Apparent Age Estimation with Relational Networks

Eu Wern Teh
School of Engineering, University of Guelph
Vector Institute for Artificial Intelligence
eteh@uoguelph.ca

Graham W. Taylor
School of Engineering, University of Guelph
Vector Institute for Artificial Intelligence
gwtaylor@uoguelph.ca

***Abstract*—Apparent age estimation is a newly proposed and under-studied problem of predicting the age that someone "looks" rather than their actual age. It has applications in many areas within the beauty industry[3]. Methods based on convolutional neural networks (CNNs) have proved to be state-of-the-art on the few datasets used to benchmark this task. However, such CNNs typically collapse spatial information via a Global Average Pooling operation. They do not perform any explicit treatment of spatial relationships of the higher-level features which emerge in the later stages of the network and which may correspond to facial parts or blemishes that are characteristic of age. In this paper, we consider a newly proposed CNN module called relational networks that explicitly capture spatial relationships. We hypothesize that we can estimate age better by learning such relationships in the final set of CNN feature maps where spatial information is still retained. Experiments were conducted on both ChaLearn LAP 2015 and 2016 datasets [6], [7] showing that on average, there is a 3.53% improvement on Mean Absolute Error and 3.31% improvement on ϵ-error when compared to the baseline. A test was also calculated to show that the improvement is statistically significant.**

***Keywords*-Relational Network; Apparent Age Estimation; Regression**

I. Introduction

Age estimation is historically a challenging problem in the field of computer vision. Several applications can benefit from predicting age, such as advanced video surveillance, collection of demographic statistics, and customer profiling [6]. In advanced video surveillance, age estimation can be used to query people of a certain age group in order to narrow down search results. In demographic statistics and customer profiling, age estimation allows for the determination of age groups of customers visiting a section of a shopping mall, for example. With that information, targeted advertising can be used to promote products of interest to shoppers of that age group.

A related but distinct problem known as apparent age estimation seeks to determine how old a person *looks*, instead of their actual age. It can be very useful for several applications, such as studying the effects of cosmetics, haircuts, plastic surgery, and anti-aging treatment. It is a challenging problem for several reasons: First, the data collection process is difficult. It is relatively easy to collect a person's actual age, rather than figuring out how old a person looks, as the latter depends on the collection of opinions from observers. Second, as the data is opinion-based, there will be variability in age labels from participating observers. The first problem requires the model to estimate the exact age given an image, but the later problem requires the model to estimate the distribution of age given an image. Figure 1 shows two examples of the difference between actual age and apparent age.

Rachel McAdams 2012
Actual age: 34, Apparent age: 31

Maria Sharapova 2008
Actual age: 21, Apparent age: 29

Figure 1. Two examples of the difference between actual age and apparent age. These images are taken from the ChaLearn LAP 2015 dataset and the actual ages are taken from Wikipedia.

The current state-of-the-art architecture on the ChaLearn LAP 2015 dataset, DEX [13], tackles the problem of apparent age estimation by using the entire face image to estimate apparent age. We hypothesize that we can better estimate age by learning the spatial relations between the final set of convolutional neural network (CNN) feature maps where spatial information is still retained. We propose the use of a recently proposed neural network module, the relational network [15], to learn these relations. Experimental results show that the proposed network is able to outperform the baseline consistently on both the ChaLearn LAP 2015 and 2016 datasets with and without the aid of external dataset.

Our paper is organized as follows: Section II discusses related work, Section III describes the architecture, and Section IV details the dataset, evaluation metrics, experimental design, and the results of our experiments. Lastly, Section V presents the final conclusions from these experiments.

II. Related Work

Over the years, the research community has created several datasets to tackle the problem of actual age estimation; these include the MORPH2 [12], FRGC [11], and HOIP [17] datasets. This has resulted in significant research in estimating actual age or actual group age. For example, Zhang et al. [20] try to quantify facial age using posterior age distribution. Xi et al. [18] use a multiple instance learning paradigm to tackle this problem. Wei-Lun et al. [2] tackled age estimation using distance metric learning. Wenjie et al. [10] propose an end-to-end spatial attention network to tackle age estimation on a video dataset [5].

