ZKG INTERNATIONAL

ISSN: 2366-1313

Vision Transformer Method to Predict Author's Gender from images

M.Swapna[1]
Research Scholar Dept. of Informatics,
Osmania University, Hyderabad
Assistant Professor
Department of Computer science and Engineering
Matrusri Engineering College Hyderabad , Telangana, India
Email: mandaswapnareddy@gmail.com

Dr.K.Nikitha[2]
Govt.Degree College Lecturer,
Dept. of Computer Science, Mahatma Jyothiba Phule,
Warangal, Telangana, India
Email: nikithakukunuru@gmail.com

Abstract

The ability to predict an author's gender from visual data presents a novel challenge and opportunity in the realm of machine learning and computer vision. This study explores the use of Vision Transformers (ViT), a cutting edge deep learning model, to infer the gender of authors based on images.

Traditional methods of author profiling primarily relied on textual analysis, but the increasing availability of visual content associated with authors offers a new avenue for exploration. Vision Transformers, known for their powerful image processing capabilities and capacity to capture complex patterns, were employed to analyze these visual inputs.

Our approach involved training a ViT model on a dataset of shared images, where gender labels were annotated for each author. We assessed the model's performance in classifying the gender of authors based on visual features extracted from the images. The results demonstrate that ViT can effectively leverage image-based information for gender prediction, achieving significant accuracy.

The study highlights both the potential and limitations of using visual data for demographic prediction. While the ViT model shows promise, factors such as image quality and dataset diversity impact performance. Future research directions include, to enhance prediction accuracy. In our study, we propose an approach that predicting gender of the author from the shared images by using Vision Transformer (ViT) mehtod. By leveraging ViT's superior image processing abilities, our approach achieves outstanding results.

ISSN: 2366-1313

Keywords: Author Profiling, Image data, Vision Transformer, Image processing, Deep learning

1. Introduction

In recent years, the intersection of computer vision and natural language processing has fostered significant advancements in various fields. One area of interest is the ability to infer characteristics of individuals based on visual data. Specifically, predicting the gender of an author from shared images of their work poses both intriguing challenges and potential applications.

Traditionally, predicting author attributes from text has relied on linguistic features and stylometric analysis. However, with the increasing prevalence of images associated with authors ranging from social media—there is a growing interest in leveraging visual information to make predictions about authors' demographic characteristics.

In this context, Vision Transformers (ViT) have emerged as a powerful tool for image analysis. Unlike conventional convolutional neural networks, ViT models utilize transformer architecture to process images, capturing complex patterns and relationships within the data. Their ability to handle large-scale datasets and extract high-level features makes them particularly suited for tasks involving image-based classification and prediction.

This study explores the application of ViT in predicting the gender of authors based on images associated with their works. By analyzing shared images through a Vision Transformer model, we aim to determine whether visual features can effectively indicate the gender of an author. This approach not only leverages state-of-the-art image processing techniques but also provides a novel perspective on author identification based on visual data.

The following sections will outline the methodology employed, including, preprocessing, and the implementation of the Vision Transformer model. We will also discuss the results and implications of our findings, contributing to the broader discussion on the role of visual information in gender prediction.

2. Dataset Description

The task undertaken in the AP challenge at PAN-2018 centred on detecting users' gender based on their tweets and photos shared on Twitter. Organizers distributed separate training datasets for three languages—Arabic, English, and Spanish. The English and Spanish datasets comprised information from

ISSN: 2366-1313

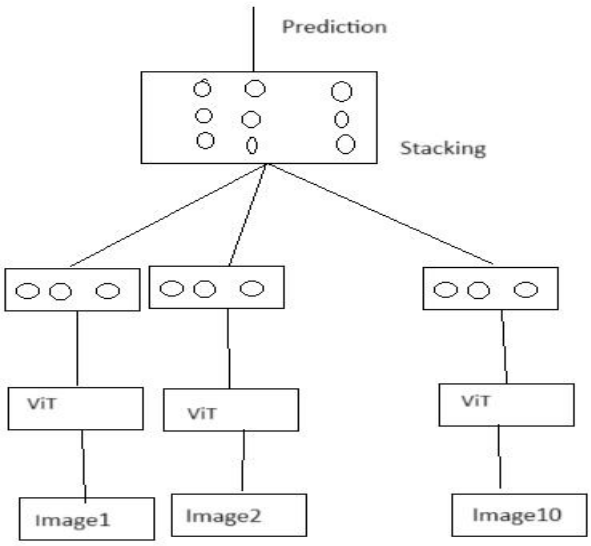
3000 users each (1500 male, 1500 female), while the Arabic dataset contained data from 1500 users (750 male, 750 female). Each user's data included 100 tweets and 10 images. The organizers also supplied a test dataset; further details are available in the overview paper of the AP task

3. Vision Transformer (ViT)

The Vision Transformer (ViT) exhibits a notably reduced inductive bias compared to Convolutional Neural Networks (CNNs) in terms of image-specific information. ViT primarily employs local and translational equivariant Multi-Layer Perceptron (MLP) layers, while Vision Transformers (ViT) the self-attention layers operate globally. The image is initially divided into patches, with position embeddings resized as required during fine-tuning. Alternatively, the input sequence can be derived from a CNN's feature map, onto which patch embedding projection is applied.

