Creating Age, Gender and Ethnicity Predictions through the use of CNN's

Arjun Khullar

UWCSEA Dover, Singapore
Email: *arjunkhullar2006[at]gmail.com*

Abstract: ***Humans are remarkably capable at detecting age, gender and ethnicities solely from a picture of an individual's face. For a computer to be able to identify individuals from images of faces would be challenging for many reasons - the variation in proportions and resolution of images, alongside the presence of facial expressions amongst them. Additionally, this type of facial recognition has several applications, including security, entertainment, and virtual reality. In particular, we will be designing an algorithm to predict the age, gender and ethnicity of an individual based solely on a headshot image of them. We were drawn to this particular application by its seeming abstractness and challenge. We used a convolutional neural network for this, as it is able to learn adaptively using facial features, and has been used in the past for similarly purposed algorithms. The algorithm was deployed on a data set of 23,000 images, with 70% being used for training, 10% for validation, and 20% for testing. We achieved an 80% accuracy in predicting age, gender and ethnicity with this algorithm, and have demonstrated its ability to be deployed in a real-world setting.***

Keywords: Age Prediction, Gender prediction, Ethnicity prediction, Machine learning, Artificial intelligence

1. Introduction

Humans are remarkably capable at detecting age, gender and ethnicities solely from a picture of an individual's face. For a computer to be able to identify individuals from images of faces would be challenging for many reasons - the variation in proportions and resolution of images, alongside the presence of facial expressions amongst them. Additionally, this type of facial recognition has several applications, including security, entertainment, and virtual reality. In particular, we will be designing an algorithm to predict the age, gender and ethnicity of an individual based solely on a headshot image of them.

We were drawn to this particular application by its seeming abstractness and challenge. Age and gender prediction have significant applications in advertising and demographic collection. Photos can be digitally analysed to produce data that traditionally would not be possible to produce, as humans are unable to comb through and identify possible thousands of pictures for the creation of data. This data can be further used for targeted advertisement profiling. Not only this, but age prediction can be utilised for the protection of minors in online settings,

Convolutional Neural Networks (CNN's) are a class of algorithms that are incredibly powerful for tasks involving image analysis, object detection and image classification. They are designed to learn features from raw data, reducing the need for manual feature engineering and significantly improving the efficiency and effectiveness of various visual tasks. CNNs consist of convolution layers, pooling layers and fully connected layers. Convolution and pooling layers are responsible for feature extraction, whilst the fully connected layers conjoin the features together into a singular output. We decided CNN's feature extraction would be perfect for identifying facial features and mapping them together to produce accurate estimations of age, ethnicity and gender. CNNs are also fairly independent, not requiring human input to aid feature extraction, as CNNs are capable of producing their own filters. Thus, for this application of age, gender and ethnicity prediction using images, it was determined that a Convolutional Neural Network (CNN's) would be the optimal algorithm.

2. Literature Review

Convolutional Neural Networks (CNN)

Convolutional neural Networks (CNNs) are a class of machine learning algorithms which are inspired from Artificial Neural Network (ANN). CNNs have had many successful applications over the years, including AlexNet, VGG-16, GoogLe Net, ResNet and more, CNNs are a class of computer vision capable of learning spatial hierarchies through the utilisation of backpropagation and a multitude of layers, including convolution layers, pooling layers and fully connected layers. The Convolution Layer is the core building block of a CNN. It involves applying a convolution operation using a kernel (also known as a filter) to the input image. The kernel slides over the input image, performing element-wise multiplication and then summing the results to produce a feature map. These feature maps capture different aspects of the input data, such as edges, textures, or more complex patterns. After one or more convolution layers, pooling layers are introduced to reduce the spatial dimensions of the feature maps. Pooling involves selecting representative values from a local region (e.g., max pooling selects the maximum value). This helps to decrease the computational burden and makes the network more robust to slight variations in the input. Once the features have been extracted and downsized through convolution and pooling layers, they are flattened and passed through fully connected layers. These layers are similar to those in a traditional neural network, where each neuron is connected to every neuron in the previous and following layers. The fully connected layers learn to make decisions based on the extracted features and provide the final output.

