Age and Gender Estimation based on the Visual Transformer

*Note: Sub-titles are not captured in Xplore and should not be used

Aldiyar Bolatov Arsen Abzhanov Adilet Mukashev

***Abstract*—Automated age and gender estimation becomes relevant in many applications. There are numerous ways to predict age and gender based on human voice, face features and posture. In this paper an image based approach will be studied. The method requires 2 dimensional images of people's faces. The challenging problem with this method is that its performance reduces severely when experiments applied to the faces in the unconstrained environment. Another problem is aging differences based on personal lifestyle, genetics and environment. To put it simply different people age differently. Yet another challenge is the distinction between biological age and apparent age. The methods based on facial images are of 2 types. First one is hand-crafted feature extraction and classification, the second using deep neural networks. In our paper, we also proposed the vision transformer for age estimation. It was the first vision transformer proposed for facial tasks, and thus there were no pretrained models. However, we still managed to achieve some results in a low-data environment. The future steps would be to pre-train the model on larger face datasets.**

***Index Terms*—Age, Gender, Estimation, Visual, Transformer, VGG CNN**

I. Introduction

The number of applications where automatic face recognition is employed increased tremendously. They are used for intelligent video surveillance, targeted marketing, user authentication, content filters for minors [1]. The age and gender estimation is a task of determining a person's age from biometric features. There are 2 main methods for age estimation, the first is age multi-class grouping and the second is age regression. There are also label distribution and ranking methods [2]. For the gender recognition task, the type of problem is binary classification.

The approach based on images where faces are extracted have several challenges. The first challenge is unconstrained conditions of the environment where images were taken. In literature the unconstrained images are referenced as Faces in the Wild. While constrained conditions with adience dataset achieve accuracy of 66.6%, the estimation in unconstrained conditions with 4 age groups based on KNN method achieves about 90% [3] [4]. However the comparison does not tell us about difficulties in estimation for unconstrained environments. The research done by Zhang shows us that unconstrained conditions significantly reduce the accuracy of the model [5]. For the gender estimation in a constrained environment the best results are 99.3% of accuracy. While unconstrained conditions yield results of lower accuracy [6]. The good example of constrained face images is 'The extended Yale B Database'.

Another challenge is aging differences across the population of the World. There were numerous attempts to tackle that problem by extracting aging features or by extracting facial landmarks [7] [8]. The challenge of aging is connected to the problem of biological age and apparent age. The apparent age estimation was attempted to be solved on the ChaLearn LAP competition. The apparent age task is different because the dataset is composed of age estimates given by respondents. Additionally, the age had different values, the first is mean value, the second is standard deviation. Hence the metric score was different. In 2016 competition the best result was achieved by a VGG16 model pretrained on ImageNet dataset as for biological age estimation, and fine-tuned on the dataset with apparent age [9].

The third challenge is cosmetics, clothes, and jewelry. The research shows that the impact of cosmetics is significant. In the paper done by Chen they analyze the effect of cosmetics on automatic gender classification models [10]. Their results showed that classification rate drops significantly. As an example AdaBoost system had classification rate drop from 78.33% to 30% after makeup has been applied on the same experiment participants.

The last challenge is the lack of appropriate datasets. Most of the datasets are taken from the Wild and models can not adequately learn because of the noise. Also the images were not distributed evenly across the ages, there were less images of children [11]. Additionally, some of the databases are biased because they contain images of specific races.

II. Related Work

The 2 main methods for automated age and gender estimation are hand-crafted feature extraction and deep neural networks. In the first approach some face patterns are extracted using predetermined algorithms, then these features are fed into the neural network. For example, there was a research on aging pattern extraction that described the shape of the face [2]. The accuracy improved over time by exploring more features and combining them with the already discovered. Another method was Geometric. It was based on the Gabor feature classifier. It is a filter designed to retrieve important

Identify applicable funding agency here. If none, delete this.

properties as spatial localization, orientation selectivity and others [12]. The Geometric method focuses on individual features as eyes, nose and mouth [13]. Another approach is to use local binary patterns, it is used as a texture descriptor [14].