On the other hand, there are very few publicly available datasets for apparent age estimation due to the nature of the problem. The ChaLearn LAP 2015 Apparent Age dataset is the first publicly available dataset for this task. This dataset was collected using a web-based game using the Facebook API. In the game, participants score points by guessing the apparent age, which is the mean age of all participants. Participants gain a higher score if their guess age is closer to the mean age. A leader board was added to increase participant engagement. There are a total of 110 male and 44 female participants in this dataset.

While this problem has received much less attention from the computer vision community than actual age estimation, the current state-of-the-art architecture for the ChaLearn LAP 2015 Apparent Age dataset is DEX [13] (Deep EXpectation of apparent age from a single image). The DEX network follows the VGG16 architecture [16], and its weights are initialized using an existing VGG16 network architecture pretrained on the ImageNet 2012 dataset [4]. The authors further fine-tune this network on an external dataset (IMDB-WIKI) before performing its final training on the ChaLearn dataset.

Our proposed solution is inspired by Santoro et al. [15], who propose a simple neural network module for relational reasoning tasks, namely Relational Networks. By concatenating each spatial feature with all other spatial features, the authors show that the network is able to discover and learn to reason about entities and their relations.

III. Proposed Approach

In this section, the proposed architecture is introduced. Our method uses the ResNet-50 architecture [8], which has shown impressive results on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [14]. An overview of the approach is illustrated in Figure 2.

A. Face Detection

The pipeline starts with face detection. An existing face detection model [19] is used to detect a face within a given image. If multiple faces are detected, the face with the largest area is used. In the experiments, it is observed that a few cases exist where the network fails to return faces. In such instances, the entire image is used as input. Most of these cases happen to be images with a very large face-to-image ratio.

B. Baseline Model

The ResNet-50 architecture is used as a baseline. The parameters were initialized with an existing ResNet-50 network pre-trained on the ILSVRC 2012 dataset. The last layer of ResNet-50 is discarded, and two parallel fully-connected layers were added. The first fully-connected layer has an input size of 2048 and an output size of 1. The second fully-connected layer has an input size of 2048 and an output size of 101. The first fully-connected layer is responsible for predicting age as a regression, and the second fully-connected layer is responsible for predicting age by treating it as a classification problem.

C. Relational Network

The proposed architecture is a modification of the baseline model. For the i^{th} input face image, $x_i \in \mathbb{R}^{224\times224}$, there is a corresponding activation output, $a_i \in \mathbb{R}^{2048\times7\times7}$ which is obtained from the output of the last Conv5 layer in ResNet-50 architecture. This activation output is then rearranged to be $a_{ij} \in \mathbb{R}^{2048\times49}$, where j corresponds to a feature vector at the j^{th} location of the activation output.

Next, the feature at each location j should interact with all other features k in the activation output. This is done by concatenating each and every feature vector with each other, which yields an output of $a_i^* \in \mathbb{R}^{4096\times49\times49}$. Global average pooling is then performed on a_i^* to yield an output of $a_i^{**} \in \mathbb{R}^{4096\times1}$:

$$\text{RN}(a_i) = \frac{1}{N}\sum_{j=1}^{49}\sum_{k=1}^{49}(a_j, a_k) = a_i^{**}. \tag{1}$$

Finally, a_i^{**} is passed to two fully-connected layers to generate two age outputs, as is done by the baseline. In detail, the first fully-connected layer has an input size of 4096 and an output size of 1, and the second fully-connected layer has as an input size of 4096 and an output size of 101.

Compared to the relational network proposed by Santoro et al. [15], our relational network differs in two ways: First, [15] concatenates text features (extracted from a natural language question) with the visual features from each concatenated region (a_j, a_k). The design choice is obvious as [15] attempts to solve relational reasoning problem, where the inputs are paired images and questions. Our method only uses the visual modality. Second, [15] used a shallow network with only four convolutional layers. However, our relational network uses a deep network (50 convolutional layers). Each feature grid in a shallow network has a smaller receptive field compared to the deep network counterpart.