Age, gender and Ethnicity prediction using CNNs

Age, Gender and Ethnicity prediction through the use of computer vision lays the foundation for several different applications. For instance, this technology could be applied to video files to create identification systems for CCTV systems, or as a means of gaining consumer information for the purpose of consumer profiling. The age detection system can also be utilised as a means of age-locking certain digital content.

The study by Ionescu explores the use of a ‘Bag of Visual Words’ model to identify facial expressions. The Bag of Words model is generally used on text, to break down sentences into “presence vectors”, counting the frequency of unique words. This can also be applied in a visual context - Radu et al utilised the bag of words approach to detect facial expressions in low resolution images. A grid is placed over each image, and within each box a SIFT descriptor is generated (e.g. cheek dimples, Duchenne markers). The presence of these visual words is used to create a histogram, and spatial presence vectors are used to identify the positioning of these visual words. The program was trained against 28,709 examples, and tested upon two sets of 3589 photographs. The “nearest neighbours” of the test images were identified, by comparing vectors and identifying the images with the greatest amount of similarities. This was then used to determine the emotion being expressed in the test picture.

The study by Srivanas explores the usage of CNNs for Age, Gender and Ethnicity detections, specifically using the east-Asian dataset. This algorithm used two parallel approaches. One network would observe the image as a whole whilst the other would break the image down into image regions. Features were identified from the image using convolutional filters, and this information was inputted into a series of fully connected layers providing class categories. However, this study did not find much success in results, with ethnicity being by far the hardest of the 3 parameters for the computer to gauge. The algorithm often repeatedly returned the answer that made up the majority of the dataset.

The study by Nyaupane similarly approached age, gender and ethnicity prediction with the use of deep separable convolutional neural networks. Utilising the UTKFace dataset of 23,705 images, the study was able to produce impressive results, with accuracies of 79% and 89% for ethnicity and gender respectively. The study utilised 3 separate models, with the 3rd model performing the best by far, with the researchers adjusting the hyperparameters between each model.

The study by Shoeran explores only age and gender through the same UTKFace dataset, also utilising CNNs. In addition, the study uses transfer learning to develop pre-trained models for the dataset, achieving impressive results of 95% accuracy when predicting gender. However, the model struggled to predict the age of individuals over 70, and the researchers theorised this was due to the scale and lack of balance in the dataset utilised.

The study by Yamashita explores the applications of CNN’s in the medical field, specifically in radiology and medical imaging. Yamashita experimented with using CNN’s to produce diagnoses based upon CT scans, creating a classification system. CNN’s utilised convolutional and pooling layers to perform a process of segmentation, isolating relevant organs for further inspection. The study mentions the struggle of acquiring data for training, due to the absence of large public databases of medical images, and the ethical issues alongside creating one.

3. Methodology

Data

For this study, we used the Age, Gender and Ethnicity dataset on kaggle.com - this dataset consisted of 23,705 images of people’s faces, across 5 different ethnicities, 2 genders and 16 age ranges. The data was slightly male dominant, with the age distribution being heavily concentrated between the ages 20 and 40. The data also contains a majority of white individuals. All of this can be seen in Figures 1-3