The second approach is to use deep neural networks. In this method, network is usually pretrained for face feature extraction, so that model has properties of low intra-personal variations and high inter-personal differences [4]. After that model is fine-tuned for specific tasks like gender and age classification.

There are different network designs for the approach of neural networks. The most popular design choice is convolutional neural networks. The first architecture is VGG which is used to extract facial features. The VGG model comprises 11 blocks followed by linear operator and RELu with max-pooling operations. The first 8 blocks are convolutional layers and the last three are fully connected layers. The model was trained on 2.6m images of 2.6K different people [15]. In the research done by Qawaqneh they replaced the last fully connected layers with 4 new fully connected layers. The last layer had 8 outputs corresponding to the age class [16]. It is noteworthy to mention other architectures that are extensively used for age and gender classification. They are ImageNet, DEX-IMDB-WIKI and DEX-ChaLearn-ICCV2015. DEX-IMDB-WIKI network was pretrained on ImageNet, then further trained using 500K images from IMDB and wiki. After that it was fine-tuned using a dataset with apparent age annotations [17].

Another appropriate example is where trained CNN was used is a work done by Mallouh [18]. They utilized several pretrained models like VGG, GoogleNet, ResNet. They combined feature vectors from the last convolutional layers of each architecture and reduced their dimensions.

Another network design is Residual Network of Residual Networks done by Zhang [5]. Their work is based on the fact that deeper networks achieve higher results. However deep networks have the problem of vanishing gradients and over-fitting. To alleviate these they used the Residual network of Residual Networks which contained 3 kinds of residual blocks. As in the CNN architecture they firstly pretrained models using ImageNet and fine-tuned it using IMDB-WIKI datasets. At the end the whole structure is fine-tuned using the Adience dataset [5]. The architecture is further improved by the same author by utilizing LSTM network for attention. The additional LSTM layer after all the residual blocks differentiate between age sensitive features and not important ones. In their work without LSTM they achieved accuracy of 61.68% for Adience dataset, but with LSTM they improved their results and achieved 67.83% on the same dataset [19].

Even more fine-tuning was done to meet more specific results in the work done by Antipov [9]. In their research they trained CNN model on ImageNet database then fine-tuned it using IMDB-WIKI dataset for biological age. After that they fine-tuned the model for apparent age estimation given by the competition dataset. Then they made a copy of network that was trained on face feature extraction and biological age, and fine-tuned it using a dataset with images of children in age between 0 and 12 for apparent age estimation. During the test they determined the biological age of a person on the image and based on the age guessed they forwarded it to the appropriate network.

Also there are methods that use facial images with additional information to estimate the age and gender. For example some studies focused on other features like body proportions. In the work done by Collins et al, they studied gender recognition by shape descriptors [20]. This method is found to be more useful since surveillance cameras can not capture people's faces accurately, because of the angle and occlusions. Hence the approach done Gonzalez-Sosa combined both of the techniques to get more reliable results in gender classification [6]. For their research they combined LBP face feature extraction method with shape-based features extracted from the person's silhouette, vectors from both methods are normalized to the same length and fused together.

A. *Datasets*

The first dataset is Adience. It was designed specifically for age and gender estimate models. Authors that designed this benchmark database included images of different ages, races, occlusions, angles [3]. The dataset contains 26K photos from 2284 subjects. The dataset is divided into 8 non-overlapping classes. The dataset designed for age and gender estimation in the wild. Images are not constrained and have different angles. They were taken from the Flickr website.

The second dataset is FG-NET. It is a database that was made in 2004 to initiate research in the learning ageing process [21]. Later it was used for age estimation tasks. The dataset has 1002 images of 82 subjects. The images are in the range of 0 to 69 years old, the highest intensity of images for the age of 40. This dataset is useful for extracting important face features, since subjects' images are taken throughout their ageing process. However, it does not meet the requirements of the unconstrained nature of the images. All of the images taken were in a controllable environment where the light and angle was adjusted. The MORPH dataset is similar to FG-NET, it contains images of adults taken in the range of a few months to 20 years. The same problem occurs for the MORPH dataset, it is not designed for unconstrained age and gender estimation [22].