The relational network allows the visual features to interact spatially with one another, hence giving it the ability

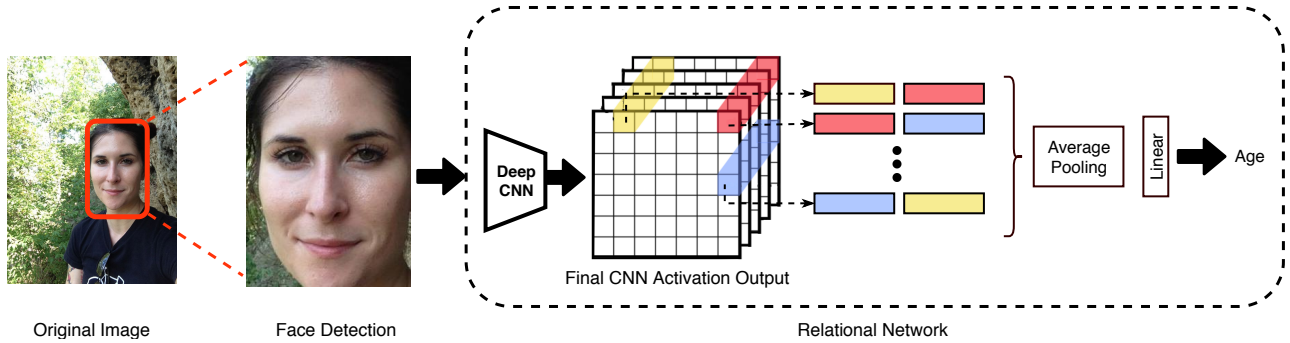

Figure 2. Our relational network architecture and pipeline. First, an existing face detection model is used to find a face in an image. Next, the image is cropped and passed to a relational network to estimate apparent age. Spatial features from each grid is represented by color blocks. The relational network allows the features to interact with one another spatially and this interaction is performed via concatenation. Next, average pooling is performed on the relational features before passing it to two parallel fully-connected layers, which output apparent age, apparent age class and expected apparent age values.

Error	Method	2015	2016	2015+	2016+	2015++	2016++	Mean
MAE	Baseline	3.949	5.216	3.593	4.721	3.527	4.401	
	Relational network	3.845	4.930	3.428	4.569	3.397	4.333	
	Difference	0.104	0.286	0.165	0.152	0.130	0.068	
	Improvement (%)	2.634	5.483	4.592	3.220	3.686	1.545	3.527
ϵ-Error	Baseline	0.373	0.363	0.331	0.390	0.326	0.357	
	Relational network	0.354	0.390	0.313	0.380	0.313	0.351	
	Difference	0.009	0.015	0.018	0.006	0.013	0.006	
	Improvement (%)	2.478	3.704	5.438	2.564	3.988	1.681	3.309

Table I
Improvement summary of our relational network over the baseline on both MAE and ϵ-Error. + means training includes the WIKI dataset. ++ means training includes the WIKI and IMDB dataset.

to uncover spatial relationships among the features. The explicit feature concatenation of each grid cell enforces pairwise relationship to be considered between each region. We hypothesize that learning these relations will yield improved results in apparent age estimation as it can more explicitly capture facial (a)symmetries which may develop over time or be perceived as age-relevant.

D. Age Estimation and Loss Functions

At the later stages of the CNNs, there are two parallel fully-connected layers. The first fully-connected layer outputs an age value, while the second fully-connected layer outputs a probability scores of 101 continuous classes. Similar to DEX [13], we also compute an expected apparent age E from the second fully-connected layer, as follows:

$$E(o) = \sum_{c=0}^{100} y_c o_c, \tag{2}$$

where $o = 0, 1, ..., 100$ is the 101 dimensional output layer, representing softmax output probabilities $o_c \in O$, and y_c are the discrete years corresponding to each class c.

We penalize both the regression output and expected value with mean squared error loss. At the same time, we penalize the classification output with cross entropy loss. All losses are penalized equally during training and the expected value is used as the final prediction at test time.

