VisionFashion: Multi-Modal Style Embedding Learning with Vision Transformers and BERT for Fashion Image Analysis and Recommendation

Tuğcan Topaloğlu
Yıldız Teknik Üniversitesi
Istanbul, Turkey
tugcan.topaloglu@std.yildiz.edu.tr

***Abstract*—In 2016, a fashion recognition model paper is published in CVPR. The dataset contained over 800,000 images, which was richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. In today's e-market and shopping world, this method remains outdated. So in this paper, a modern and more robust model is created with Vision Transformers (ViT) for image feature extraction and a BERT model for processing textual descriptions from the DeepFashion-MultiModal dataset. A method was created similar to the CLIP method. Embeddings are evaluated on image-to-text and text-to-image retrieval tasks. Classification heads are added to the ViT image encoder and fine-tuned to predict fashion categories (based on shape) and attributes (based on fabric and pattern). The fine-tuned classification tasks yielded strong results, with Top-1 Category (Shape) Accuracy reaching 0.9470 and Average Recall@5 for Attributes (Fabric+Pattern) achieving 0.7291, outperforming or being highly competitive with the reported results in the original paper, which was created in 2016. R@10 scores were approximately 0.55 for both the image-to-text and text-to-image directions. The results of this paper show new methods, like ViT performance of image retrieval, relatively better than older methods.**

***Index Terms*—Deep Learning, Multi-Modal Learning, Vision Transformer, BERT, Contrastive Learning, Fashion Recommendation, Style Embedding, Image Retrieval, Attribute Prediction**

I. Introduction

Due to increasing demand in marketing and e-commerce, understanding and retrieving fashion designs is very important. Detecting and classifying fashion products is a great challenge for SEO and marketing. Traditional methods remain limited because of the increasing product numbers. This model aims to develop a multi-modal system capable of learning a shared embedding space for fashion images and their corresponding textual captions. By using Vision Transformers (ViT) for visual understanding and BERT for natural language processing, this models employ a contrastive learning strategy to align these modalities.

Furthermore, to contextualize this model's visual understanding capabilities, this model extends its application to fashion category and attribute prediction, drawing parallels with the benchmark tasks defined in the seminal DeepFashion paper [1]. This allows for a partial comparison against established results, highlighting the advancements brought by modern architectures and learning paradigms.

This paper details the methodology, including all aspects of data processing, model architecture, and the two training processes (contrastive learning and then classification fine-tuning). First, our model will be trained with ViT, and then it will be fine-tuned for classes. After that, our BERT model is trained with that model too. We are going to find our final model with image-to-text and text-to-image capabilities.

II. Related Work

Today's fashion analysis field is mostly dominated by deep learning models. Early works focused on CNNs for tasks like category classification and attribute recognition [1]. Introduced the large-scale DeepFashion dataset and proposed CCN for various fashion-related tasks.

Multi-modal learning, especially for vision and language, has seen remarkable progress with models like CLIP [2] and ALIGN [3], which learn rich joint embeddings through contrastive learning on massive image-text datasets. These models have demonstrated impressive zero-shot capabilities. In our study, Vision Transformers (ViT) [4] have emerged as a powerful alternative to CNNs for image recognition, capturing global context effectively. BERT [5] remains a standard for various NLP tasks due to its deep bidirectional representations.

My model draws inspiration from these advancements with ViT and BERT. Applying a ViT-BERT architecture with a CLIP-style contrastive loss to the fashion dataset, specifically using the DeepFashion-MultiModal dataset. Our model improves these past studies and exceeds their limits with the newest methods.

III. Methodology

A. Dataset

In the DeepFashion study [1] they contribute DeepFashion, a large-scale clothes dataset, to the community. DeepFashion has several appealing properties. First, it is the largest clothing dataset to date, with over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos, making it twice the size of the previous largest clothing dataset. Second, DeepFashion is annotated with rich information on clothing items. Each image in this dataset is labeled with 50 categories, 1,000 descriptive attributes, and clothing landmarks. Third, it also contains over 300,000 cross-pose/cross-domain image pairs. Some example images along with the annotations are shown in Fig.1.

