, Zhang, Ao, Zhang, Liang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

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

Hossin, Mohammad and Sulaiman, Md Nasir. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

Huang, Jin and Ling, Charles X. Constructing new and better evaluation measures for machine learning. In *IJCAI*, pp. 859–864, 2007.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, and Tang, Ping Tak Peter. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization, 2017.

Krogh, Anders and Hertz, John. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Lee, Seung Hoon, Lee, Seunghyun, and Song, Byung Cheol. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Lei, Cheng, Hu, Benlin, Wang, Dong, Zhang, Shu, and Chen, Zhenyu. A preliminary study on data augmentation of deep learning for image classification. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, pp. 1–6, 2019.

Liu, Shuying and Deng, Weihong. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, 2015. doi: 10.1109/ACPR.2015.7486599.

Liu, Yahui, Sangineto, Enver, Bi, Wei, Sebe, Nicu, Lepri, Bruno, and Nadai, Marco. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021a.

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

Liu, Ze, Hu, Han, Lin, Yutong, Yao, Zhuliang, Xie, Zhenda, Wei, Yixuan, Ning, Jia, Cao, Yue, Zhang, Zheng, Dong, Li, Wei, Furu, and Guo, Baining. Swin transformer v2: Scaling up capacity and resolution, 2022.

Mascarenhas, Sheldon and Agarwal, Mukul. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, volume 1, pp. 96–99. IEEE, 2021.

Shihab, Md, Hossain, Istiak, Tasnim, Nazia, Zunair, Hasib, Rupty, Labiba Kanij, and Mohammed, Nabeel. Vista: Vision transformer enhanced by u-net and image colorfulness frame filtration for automatic retail checkout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3183–3191, 2022.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sinha, Ankit, Banerjee, Soham, and Chattopadhyay, Pratik. An improved deep learning approach for product recognition on racks in retail stores. *arXiv preprint arXiv:2202.13081*, 2022.

Sun, Hao, Shen, Li, Zhong, Qihuang, Ding, Liang, Chen, Shixiang, Sun, Jingwei, Li, Jing, Sun, Guangzhong, and Tao, Dacheng. Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *arXiv preprint arXiv:2303.00565*, 2023.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Talebi, Hossein and Milanfar, Peyman. Learning to resize images for computer vision tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 497–506, October 2021.

Tan, Mingxing and Le, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Voulodimos, Athanasios, Doulamis, Nikolaos, Doulamis, Anastasios, Protopapadakis, Eftychios, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

Wang, Wenhao, Sun, Yifan, Yang, Zongxin, and Yang, Yi. VΘ2 l: Leveraging vision and vision-language models into large-scale product retrieval. *arXiv preprint arXiv:2207.12994*, 2022.

Wei, Xiu-Shen, Cui, Quan, Yang, Lei, Wang, Peng, and Liu, Lingqiao. Rpc: A large-scale retail product checkout dataset, 2019.

Wei, Yuchen, Tran, Son, Xu, Shuxiang, Kang, Byeong, Springer, Matthew, et al. Deep learning for retail product recognition: Challenges and techniques. *Computational intelligence and neuroscience*, 2020, 2020.

Yun, Sangdoo, Han, Dongyoon, Oh, Seong Joon, Chun, Sanghyuk, Choe, Junsuk, and Yoo, Youngjoon. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.

Zahavy, Tom, Magnani, Alessandro, Krishnan, Abhinandan, and Mannor, Shie. Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce. *CoRR*, abs/1611.09534, 2016. URL http://arxiv.org/abs/1611.09534.

Zhang, Hongyi, Cisse, Moustapha, Dauphin, Yann N., and Lopez-Paz, David. mixup: Beyond empirical risk minimization, 2018.

Zhu, Boda. Retail commodity image recognition based on ws-dan. In *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, pp. 310–316, 2022. doi: 10.1109/MLISE57402.2022.00068.