Error	Method	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Mean	Std
MAE	RN1	3.967	3.834	4.099	3.880	3.967	3.949	0.101
	RN2	**3.806**	**3.850**	**3.883**	**3.883**	**3.776**	**3.840**	**0.048**
ϵ-Error	RN1	0.359	0.352	0.377	0.356	0.364	0.362	0.010
	RN2	**0.351**	**0.352**	**0.360**	**0.358**	**0.347**	**0.354**	**0.005**

Table II
MAE and ϵ-Error results of experiments on the ChaLearn LAP 2015 dataset. RN1 is a relational network model that outputs apparent age class and expected apparent age. RN2 is our proposed relational model where it outputs apparent age, apparent age class and expected apparent age. The experiment was run 5 times with different random seeds, and we report the mean and standard deviation for each method.

Our architecture is slightly different from DEX as our models output apparent age and apparent age class as well as expected apparent age, while DEX only computes apparent age class and expected apparent age. We find that having a separate fully-connected layer to regress the apparent age value helps to improve the overall estimation because it acts as a form of regularization, which allows the network to estimate age independently from the classification layer. At test time, we take expected apparent age to be our final decision. This result is illustrated in Table II.

IMDB	WIKI	Total	
461,871	62,359	524,230	Original
386,316	42,119	428,435	Our models
N/A	N/A	260,282	DEX [13]

Table III
Number of images used in the IMDB and WIKI datasets to pre-train our models. The reason for the discrepancy with [13] was that the authors did not provide details on the subset of data used for their experiments. Details of how we selected images to include are provide in the main text, but we speculate that more careful curation of the data may lead to better results for our model.

IV. Experiments

This section describes the dataset, experimental setup, evaluation metrics, and experimental results. The improvement of our relational network-based approach over the DEX-style CNN baseline is summarized in Table I.

A. ChaLearn LAP 2015 and 2016 Apparent Age dataset

The models were trained using the ChaLearn LAP 2015 and 2016 Apparent Age datasets [6], [7]. The ChaLearn LAP 2015 dataset consists of 4,699 images, which are split into 2,476 training images, 1,136 validation images, and 1,087 test images. The ChaLearn LAP 2016 dataset consists of 7,591 images, which are split into 4,113 training images, 1,500 validation images, and 1,978 test images. Figure 3 shows the age distribution for both datasets.

B. IMDB and WIKI dataset

Inspired by [13], we pre-trained our models on the actual age estimation task using the IMDB and WIKI age dataset. This is an actual age dataset collected by [13] and consists of 542,230 images. Only images where faces are found by the face detector were used, and the dataset was further pruned by discarding images with a height or width smaller than 32 pixels. With that criteria, 90% of the images were used for pre-training. See Table III for details.

C. Experimental Details

In all experiments, the weights of the model were initialized with ResNet-50 model pre-trained ResNet-50 on the ILSVRC 2012 dataset. The learning rate started at 10^{-4}, and the RMSprop [9] optimizer was used to train all models. The learning rate was reduced by 10^{-1} at epoch 10 and 15 and training was stopped at epoch 20. All hyperparameters were selected based on the given validation set of the respective dataset. Next, we use both the training and validation set to train our final models. During training, the input images were resized to 256×256 and a random cropping of 224×224 was applied. Horizontal flipping was also used to augment the data.

For experiments with the external actual age estimation datasets (IMDB and WIKI), the models were first trained on the IMDB and WIKI dataset for 20 epochs. During training, the learning rate was reduced at epoch 10 and 15. After that, the same models were further trained on the ChaLearn LAP 2015 and 2016 datasets respectively for another 20 epochs, where the learning rate was also reduced at epoch 10 and 15.