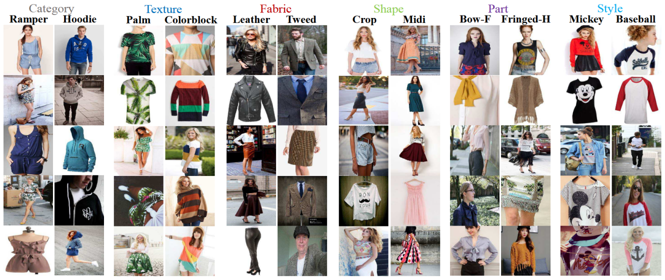

Fig. 1. Example images of different categories and attributes in DeepFashion. The attributes form five groups: texture, fabric, shape, part, and style.

The dataset has been taken from common sources like shopping websites and firm databases. Mostly from *Forever212* and *Mogujie* [1] and they collected 1,320,078 images of 391,482 clothing items. Collecting data has multiple states in the related paper [1]. First they got from Google Images, and then they fed AlexNet with those images. With AlexNet output, they retrieved more clean versions of these images. After that, they annotated and cleaned the outputs. So the final dataset has been created. How they pointed needed anchors can be seen in Fig.2.

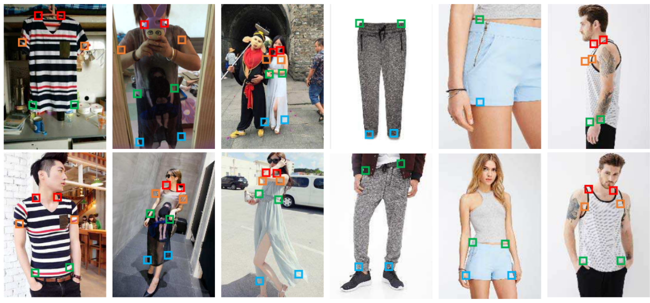

Fig. 2. Example images for different points and anchors in images.

For our study, the dataset was split into training (70%), validation (15%), and test (15%) sets. Approximately 42,544 image-text-label instances were used.

B. Model Architecture

Our architecture relies on different aspects. One core aspect is ViT and the other one is BERT. ViT is the most important and the main reason for this study.

1) Image Encoder: A pre-trained Vision Transformer (*google/vit-base-patch16-224-in21k*). The output embedding (pooler output) is used. This architecture can be seen in Fig.3.

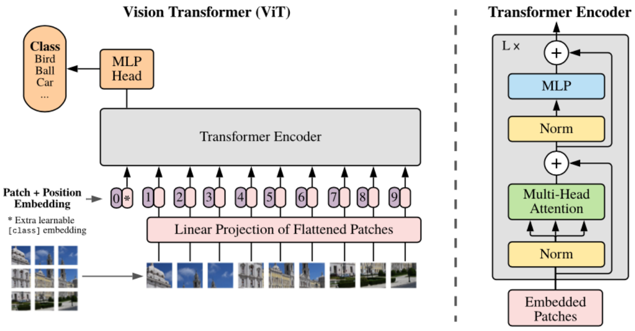

Fig. 3. Example ViT architecture with example input.

2) Text Encoder: A pre-trained BERT model (bert-base-uncased). The [CLS] token embedding (pooler output) is used.