The third dataset is IMDB-WIKI dataset. It contains 500k images of celebrities taken from IMDB and Wiki pages. It contains the age of the person on the photo and location of the face. It is the biggest dataset for facial age estimation nowadays. It was used in the 2015 ChaLearn LAP competition and achieved an error of 0.264975 which is the best result [17]. This dataset is more relevant for training model in unconstrained conditions since photos were taken in the wild.

III. Setup

For our project we used Adience dataset. We used traditional approach and trained our model on 5 fold dataset [3]. The model is based on Visual Transformer and attention mechanism. Also we reproduced the results done by Qawaqneh. [16]. The model was fine tuned using Adience dataset and results are shown in the Table I. The further details will be given in the next sections.

IV. VGG-Face model reproduced

VGG16 - a Convolutional Neural Network created for face recognition can be fine-tuned for age classification. There is an abundance of data for face recognition, as opposed to age classification. VGG16 extracts facial features that could assist in tackling related tasks.

One of the concerns regarding age classification is that there are not enough datasets for learning. It could lead to overfitting. To counterbalance this issue, some suggest using pre-trained CNN designed for similar tasks and fine-tuning them [16]

The VGG16 model was trained and tested on an Nvidia Geforce Gtx 1060 with 14GB of video memory. 1-off accuracy in age classification of the given model and VGG16 model is shown in tables 1 and 2.

We experimentally proved that the problem of insufficient datasets can be mitigated by using pre-trained neural networks designed for similar tasks.

Visual Transformer

The transformer architecture previously surpassed the RNN-based models in natural language task (NLP) and became the standard model for text-related problems [23]. However, in the computer vision domain, transformer networks were only used with convolutional layers. The reason was to make the overall convolutional neural network (CNN) more expressive [24]. However, only recently, it was shown, that reliance on convolutional layers is not essential [25]. Dosovitskiy et al., in their work, presented a pure transformer model that uses sequences of image patches as an input. They heavily pretrained the model, and then by transfer learning, they mapped the pretrained transformer model to image recognition benchmarks like ImageNet, and CIFAR [25]. Their vision transformer achieved comparable, with the state-of-art, results while consuming fewer computational resources during training.

Benefits of transformer architectures are their computational efficiency, lesser number of hyperparameters to tune, and scalability [25]. This lead to a lack of saturating performance with an increase in the model and dataset.

However, image recognition and face tasks have some dissimilarities. For example, to differentiate between two different animals, it is enough to ”understand” animals’ shape and color. ”Face” tasks are a little bit trickier and have more underlying complexities such as wrinkles and eye bags. Furthermore, pure vision patch transformers were not tried on face tasks such as face recognition or age estimation. This leads us to the question: how satisfactory vision transformers can perform more complex vision tasks such as face classification/recognition.

To fill the research gap, we proposed a patched vision transformer for age estimation. Age estimation is a complex task and relies on the valid feature extraction capabilities of the network. Thus, age estimation can also give us a forecast on how well vision transformers will perform on different face tasks.

Since this work first presents the patch vision transformer for face tasks, there were no available pre-trained models. Therefore, to partially solve this problem, our dataset was augmented. A further complication, results, and dissuasion is given in subsection IV-B. In subsection IV-A, the proposed approach is described in more details.

A. Network Architecture

The overall model architecture can be seen in figure 1. Firstly, since the transformer accepts only a one-dimensional sequence of token embedding, the input image needs to be handled. Dosovitskiy et al. proposed to divide two dimensional images $im \in R^{C\times H\times W}$ into N patches of size $patch_n \in R^{P^2C}$, where C is number of channels; H, W is stand for original image resolution; P, P is the resolution of only one patch and number of patches $N = HW/P^2$ [25]. Then patches are flattened, and N serves as an input sequence length. Afterward, every patch is mapped to some constant dimensionality using the trainable linear projection.