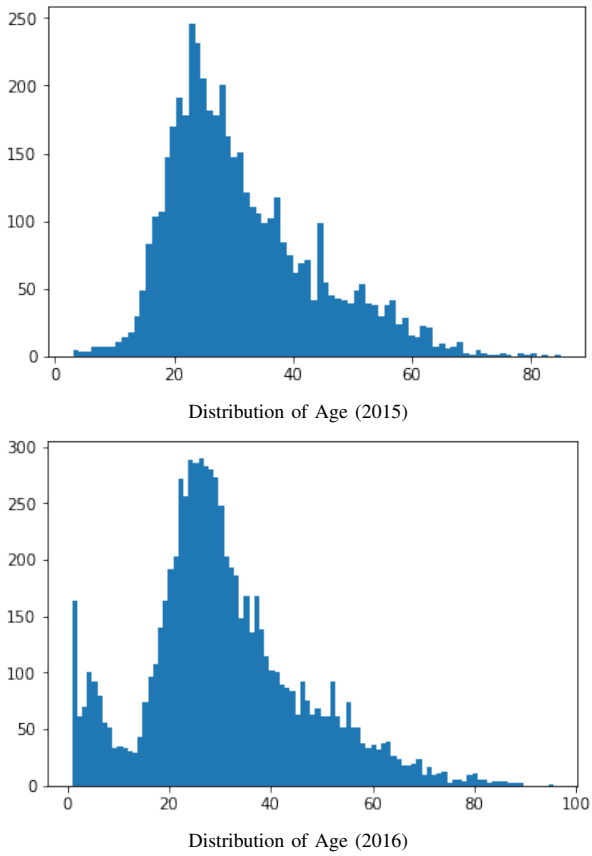

Figure 3. Age distribution for the ChaLearn LAP 2015 and 2016 datasets, respectively.

Experiments were implemented in Python using the PyTorch framework, and experiments were performed on a NVIDIA GTX 1080 Ti with 1480 MHz core speed. The training time for Relational network (+WIKI) and Relational network (+IMDB, WIKI) are 7.5 hours and 57 hours respectively. The test time on Relational network is 13ms per image. To be consistent with [13], [1], experiments were also performed on a NVIDIA Tesla K40c. The training and testing time of the relational model are shown in Table IV.

D. Evaluation Metrics

We evaluated each model using the protocol and metrics established by [13]. These metrics include the Mean Absolute Error (MAE) and ϵ-Error, and are described in Equations 3 and 4, respectively, where given an image i,

	Training	Testing
Relational network (+WIKI)	**15 hrs**	**38 ms/img**
Relational network (+IMDB,WIKI)	133 hrs	**38 ms/img**
DEX [13] (+IMDB, WIKI)	123 hrs	200 ms/img
OrangeLab [1] (+IMDB, WIKI, Private data)	44 hrs	6.34 s/img

Table IV
Training and testing time of our relational network trained on a NVIDIA Tesla K40c. We compare our results with DEX and OrangeLab. From Table V, there is a drop of 6% in MAE when we compare Relational Network (+WIKI) with DEX but there are over 8× and 5× speed improvement on train and test time with respect to DEX.

x_i is the predicted mean age, μ_i is the mean ground truth age, and σ_i is the standard deviation of the ground truth age. Under the ϵ-Error metric, apparent age estimation error is penalized less on examples with a larger σ_i (i.e. on examples on which the human votes are far apart from each other) and vice versa.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu_i|, \tag{3}$$

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N} 1 - \exp(\frac{-(x_i-\mu_i)^2}{2*\sigma_i^2}). \tag{4}$$

Method	MAE	ϵ-Error
Baseline	3.95	0.363
Relational network	3.84	0.354
DEX [13]	5.01*	0.431*
Baseline (+WIKI)	3.59	0.331
Relational network (+WIKI)	3.43	0.313
Baseline (+IMDB, WIKI)	3.53	0.326
Relational network (+IMDB, WIKI)	3.40	0.313
DEX [13] (+IMDB,WIKI)	**3.22***	**0.265**

Table V
Results on the ChaLearn LAP 2015 dataset. * indicates evaluation obtained from validation set.

Method	MAE	ϵ-Error
Baseline	5.22	0.405
Relational network	4.90	0.389
Baseline (+WIKI)	4.72	0.380
Relational network (+WIKI)	4.57	0.366
Baseline (+IMDB, WIKI)	4.40	0.357
Relational network (+IMDB, WIKI)	**4.33**	0.351
OrangeLab [1] (+IMDB,WIKI, Private children dataset)	N/A	**0.241**

Table VI
Results on the ChaLearn LAP 2016 dataset.