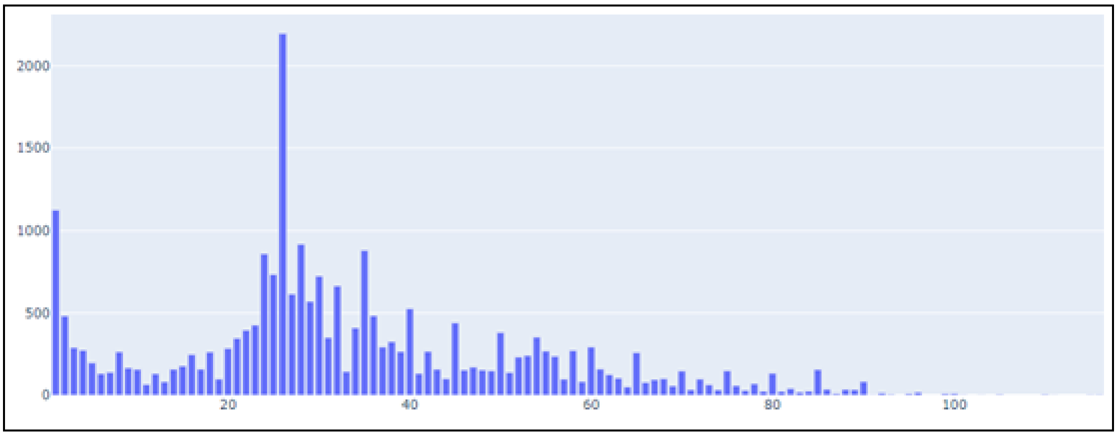


Figure 1: Age distribution of all the individuals in the dataset

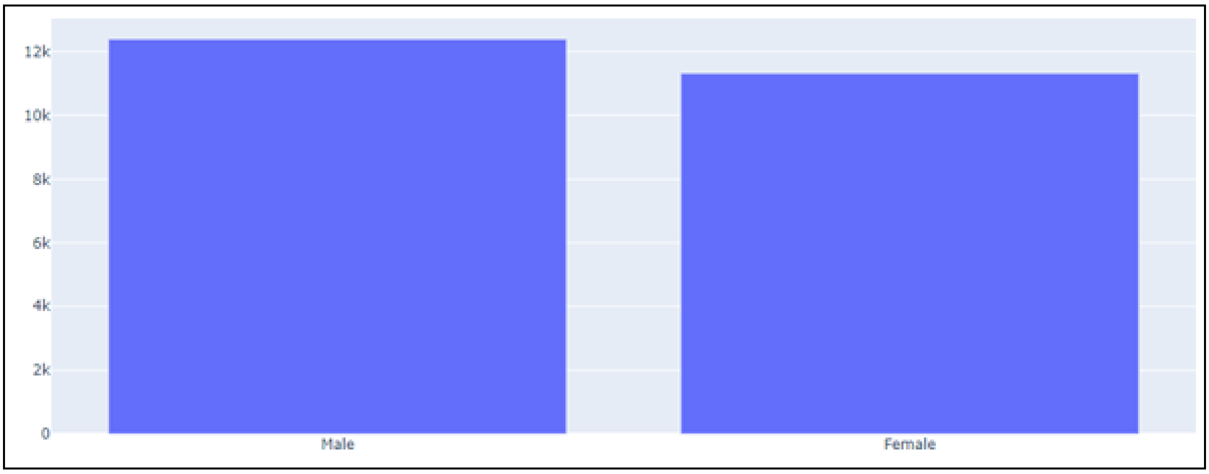


Figure 2: Gender Distribution

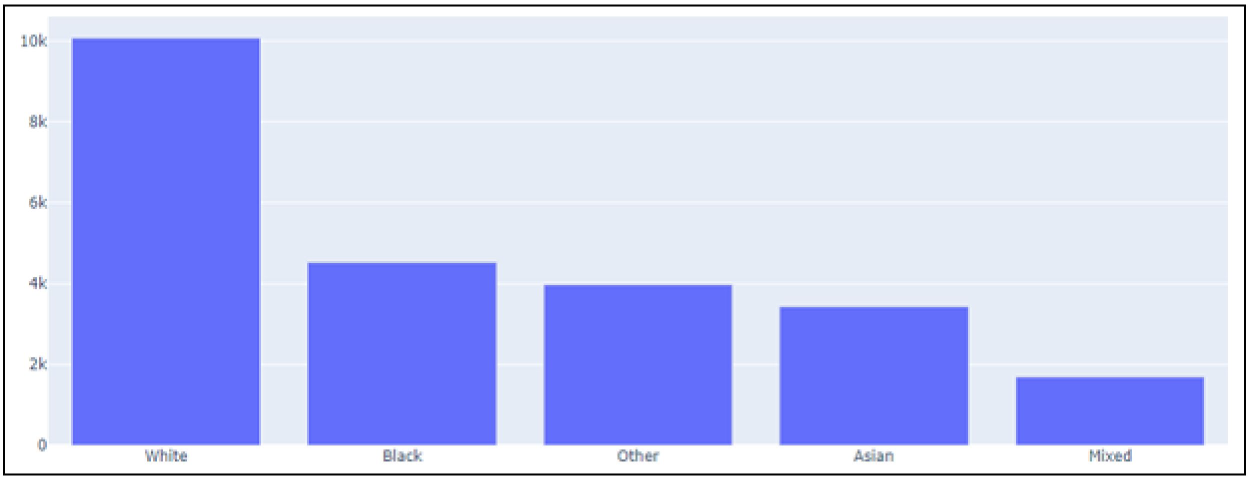


Figure 3: Ethnicity Distribution

The images were split in a ratio of 70:30, with 70% of the images being utilised for training, and the other 30% for training. The images were inconsistent in size and scaling, so all images were resized to 48x48 pixels. To reduce the computation time of the model, all the images were converted to grayscale by dividing the RGB values by 255.

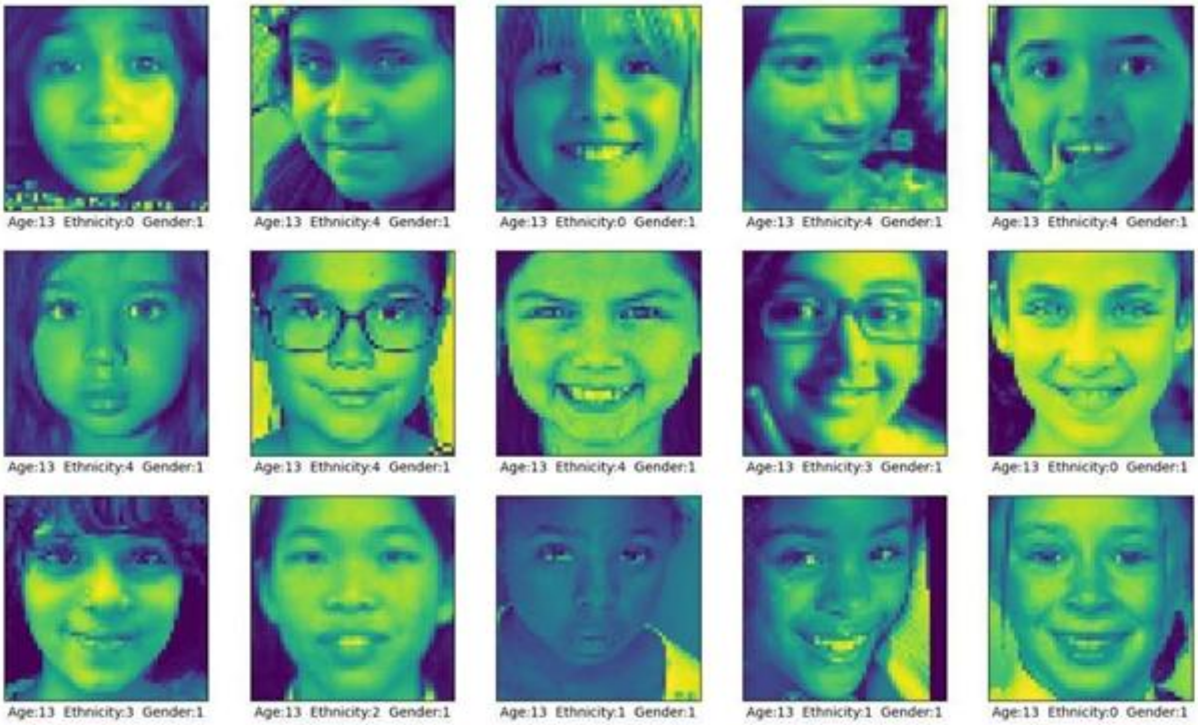

Age:13 Ethnicity:0 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:0 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:4 Gender:1
Age:13 Ethnicity:3 Gender:1
Age:13 Ethnicity:0 Gender:1
Age:13 Ethnicity:3 Gender:1
Age:13 Ethnicity:2 Gender:1
Age:13 Ethnicity:1 Gender:1
Age:13 Ethnicity:1 Gender:1
Age:13 Ethnicity:0 Gender:1

Figure 4: Photo samples

4. Model

For this model, we utilised Google's Tensorflow, specifically a python interface for tensorflow named 'keras'. 10% of the training images were used as validation images, to avoid overfitting. First the model ran a convolutional layer of size 32, with a filter size of 3x3. After the batch was normalised, we applied pooling of size 2x2. The activation type used for all convolutional layers was 'Rectified linear unit' (ReLU). Two more pairs of convolution and pooling layers were applied, with respective sizes of 48 and 64. The resultant output was flattened, and ran through 2 fully connected dense layers, of respective sizes 64 (ReLU) and 1 (Sigmoid)., with a dropout rate of 0.5 between the two.

