Instagram post classification and likes prediction

The idea behind the proposed task

The task was to create a multimodal system capable of predicting the specific category of a post and the possible number of likes that post can get from Instagram. The core objective was to leverage the visual and numerical data from social media posts to predict the popularity of content, measured in terms of number of likes. This involved understanding the relationship between the content of a post (landscapes or animals) and its engagement metrics (likes, comments, and distribution).

The inspiration

The approach was inspired by the the paper "Multimodal Post Attentive Profiling for Influencer Marketing", which utilized Inception-v3 for image encoding and BERT for the text of a post description. In the paper, a multimodal approach was employed for the task of influencer profiling [1]. The change to a Vision Transformer architecture was motivated by its recent successes in various image-related tasks and its ability to handle images as sequences of patches, which aligns with the Transformer's sequence processing capabilities.

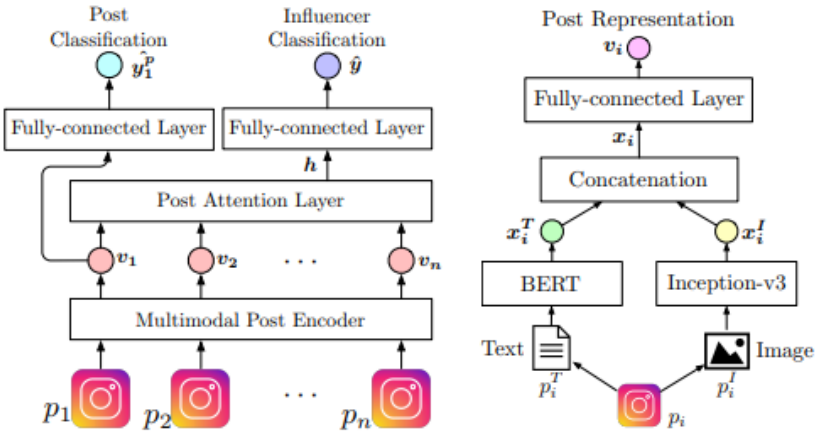

Figure 1. Proposed approach in the paper "Multimodal Post Attentive Profiling for Influencer Marketing" [1].

The solution for the task

To address the prediction and classification aspects of the task, a multimodal approach was also proposed, that combines image and numerical data processing. However, for this project, a Vision Transformer architecture was used for image processing and a separate regression network to predict the number of likes based on comments and post distribution was implemented. This regression network was designed under the assumption that landscape images generally garner more likes, therefore being more popular, than animal images [2][4][6].

The architecture of the proposed algorithm

The model's architecture comprises two main components: a Vision Transformer, also known as ViT, for image encoding and a regression network for predicting the popularity of a post. The ViT processes images converted into patches to create embeddings, with its structure consisting of 4 blocks of multi-head attention layers, normalization layers, multi-layer perceptrons, and skip connections. The regression network features a convolutional layer for data vectors of one dimension, a fully-connected layer, and a batch normalization layer. In total, the model has approximately 137 million parameters. For training purposes, images were resized to a resolution of 512x512 pixels and the model was trained over 10 epochs [3].

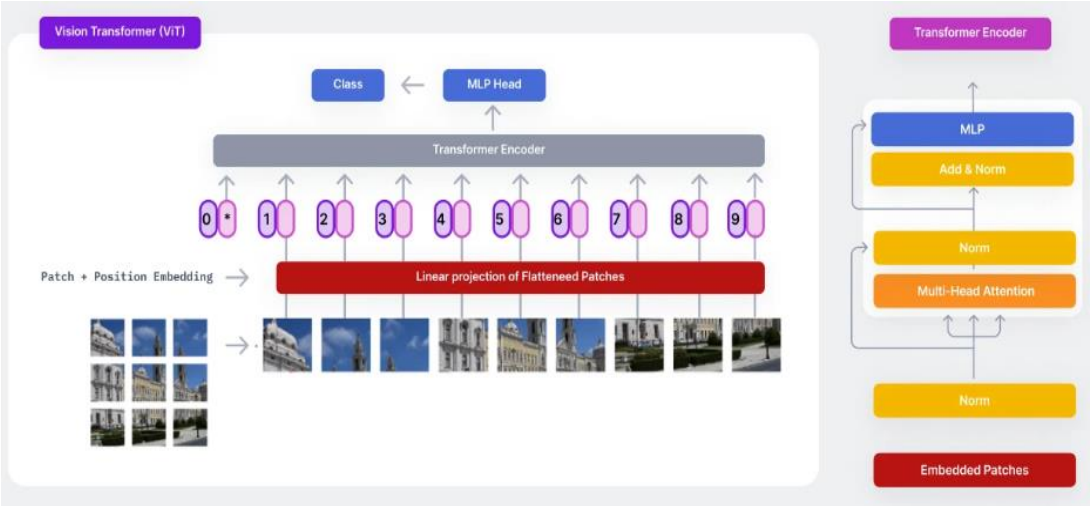

Figure 2. The architecture of a Vision Transformer [5].