Error	Method	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Mean	Std
MAE	Baseline	3.825	3.892	4.051	3.987	3.991	3.949	0.090
	Relational network	**3.806**	**3.850**	**3.883**	**3.883**	**3.776**	**3.840**	**0.048**
ϵ-Error	Baseline	**0.351**	0.356	0.372	0.371	0.366	0.363	0.009
	Relational network	**0.351**	**0.352**	**0.360**	**0.358**	**0.347**	**0.354**	**0.005**

Table VII
MAE ϵ-Error results of experiments on the ChaLearn LAP 2015 dataset. The experiment was run 5 times with different random seeds, and we report the mean and standard deviation for each method.

Error	Method	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Mean	Std
MAE	Baseline	5.085	5.083	5.392	5.363	5.156	5.216	0.151
	Relational network	**4.981**	**4.950**	**4.874**	**4.766**	**4.930**	**4.900**	**0.085**
ϵ-Error	Baseline	0.402	0.397	0.408	0.416	0.400	0.405	0.008
	Relational network	**0.390**	**0.390**	**0.389**	**0.384**	**0.391**	**0.389**	**0.003**

Table VIII
MAE and ϵ-Error results of experiments on the ChaLearn LAP 2016 dataset. The experiment was run 5 times with different random seeds, and we report the mean and standard deviation for each method.

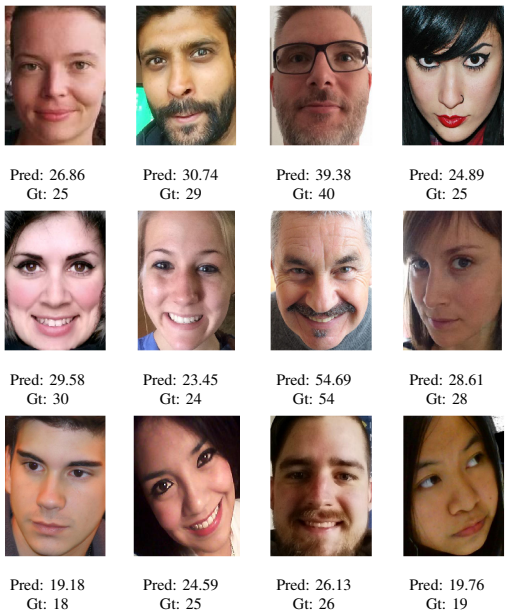

Figure 4. Qualitative results: examples of images that were accurately estimated by our approach, taken from the ChaLearn LAP 2015 dataset. The model was not pre-trained with an external dataset.

E. Results

Table V shows the results of experiments on the ChaLearn LAP 2015 dataset. The relational network consistently outperforms the baseline that does not use explicit spatial reasoning, with or without the external IMDB and WIKI datasets. The performance of the state-of-the-art architecture DEX [13] is included for comparison; however, it is not a fair comparison, as DEX uses an ensemble of 20 CNN models. The baseline and relational network in this work are able

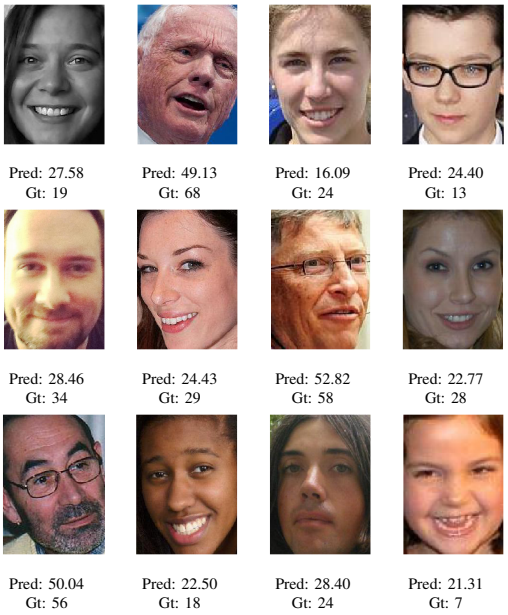

Figure 5. Qualitative results: examples of images that were estimated far from the ground truth, taken from the ChaLearn LAP 2015 dataset. The model was not pre-trained with an external dataset.

to outperform DEX without the external IMDB and WIKI dataset, while being about 8× faster at test time. However, it narrowly loses to DEX with the external dataset. One possible reason for this is because of subtle differences in the particular subsets of images from these datasets used to pre-train each model. In their paper, Rothe et al. [13] mention that they only use a subset of the IMDB and WIKI dataset to pre-train their model, but do not mention the specific selection criteria.