Further, as in BERT [26], a learnable class token is prepended to the embedded patches, which output states of the transformer is used for the classification. Also, because of the fact, those transformers are not aware of the spatiality of the information, learnable positional embedding was augmented into every patch embeddings. In turn, this helped to keep positional information.

At the end we get a sequence x of length $N+1$, where each sequence element is $x_n \in R^D$ and encode spatial information. This sequence is then fed into a six-layer transformer network. The transformer network itself consists of two main layers: multi-head self-attention (SA) and feed-forward neural network (FF).

The self-attention function is a process of quantifying the representation of the relative importance of each sequence element. Here, the self-attention mechanism relates each embedded patch with all other patches of the sequence. Then, this new information and relative importance added to the value representation [23]. This is done using the compatibility function. In our work, we used scaled dot-product because our input sequence length N was relatively low. Different functions are focused on the optimizing quadratic complexity concerning the sequence length [27]:

$$f_{scaled}(Q, K) = \frac{Q, K^T}{\sqrt{d_k}} \tag{1}$$

Then new value representation is then computed using:

$$f_{at}(Q, K, V) = V' = softmax(f_{scaled}(Q, K))V \tag{2}$$

As Vaswani et al. [23] specified, Q, K, V are three different representation of the input sequence, which were acquired

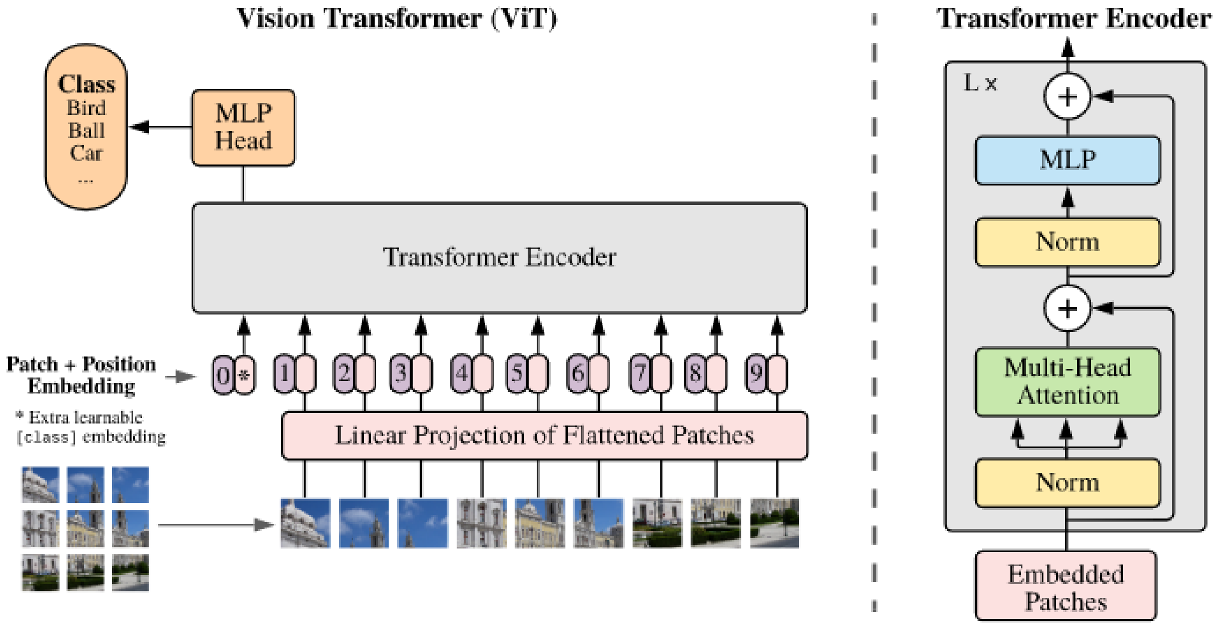

Fig. 1. The Model Structure

using separate learned linear transformation. Usually, the self-attention mechanism is split between heads, which are then concatenated. The purpose is that each head can capture unique features by having individual parameters.

