Expertblog

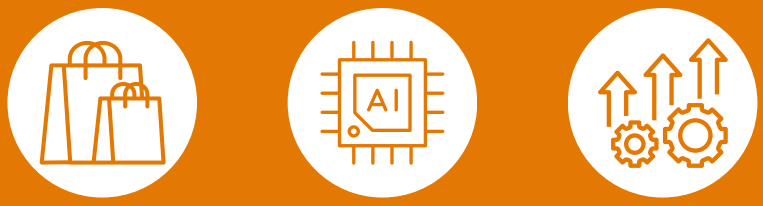
Fine-tuning CLIP for automatic extraction of product features from images

Written by
Britt Deckers

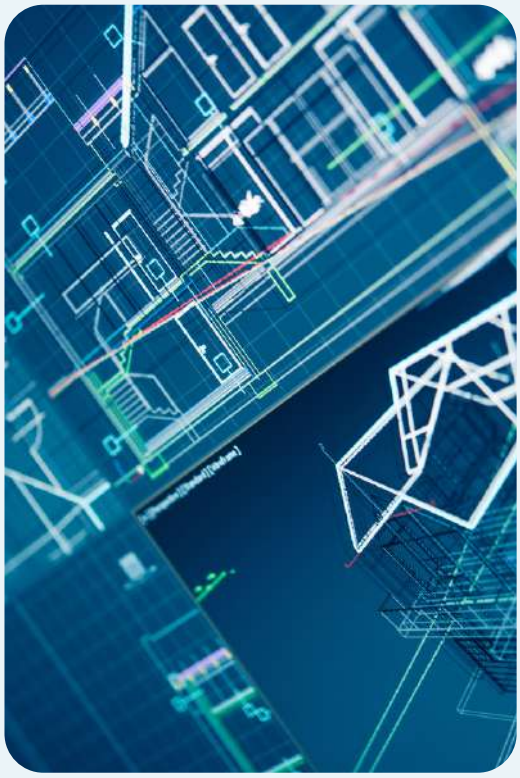
10-03-2023

Introduction

Over the last 10 years, the amount of data in the retail industry has grown significantly. Retailers are now collecting a wide range of data points about the products they sell to improve customer information, enhance data quality and target products to buyers. However, the quality and quantity of this product data can vary greatly between businesses, with some companies having high-quality and ample data while others have significant gaps. For instance, some retailers may only have product images and no product feature list (e.g. color = red). Extracting these features can be useful for various purposes such as data enrichment, data quality issue detection and data quality improvement. In recent years, there have been significant advancements in language and vision models. These multimodal models, such as OpenAI's CLIP (Radford et al., 2021), can bridge the gap between text and images by identifying similarities between the textual and visual spaces. In this article, a method for fine-tuning CLIP for automatic extraction of product features from images will be described.



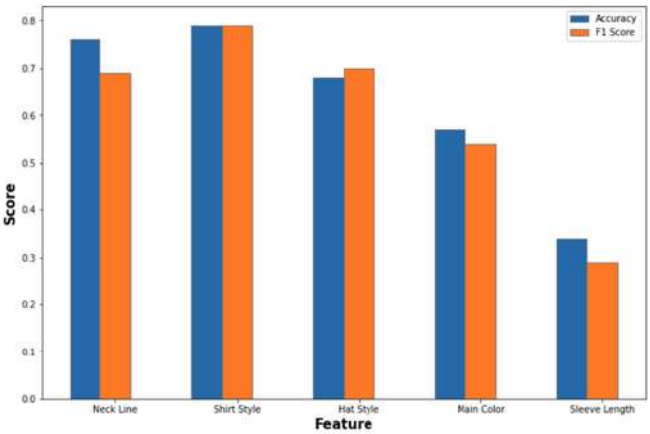
Introduction to CLIP



CLIP (Contrastive Language-Image Pre-training) is a model that has been developed to connect language to vision. It is a large multi-modal Neural network that has been trained on a wide variety of images that are labeled with natural language in order to understand the relationship between natural language and visual objects. CLIP is trained using a process called contrastive learning. In this process, the model is presented with a set of positive and negative examples, and it learns to distinguish between the two. In the case of CLIP, the positive examples consist of pairs of images and text captions that describe the images, while the negative examples consist of pairs of images and text captions that do not match. During training, the model is presented with an image and a text caption, and it must determine whether or not the text caption accurately describes the image. The model is trained to maximize the similarity between the image and text caption when they match and minimize the similarity when they do not match. CLIP combines the strengths of both image recognition models and language models and is capable of generalizing to novel objects and scenes. CLIP can be used by formulating natural language prompts to perform a wide range of tasks, such as image captioning, object detection and visual question answering.