Table VI shows the results of experiments on the ChaLearn LAP 2016 dataset. Similarly, the relational network consistently outperforms the baseline with or without the external IMDB and WIKI dataset. The state-of-the-art architecture OrangeLab [1] is also included for comparison, but this is also not a fair comparison for several reasons: (1) OrangeLab uses an ensemble of 14 CNN models; (2) the model was pre-trained on a privately collected and unreleased dataset of children; and (3) they use a pre-trained model from the VGG Face dataset as a starting point, which gives them an unfair advantage when comparing to a pre-trained model from ImageNet.

From Table V and Table VI, we observe that there is a huge improvement on the performance of the models when the WIKI dataset is used for pre-training, even though it consists of only 42,119 images. However, there is only very little incremental in performance when both IMDB and WIKI dataset are used for pre-training.

Figures 4 and 5 show the qualitative results of the relational network on the ChaLearn LAP 2015 dataset. Table VII and Table VIII show the results of experiments on

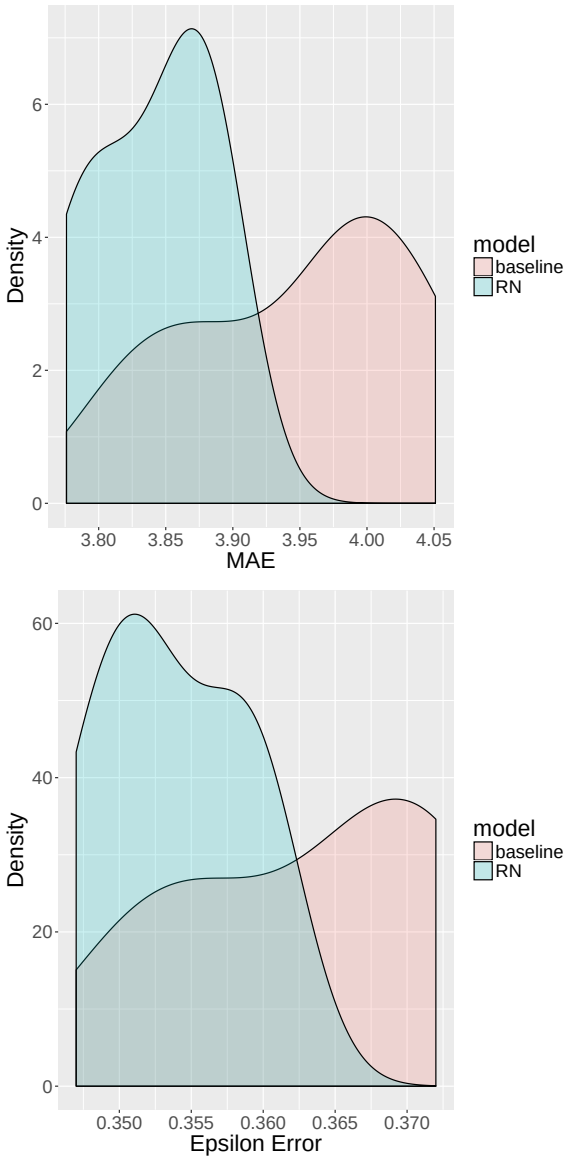

Figure 6. Kernel density estimate of MAE and ϵ-Error on five runs of our experiments on the ChaLearn LAP 2015 dataset.

five different splits. Figures 6 and 7 show the distribution of MAE and ϵ-Error for both the baseline and relational network on the ChaLearn LAP 2015 and 2016 datasets. We use the Shapiro-Wilk test to perform normality testing and all results pass the this test, so a t-test was used for significance testing.