The model was compiled with the stochastic gradient descent algorithm (SGD), which utilises backpropagation to continuously update the weights of the fully connected layers to increase accuracy.

The model splits the images into 20 separate epochs, with backpropagation performed after the completion of each epoch. To avoid overfitting, the program runs until either all 20 epochs are completed or the val_loss dips below 0.2700. The val_loss value represents the error rate of the model when attempting to predict the age/gender/ethnicity of the validation images - all of which the model is never trained on.

Experiment

- Images are inputted, and are all of size 48x48 due to resizing applied earlier
- The images are passed on to the first convolution layer, which identifies the prominent features of the images using 3x3 filters/kernels
- The images are then passed on to the MaxPooling layer, which reduces the image dimensions to increase the speed of further computation.
- The images are run through two more pairs of convolution & pooling layers, before being passed to Fully Connected (FC) layers.
- The FC layers learn to predict the age/gender/ethnicity of the individual in the image, and then based on the error perform back propagation to adjust weights and improve predictive accuracy
- The algorithm runs for each epoch until either all 20 are run through, or a val_loss of below 0.2700 is achieved for a certain epoch.

5. Results

Gender Results

In our gender prediction model we experimented with different convolutional layers ranging between 3 and 5 We performed a validation of these models to observe any overfitting, and found that 5 convolutional layers caused overfitting. We were able to achieve the best results with 4 convolutional layers. Our model was run over 20 epochs, and the training loss graph can be seen below in Figure 5.

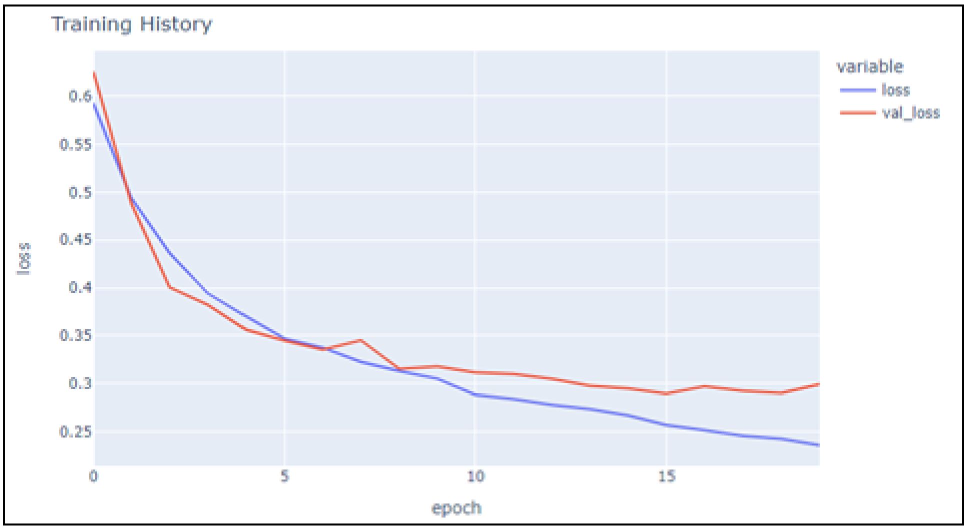
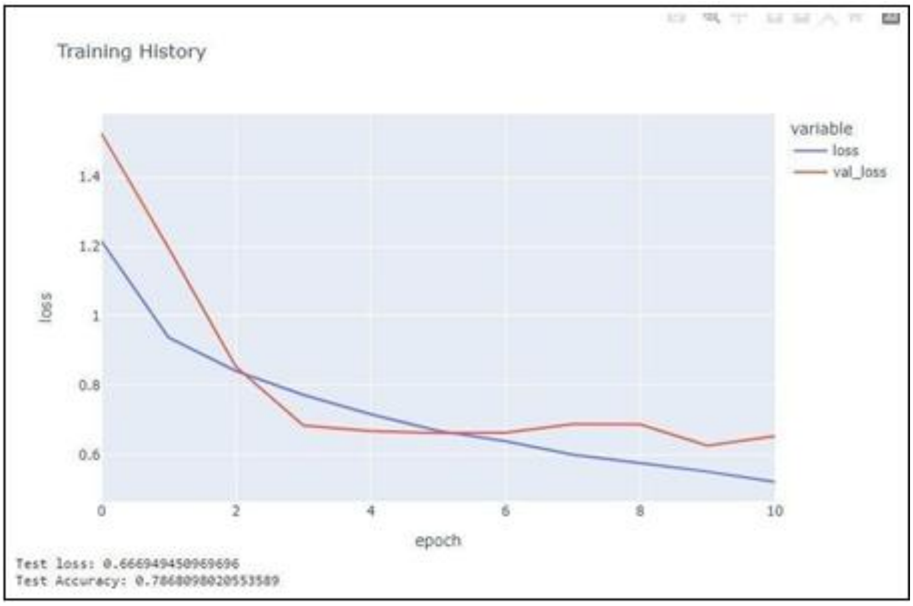