CLIP for extracting product features from images

For the task at hand, extracting product features from images, an image is presented to the model with a list of possible captions, e.g., ["A photo of a long-sleeved top", "A photo of a short-sleeved top", "A photo of a sleeveless top"], and the model outputs a list of probabilities for every given caption. It is important that the prompts are formulated in a similar manner that captions within the training data set are. Zhou et al. (2022) show that it is important to add context to the prompt, which is why extra information is added to the prompt, and not only the feature. They show that tuning words to a specific domain can have a huge impact on the performance of these models. In order to test the performance of standard CLIP for image feature extraction, a data set
was used that contains images of clothing with their features, such as color, sleeve length, neckline, etc. CLIP was implemented and tested on multiple features of the data



Fine-tuning CLIP

To improve CLIP's performance on the extraction of product features, we fine-tuned CLIP for the domain of product images. In order to fine-tune CLIP, multiple tests were done on the aforementioned clothing data set in order to perfect the training loop. First, the data were split into train, validation, and test sets, and a balanced batch sampler was implemented to make sure the different feature values are balanced for every batch to prevent bias after training. A cross-entropy loss was used, and the model with the smallest loss, calculated on the validation set during training, out of all epochs is saved as the best model. The trained model was then tested on the test set to see whether fine-tuning CLIP actually improves the scores of the model. The figures show the amount of training data needed to fine-tune CLIP to reach decent accuracy and F1 scores. For the clothing data set, the figures show that a lot of training data is needed to reach an accuracy score of ~0.7. This seems somewhat inefficient, as you would thus need a lot of labeled data to begin with. As seeing certain features on clothing items might even be ambiguous to the human eye, e.g., there are often multiple colors, which makes it harder to decide on a main color, it is hypothesized that CLIP needs less training data when the features are more clearly visible. This hypothesis is tested on a data set containing pots and pans, where the desired features were whether the pans have a lid or not and whether they have a handle or not. The results are shown in the figure 1. For fine-tuning CLIP on these features, significantly less training data were needed, with ~60 training data points, an accuracy and F1 score of ~1 could be reached.

As mentioned before, the formulation of the prompts is important for gaining an accurate outcome from models such as CLIP. Therefore, it is hypothesized that this it would also give better results for training CLIP on natural language captions generated around the feature labels. However, after running tests on training CLIP with multiple forms of captions that include the feature, it was concluded that the results for fine-tuning are better when the captions only consist of the desired feature. So instead of training an image with the text "A photo of a top with long sleeves", the image was trained on "Longsleeved".

Figure 1

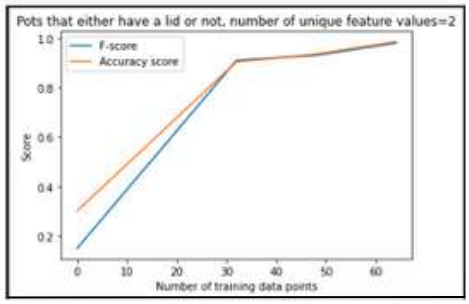
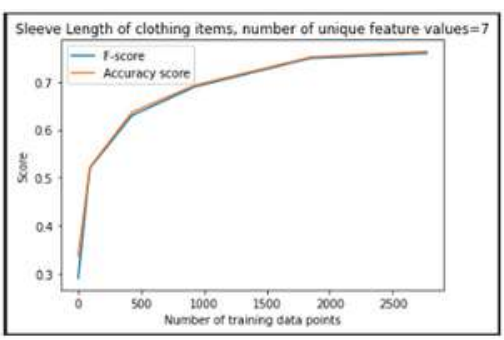


Figure 2



Conclusion

In conclusion, OpenAI's CLIP can be leveraged for automatically extracting product features from images. However, for many categories of products and features, this task is too domain-specific, and CLIP needs to be fine-tuned in order to gain reasonable results. Therefore, this article provides a guidance on how CLIP can be fine-tuned for extracting product features from images. For this, you will need a set of labeled images, the size of this set depends on how visible the features are on your images. The more clearly visible, the less training data is needed.

Contact Squadra MLC