The feed-forward neural network in our work has a form of gated activation linear unit, which is followed by linear projection to the same space as input. There, swish activation function was used [28]. In the formula 3, $LN \in R^{(}D \times D')$ stands for learned linear projection.

$$f_{ff}(X) = LN_D(LN_{glu1}(X) \times swish(LN_{glu2}(X))) \quad (3)$$

The transformer utilizes Layernorm (LN) and residual connections, which both help to better gradient flow. The overall equations that took place in every block (t) of the transformer are presented below with multi-layer perceptron (MLP), which takes output states of the class token.

$$X_t^{'} = SA(LN(X_{t-1})) + X_{t-1} \quad (4)$$

$$X_t = FF(LN(X_t^{'})) + X_t^{'} \quad (5)$$

$$class = MLP(LN(X_6^0)) \quad (6)$$

B. Results and Analysis

The results of the transformer model were below the results of the CNN-approaches, though, most of the CNN-approaches were firstly heavily pretrained. Due to the limited hardware and time resources, we were not able to heavily pre-train the network. Also, pretraining of the transformer model gives more significant advantages than CNN, due to scalability and computational efficiency of the transformer model. CNN model also has an edge because of inductive biases such as translation equivariance and locality. But the transformer network is an almost unbiased network, where the large scaling of the training data trumps inductive bias.

For the comparison reason, we tried the performer model with local attention heads [], which is an efficient network designed for the task with ling sequences. As expected, it under-performed and had an accuracy of 39%. Our model had an accuracy of 39% with one-off accuracy of 69.16. We also tried to add more augmented data, and as expected, the result went up. The graph 2 shows in yellow the validation accuracy when additional data were presented.

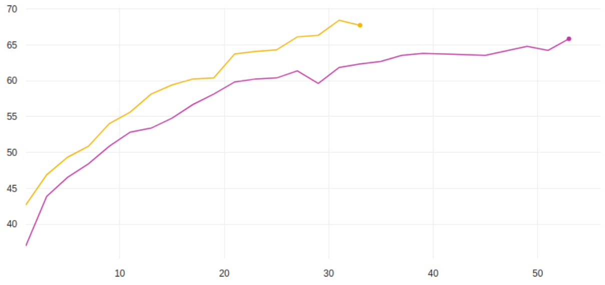

Fig. 2. The Augmented Data

Overall, it can be said that a patched vision transformer is capable of learning and understanding facial features. In the low data environment, it still managed to pull intriguing results. Thus, to continue the research, the model needs to be pretrained on more significant datasets.

Comparison

Comparing with other methods, our method achieved lowest results. The comparison should be made with simple CNN model presented by Gil Levi and Tal Hassner. Despite lower accuracy our model were able to learn faster that their model. Their model took 4 hours, and ours for 3. Other approaches used pretrained networks that improved their results in feature extraction. Additionally, the hand-crafted filters like LBP achieved better results. However we can see that CNN network based on VGG that was pretrained on ImageNet dataset achieves much higher results than LBP which used AdaBoost for face detection. The comparison table is given by the label Table 1.

Conclusion

The work presented several CNN-based approaches and showed that with sufficient resources, it achieves a high score while relying on inductive biases such as locality and translation equivariance.

We also presented the first vision transformer for face tasks, and in our case, age classification. The advantage of the transformer-based approach over the CNN-based is a speed of training, scalability, and exemption of saturation when model or data capacity increased. However, due to limited resources, we trained it in a low-data environment, where it achieved a solid score, which still lower than CNN-based models.

The problem was that CNN-based methods were heavily pre-trained and also received the advantage of their biases in a low-data environment.

The pre-trained vision transformer does not exist still for the facial tasks, and we were not able to pre-train it effectively with limited resources. However, as it can be seen, if new data would be presented in a sufficient amount, the transformer model can significantly increase its performance.

So the plan is to continue the research, where one of the main concerns will be pretraining. The addition of local attention can also improve the result if added in later layers of transformers.