Figure 5: Gender Training History

Our best results were achieved with 4 convolutional layers, with a test accuracy of 88.1%.

Ethnicity Results

In our ethnicity prediction model, we experimented with different dropout rates between 0.1 and 0.6. We found that 0.5 provided the best results. We also briefly experimented with varying quantities of convolutional layers but quickly found that more than 2 convolutional layers caused overfitting. To avoid overfitting in later epochs, the model stopped training as val_accuracy reached 79%. On testing, the model achieved a test accuracy of 78.7%.



```
Test loss: 0.666949450969696
Test Accuracy: 0.7868098020553589
```

Figure 6: Ethnicity Training History

Age Results

For age, it was required that all ages be separated into categories. All images were split into 13 groupings by age - <5, <10, <15, <20, <30, <40, <50, <60, <70, <80, <90, <100, <150. Experimentation with convolutional layers and normalisation failed to produce a highly successful model, with performance peaking at 50% accuracy. We decided this was likely due to a large quantity of age categories, with distinctions between 20/30 year olds or 5/10 year olds being too challenging for the computer to make. To combat this, the groupings were changed to be <5, <20, <40, <60, <80 and >=80. With this alteration we were able to achieve a Test Accuracy of 73.2%.

5. Conclusion

This research paper explored some previous uses of CNN to identify age, gender and ethnicity, and then produced a model capable of performing the same task with high accuracy. The model developed is capable of detecting the Age, Gender and Ethnicity of an individual based solely on a 48x48 image of their face. It's able to do this with accuracies of 73.2%, 88.1% and 78.7% respectively. We were able to achieve these results through variation in the quantity of convolutional and pooling layers, the dropout rates, and the categorization of ages into groups.

References

[1] Srinivas, N., Atwal, H., Rose, D. C., Mahalingam, G., Ricanek, K., & Bolme, D. S. (2017, May). Age,

[2] Gender, and Fine-Grained Ethnicity Prediction Using Convolutional Neural Networks for the East

[3] Asian Face Dataset. 2017 12th IEEE International Conference on Automatic Face & Gesture

[4] Recognition (FG 2017). https://doi.org/10.1109/fg.2017.118

[5] Sheoran, V., Joshi, S., & Bhayani, T. R. (2021). Age and gender prediction using deep cnns and transfer learning. In Computer Vision and Image Processing: 5th International Conference, CVIP

[6] 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5 (pp. 293- 8) 304). Springer Singapore.

[7] Nyaupane, B. K., & Shakya, S. (2022, March). Age, Gender, and Ethnicity Prediction using Deep

[8] Separable Convolutional Neural Networks. In SP Jain School of Global Management, International Multidisciplinary conference, Dubai.

[9] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems.

[10] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9, 611-629.

[11] Ionescu, R. T., Popescu, M., & Grozea, C. (2013, June). Local learning to improve bag of visual words model for facial expression recognition. In Workshop on challenges in representation learning, ICML.