Visual Transformers undergo pre-training on extensive datasets and are subsequently fine-tuned for specific downstream tasks. During fine-tuning, a projection head is removed, and a zero-initialized D × K feedforward layer is added, where K represents the number of classes for downstream tasks. Additionally, employing higher resolutions during fine-tuning can be advantageous. ViT is capable of accommodating sequences of arbitrary lengths, with pre-trained position embeddings often proving adequate.It is important to note that adjustments in resolution and patch extraction are the sole stages at which an inductive bias regarding the 2-dimensional structure of images is manually incorporated into Vision Transformers.

Fig. 3.1 Image Classification Architecture



ISSN: 2366-1313

The image component utilizes the ViT architecture and performs the following specific preprocessing steps to process the images. Image preprocessing is essential, and the following

steps are applied to the images:
a) Random resizing to 256 × 256 pixels.
b) Random horizontal flipping.
c) Random rotation within the range of 10° to 15°.
e) Final resizing of all images to 224 × 224 pixels.
f) Normalization of the data

The features are extracted from each image by using ViT pretrained architecture.

$$Fi=ViTImagei, \text{ where } i=1, \ldots, 10$$

The extracted features are stacked and averaged as

$$F=Average(F1, \ldots, F10)$$

Finally, the average of the combined representations is fed into a fully connected layer to obtain a feature representation. This feature representation is then utilized for fusion to achieve the desired outcomes

$$ImgRepr=FC(F)$$

Finally the ImgRep is passes to the output layer, which classifies the gender by using sigmoid activation function.

Algorithm (Image data)

For each Image X in dataset

For each Author i

Apply ViT (X_i) and extract visual features

Stack the visual features

Apply Fully connected layers

Predict the Gender

ISSN: 2366-1313

4. Result Analysis

We have implemented pretrained ViT and extracted features from images. The results are presented in the below table

Table. 4.1 Performance Comparison of proposed and existing

Models	accuracy
VGG-16	76.12
ResNet-50	78.31
ResNet101	51.6
EfficientNet	82.05
ViT	83.17

Vision Transformers represent a significant advancement in the field of computer vision, offering a powerful alternative to traditional CNNs. Their ability to capture global context through self-attention mechanisms makes them particularly well-suited for complex image understanding tasks.

5. Conclusion

This study has investigated the efficacy of using Vision Transformers (ViT) for predicting the gender of authors from images. By leveraging the advanced capabilities of ViT, we explored how the shared images can be indicative of an author's gender.

Our findings demonstrate that ViT, with its ability to capture intricate patterns and high-level features from images, holds significant potential for this predictive task. The model's performance underscores the viability of using image-based data in conjunction with sophisticated deep learning techniques to infer author attributes. This approach represents a novel extension of existing methods, traditionally reliant on textual analysis or demographic information.

The results reveal that while ViT models can achieve notable accuracy in predicting gender, there are inherent challenges and limitations. Factors such as image quality, context, and the diversity of the dataset impact the model's performance. These findings highlight the importance of robust data preprocessing and model training to enhance prediction reliability.

ISSN: 2366-1313

Future work could focus on addressing these limitations by incorporating additional data sources or refining the ViT architecture. Exploring different image modalities and integrating multimodal approaches could further improve prediction accuracy.

In conclusion, the application of Vision Transformers for gender prediction from images opens up new avenues for research and practical applications. This study contributes to the growing body of knowledge on leveraging visual data for demographic inference and sets the stage for further exploration in this innovative domain.

References:

1. C. Suman, A. Naman, S. Saha and P. Bhattacharyya, "A Multimodal Author Profiling System for Tweets," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2021.3082942
2. Carmona, Miguel Angel & Villatoro-Tello, Esaú & Montes, Manuel & Pineda, Luis. (2020). Author Profiling in Social Media with Multimodal Information. Computación y Sistemas. 24. 1289–1304. 10.13053/CyS-24-3-3488.
3. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T., 2018. Text and image synergy with feature cross technique for gender identification. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Vol. 2125. pp. 10–22.
4. Farnadi, G., Tang, J., De Cock, M., Moens, M.-F., 2018. User profiling through deep multimodal fusion. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, pp. 171–179.
5. Jia, X., Song, S., He, W., Wang, Y., Rong, H., Zhou, F., Xie, L., Guo, Z., Yang, Y., Yu, L., et al., 2018. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. arXiv preprint arXiv:1807.11205.
6. Wendlandt, L., Mihalcea, R., Boyd, R. L., Pennebaker, J. W., 2017. Multimodal analysis and prediction of latent user dimensions. In: International Conference on Social Informatics. Springer, pp. 323–340.
7. Segalin, C., Cheng, D. S., Cristani, M., 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. Computer Vision and Image Understanding 156, 34–50.
8. Estruch, C. P., Paredes, R., Rosso, P., 2017. Learning multimodal gender profile using neural networks. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 577–582.

ISSN: 2366-1313

9. Farseev, A., Nie, L., Akbari, M., Chua, T.-S., 2015. Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, pp. 235–242.
10. Merler, M., Cao, L., Smith, J. R., 2015. You are what you tweet. . . pic! gender prediction based on semantic analysis of social media images. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on. IEEE, pp. 1–6.
11. Vinokur, A. I., 2015. Information technologies in culture and education: Image processing issues. Modern Applied Science 9 (5), 314.
12. You, Q., Bhatia, S., Sun, T., Luo, J., 2014. The eyes of the beholder: Gender prediction using images posted in online social networks. In: Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, pp. 1026–1030.