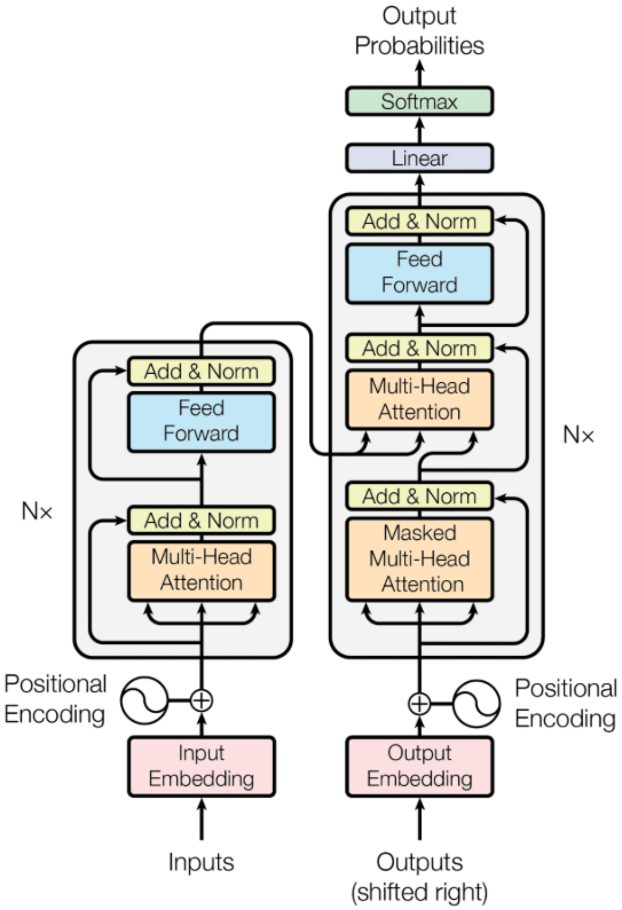

Fig. 4. Example BERT architecture.

3) Projection Heads: Separate linear layers project the ViT image features and BERT text features into a shared embedding space of dimension 512. L2 normalization is applied to these projected embeddings.

4) *Classification Heads:*

- *Category Classifier:* A linear layer Pferde on top of the raw ViT image features (before projection) to predict one of the [NUM_CATEGORIES_FOR_COMPARISON] shape-based categories.
- Attribute Classifier: A linear layer Pferde on top of the raw ViT image features to predict the [NUM_ATTRIBUTES_FOR_COMPARISON] dimensional binary attribute vector (fabric + pattern).

C. *Training Strategy*

Our training has two different phases, which makes our architecture more robust and strong.

*1) **Phase 1 (Contrastive Learning)**:* This is the phase that our contrastive learning model created. In Fig.5 this architecture can be seen.

- The image and text encoders, along with their projection heads, were trained using a symmetric contrastive loss function, similar to CLIP. Given a batch of N image-text pairs, the model aims to maximize the cosine similarity of N correct pairs while minimizing it for N2N incorrect pairs. A learnable temperature parameter (logit_scale, initialized with the logarithm of 1/0.07) was used.
- Optimizer: AdamW.
- Batch Size: 150. Epochs (tried different epochs for understanding overfitting and underfitting): 150, 100, 50, and finally 20.
- Learning Rate: 1e-5
- Weight Decay: 1e-4
- Learning Rate Scheduler: torch.optim.lr_scheduler.ReduceLROnPlateau was used, reducing the learning rate by a factor=0.2 with a patience=3 epochs if the validation contrastive loss did not improve.
- Gradient Clipping: Gradient norms were clipped at a maximum value of 1.0.
- Model Selection: The contrastive loss on the validation set was monitored at the end of each epoch. The model checkpoint yielding the lowest validation contrastive loss (best_contrastive_model.pth) was saved for the subsequent fine-tuning phase and for the retrieval evaluation tasks.

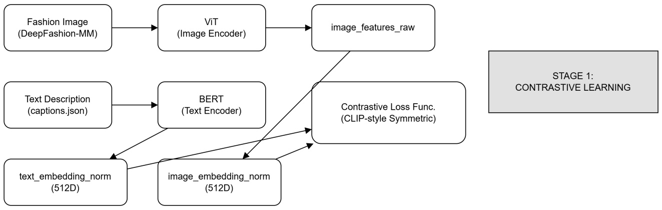

Fig. 5. Phase 1 system architecture.

*2) **Phase 2 (Classification Head Fine-Tuning)**:* In phase 2 we made an classification head fine-tuning. In Fig.6 this architecture can be seen.