References

[1] J.-H. Yoo, S.-H. Park, and Y. Lee, "Real-time age and gender estimation from face images," in *Proceedings of the 1st International Conference on Machine Learning and Data Engineering (iCMLDE2017), Sydney, Australia*, 2017, pp. 20–22.

[2] A. Othmani, A. R. Taleb, H. Abdelkawy, and A. Hadid, "Age estimation from faces using deep learning: A comparative analysis," *Computer Vision and Image Understanding*, p. 102961, 2020.

[3] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[4] J. Huang, B. Li, J. Zhu, and J. Chen, "Age classification with deep learning face representation," *Multimedia Tools and Applications*, vol. 76, no. 19, pp. 20 231–20 247, 2017.

[5] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep ror architecture," *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017.

[6] E. Gonzalez-Sosa, A. Dantcheva, R. Vera-Rodriguez, J.-L. Dugelay, F. Brémond, and J. Fierrez, "Image-based gender estimation from body and face across distances," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3061–3066.

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2234–2240, 2007.

[8] T. Wu, P. Turaga, and R. Chellappa, "Age estimation and face verification across aging using landmarks," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1780–1788, 2012.

[9] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 96–104.

[10] C. Chen, A. Dantcheva, and A. Ross, "Impact of facial cosmetics on automatic gender and age estimation algorithms," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2. IEEE, 2014, pp. 182–190.

[11] P. Grd and M. Bača, "Creating a face database for age estimation and classification," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 1371–1374.

[12] C. Liu and H. Wechsler, "A gabor feature classifier for face recognition," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 270–275.

[13] R. Rouhi, M. Amiri, and B. Irannejad, "A review on feature extraction techniques in face recognition," *Signal & Image Processing*, vol. 3, no. 6, p. 1, 2012.

[14] S. Bekhouche, A. Ouafi, A. Benlamoudi, A. Taleb-Ahmed, and A. Hadid, "Automatic age estimation and gender classification in the wild," in *Proceeding of the international conference on automatic control, telecommunications and signals ICATS*, vol. 15, 2015.

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[16] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep convolutional neural network for age estimation based on vgg-face model," *arXiv preprint arXiv:1709.01664*, 2017.

[17] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15.

[18] A. A. Mallouh, Z. Qawaqneh, and B. D. Barkana, "Utilizing cnns and transfer learning of pre-trained models for age range classification from unconstrained face images," *Image and Vision Computing*, vol. 88, pp. 41–51, 2019.

[19] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention lstm networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[20] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1235–1242.

[21] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the fg-net ageing database," *Iet Biometrics*, vol. 5, no. 2, pp. 37–46, 2016.

[22] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 341–345.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

TABLE I
MODELS COMPARISON

Name	Type	Dataset	Exact accuracy	1-off accuracy	gender accuracy	Description

VGG pretrained [18]	CNN	ImageNet(pretrain)
		Adience(fine-tune and test)
			60.60%	90.57%	None	The images were augmented, dimensionality reduction used

3 layer CNN [29]	CNN	Adience(train and test)	50.7%	84.7%	86.8%	The images were cropped and over-sampled
Deep RoR	Residual NN	ImageNet(pre-train), IMDB-WIKI(fine-tune), Adience(more fine-tune and test)	61.78%	92.15%	90.87%	The network has 82 convolutional layers and uses random drop of layer subset
LBP [30]	hand-crafted filters	Gallagher group images	48.3%	83.0%	73.6%	Adaboost was used with LBP
Visual Transformer	Transformer	Adience(training and testing)	41.56%	62.17%	None	Soft max attention mechanism was used
VGG(our try)	CNN	ImageNet(pretrain), Adience (fine-tune)	67%	92%	None	

[27] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are rnns: Fast autoregressive transformers with linear attention,” *arXiv preprint arXiv:2006.16236*, 2020.

[28] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *arXiv preprint arXiv:1710.05941*, 2017.

[29] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[30] C. Shan, “Learning local features for age estimation on real-life faces,” in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, 2010, pp. 23–28.