Final results

Training was stopped after 5 epochs due to a lack of improvement in the validation loss, indicating that the Vision Transformer component was not learning effectively. The model demonstrated a bias towards landscape images and was unable to achieve an accuracy exceeding 50%. The poor performance is attributed to the small size of the datasets, which consisted of 6064 training files, 1515 validation files, and 1600 test files. The regression data was synthetically generated, with the assumption that landscapes (annotated with the label 0) would be more popular than animal images (annotated with the label 1), which may have also contributed to the bias observed in the model's predictions.

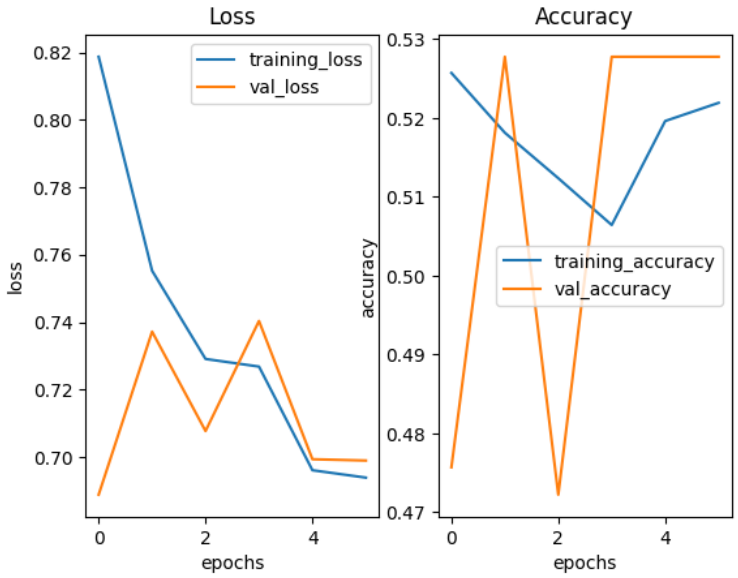

Figure 3. Accuracy and loss for training and validation datasets of the ViT.

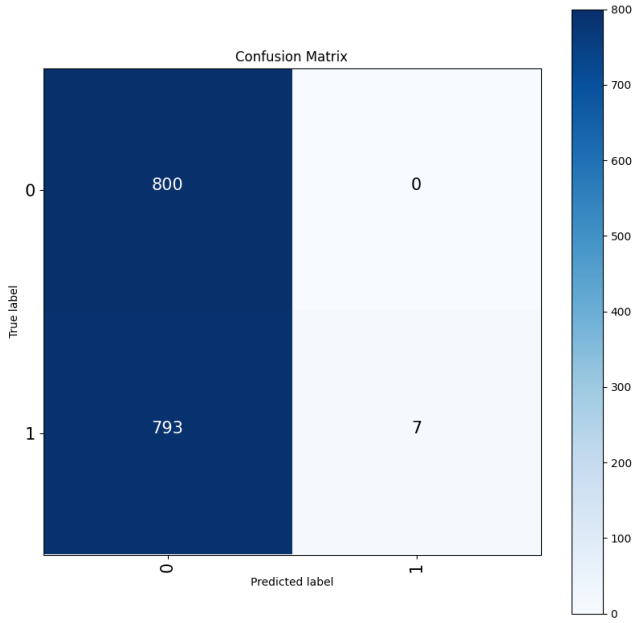

Figure 4. Confusion matrix of the classified images.

	precision	recall	f1-score	support
0.0	0.502197	1.000000	0.668617	800.000000
1.0	1.000000	0.008750	0.017348	800.000000
accuracy	0.504375	0.504375	0.504375	0.504375
macro avg	0.751099	0.504375	0.342983	1600.000000
weighted avg	0.751099	0.504375	0.342983	1600.000000

Figure 5. Classification report results.

Figure 6. Results provided by the multimodal model for “Gura Portitei” image.

References

[1] S Kim, JY Jiang, M Nakada, J Han and ..., "Multimodal post attentive profiling for influencer marketing", *Proceedings of The Web …* (dl.acm.org, 2020), https://doi.org/10.1145/3366423.3380052

[2] A Dosovitskiy, L Beyer, A Kolesnikov and ..., "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv … (arxiv.org, 2020), https://arxiv.org/abs/2010.11929

[3] Mehra, Akshit. "Vision Transformers For Object Detection: A Complete Guide." Labellerr, October 23, 2023. https://www.labellerr.com/blog/vision-transformers-for-object-detection/

[4] "Vision Transformer (ViT)". Hugging Face. https://huggingface.co/docs/transformers/model_doc/vit

[5] Shah, Deval. "Vision Transformer: What It Is & How It Works [2023 Guide]." V7 Labs, December 15, 2022. https://www.v7labs.com/blog/vision-transformer-guide

[6] "Vision Transformer Explained". Papers With Code. https://paperswithcode.com/method/vision-transformer