- The weights of the pre-trained image and text encoders (from Phase 1, using the best contrastive model) were frozen.
- Only the newly added category and attribute classification heads were trained.
- Optimizer: AdamW.
- Loss Functions: Cross-Entropy Loss for category prediction and Binary Cross-Entropy with Logits Loss for multi-label attribute prediction.
- Batch Size: 150. Epochs (tried different epochs for understanding overfitting and underfitting): 150, 100, 50, and finally 20.
- Learning Rate: 5e-5.
- Weight Decay: 1e-4.
- Category Prediction (Shape-based): torch.nn.CrossEntropyLoss (with ignore_index=-1 for samples without a valid category label).
- Attribute Prediction (Combined Fabric+Pattern, binary multi-label): torch.nn.BCEWithLogitsLoss.
- Total Loss: A simple sum of these two loss functions was used (L_total = L_category + L_attribute).
- Learning Rate Scheduler: torch.optim.lr_scheduler.ReduceLROnPlateau was used, reducing the learning rate by a factor=0.2 with a patience=2 epochs if the total validation classification loss (category + attribute) did not improve.
- Gradient Clipping: Gradient norms were clipped at a maximum value of 1.0.
- Model Selection: The total classification loss on the validation set was monitored at the end of each epoch. The model checkpoint yielding the lowest total validation classification loss (best_classification_model.pth) was saved for the final category and attribute prediction evaluations.

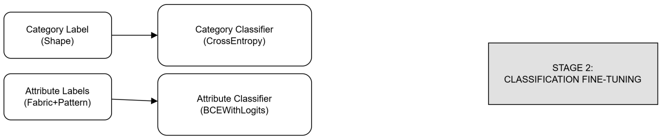

Fig. 6. Phase 2 system architecture.

IV. Experiments and Results

A. *Evaluation Metrics*

To compare our model performance with currently used ones and the Liu Z. [1] we calculated the three most important metrics. This evaluation architecture can be seen in Fig.7.

- Retrieval: Recall@K (R@1, R@5, R@10) for both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval on the test set.
- Category Prediction: Top-K Accuracy (K=1, 3, 5) on the test set.
- Attribute Prediction: Average Recall@K (K=1, 3, 5) for positive attributes on the test set, similar to Liu et al. [1].

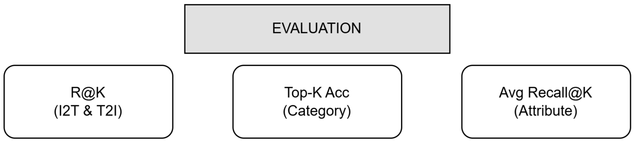

Fig. 7. Evaluate architecture.

B. Quantitative Results

1) Learning Curves: To achieve these results, the model was evaluated for about 95+ hours on an A100 GPU. First, epochs were set to 150 epochs for every training phase. With this training, our model was overfitting the data, which can be seen in Fig.8. Validation loss was starting to get higher and higher. Also, when we tried to evaluate this overfitted model on our test data, we could see our model metrics were not as good as expected in fig.9.

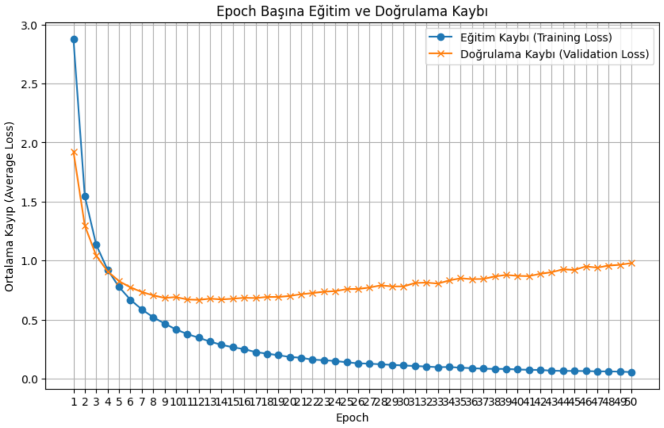

Fig. 8. 150 epoch training model training.

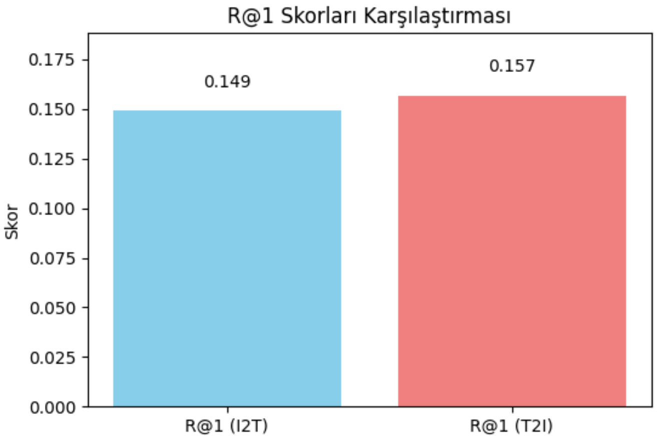

Fig. 9. 150 epoch training model results.

When the loss curve is inspected in detail, we can see our best epoch is between epochs 15 and 20. After that discovery, we set our epoch to 20 to get better results. And we created the new contrastive learning phase. The training loss consistently decreased, indicating learning. The validation loss initially decreased but started to plateau and slightly increase after epoch 15-20 (starting about epoch 10), suggesting the onset of overfitting. The model checkpoint with the best validation loss was used for subsequent retrieval evaluation. These loss results can be seen in Fig.10.

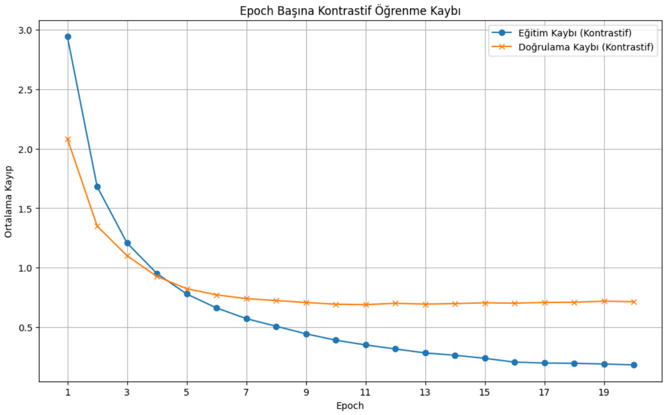

Fig. 10. Final model training loss graphic.

When our phase one is ended, it comes to the fine-tuning phase. We used the same approach here with 20 epochs. Fig. 11 and Fig. 12 shows the loss curves for fine-tuning the category and attribute prediction heads. Both training and validation losses for category prediction (shape-based) show a good decreasing trend, indicating effective learning and generalization.

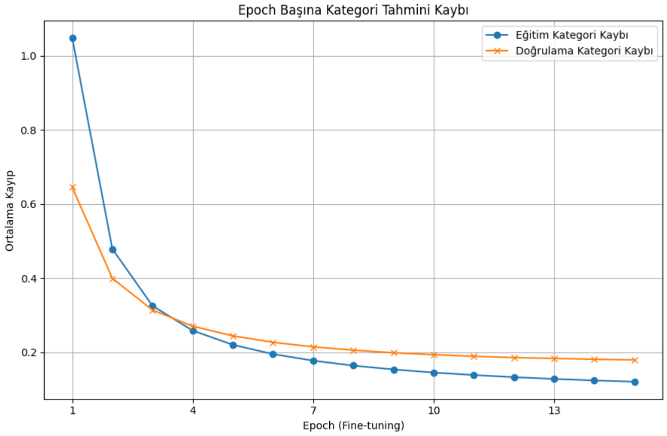

Fig. 11. Final model training category loss graphic.

2) Retrieval Performance: Retrieval performance of our model can be seen in Table 1. In the referenced paper [1] these scores weren't mentioned, so we can't really compare these. This is special to our ViT and BERT implementation. We can not compare these results with current state-of-art because this study is the first with our dataset.

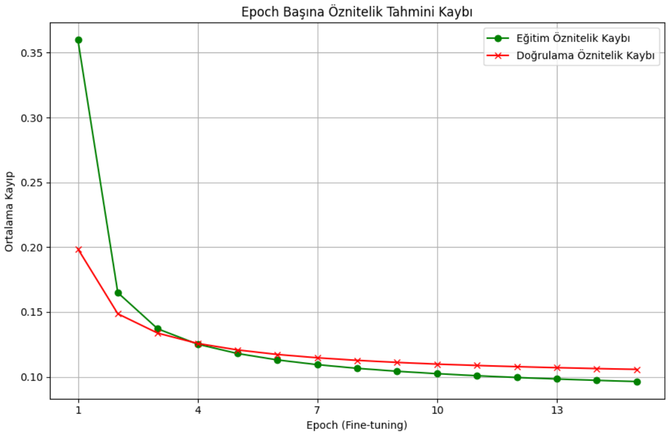

Fig. 12. Final model training attribute loss graphic.

TABLE I
IMAGE-TEXT RETRIEVAL PERFORMANCE (R@K)

Direction	R@1	R@5	R@10
I2T (Image-to-Text)	0.1914	0.4007	0.5492
T2I (Text-to-Image)	0.2175	0.4088	0.5539

The R@10 scores indicate that for a given query, the correct match is found within the top 10 results over 55% of the time.

3) Category Prediction: The category prediction task was based on the primary shape attribute. Our results are compared with FashionNet (VGG) from Liu et al. [1] in Table 2.

TABLE II
CATEGORY PREDICTION ACCURACY (TOP-K)

Metric	Our Model (ViT-based)	Liu et al. [1]
Top-1 Acc (Category)	0.9470	∼55-65%
Top-3 Acc (Category)	0.9962	∼75-85%
Top-5 Acc (Category)	0.9995	∼80-90%

Note: Dataset splits might not be identical due to random splitting.

In Table 2, it can be seen that our model significantly outperforms the reported FashionNet results, demonstrating the strength of the ViT backbone for this task. This is mostly because of the newly used transformer architecture with our model.

4) Attribute Prediction: Attributes were derived from one-hot encoded fabric and pattern annotations. Results are compared in Table 3.

TABLE III
ATTRIBUTE PREDICTION (AVERAGE RECALL@K)

Metric	Our Model (ViT-based)	Liu et al. [1]
Avg Recall@1 (Attributes)	0.1598	N/A
Avg Recall@3 (Attributes)	0.4634	∼30-38%
Avg Recall@5 (Attributes)	0.7291	∼45-55%

Our model shows competitive, and in R@5, superior performance compared to the reported FashionNet results for attribute prediction. So we can see our model's strong side with the transformer.

C. Discussion

The contrastive learning phase successfully learned aligned image and text embeddings, as evidenced by the retrieval scores being significantly better than in other studies. The overfitting observed in contrastive validation loss suggests that further regularization or more data could be better for future studies.

The classification fine-tuning phase was highly successful. The ViT image features proved very effective for predicting both shape-based categories and combined fabric/pattern attributes, outperforming the baseline CNN (FashionNet) from Liu et al. [1] on comparable tasks, despite potential differences in dataset splits and precise label definitions. This highlights the power of pre-trained ViT models for visual feature extraction in the fashion domain.

V. CONCLUSION AND FUTURE WORK

DeepFashion study presented a large-scale clothing dataset with comprehensive annotations. DeepFashion contains over 800,000 images, which are richly labeled with fine-grained categories, massive attributes, landmarks, and cross-pose/cross-domain image correspondence. It surpasses existing clothing datasets in terms of scale as well as richness of annotation. When it comes to inference and better modeling this study and overall project successfully developed a multimodal system for fashion style embeddings using ViT and BERT with contrastive learning. Our newly developed system demonstrated reasonable performance on image-to-text and text-to-image tasks. Especially fine-tuning classification heads on the learned visual features made the most difference. With the benefits from new techniques, we achieved excellent results in category (shape) and attribute (fabric/pattern) prediction, surpassing the performance of earlier benchmarks like FashionNet on the DeepFashion [1] dataset. Even though we got better results and improved all performance overall, this study can be improved in future studies. Future work could focus on:

- Improving retrieval performance through larger-scale pre-training or more advanced contrastive learning techniques.
- Implementing more sophisticated regularization methods to mitigate overfitting during contrastive learning.
- Exploring attention visualization techniques to better understand what visual and textual cues the model uses for style representation.
- Extending the model to handle more fine-grained attribute prediction or to generate textual descriptions from images.
- Developing a user interface for a personalized fashion recommendation system based on the learned style embeddings.

References

[1] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1096-1104).

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[3] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning (pp. 4904-4916). PMLR.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).