For the ChaLearn LAP 2015 dataset, a p-value of 0.041 for MAE and a p-value of 0.047 for ϵ-Error were found. For the ChaLearn LAP 2016 dataset, a p-value of 0.036 for MAE and a p-value of 0.025 for ϵ-Error were found. All cases pass the standard 95% significance testing criteria,

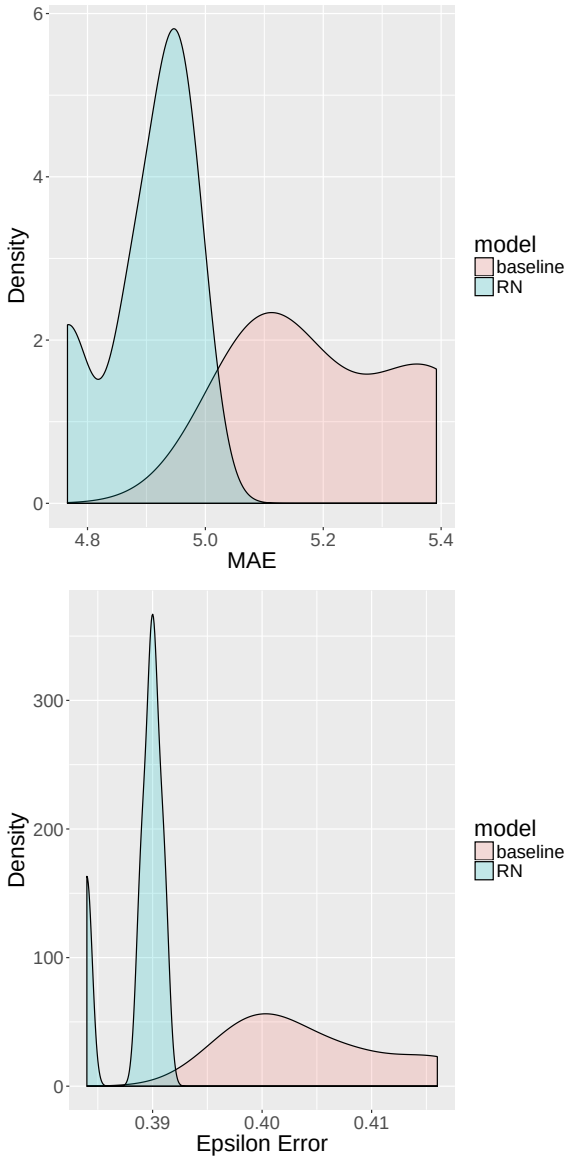

Figure 7. Kernel density estimate of MAE and ϵ-Error on five runs of our experiments on ChaLearn LAP 2016 dataset.

which demonstrates that the improvement from the relational network over the baseline is significant.

Lastly, Figure 8 shows a heat map visualization of the paired grid cells' activation scores in the relational network. The brightness of the heat map corresponds to the level of excitement of the grid cells when they are paired with one another. The brighter the color, the higher the interaction between the grids. In addition, the heat map also highlights the contribution of the paired grid cells' in estimating apparent age.

V. Conclusion

We have shown that a simple module, the relational network, can augment CNN performance on the apparent age estimation task. Our proposed architecture outperformed the baseline model on the ChaLearn LAP 2015 and 2016 datasets with and without the use of an external dataset. On average, there is a 3.53% improvement on MAE Error and a 3.31% improvement on ϵ-Error which were shown to be statistically significant. It achieved near state-of-the-art performance on ChaLearn LAP 2015 while improving test time speed by $8\times$ over the current state-of-the-art, DEX [13]. These improvements support our hypothesis that better apparent age estimation can be achieved by learning the spatial relations between high-level features in a CNN.

References

[1] 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*, pages 96–104, 2016.

[2] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[3] 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)*, volume 2, pages 182–190. IEEE, 2014.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012.

[6] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[7] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning-lecture 6a-overview of mini-batch gradient descent.

[10] W. Pei, H. Dibeklioğlu, T. Baltrušaitis, and D. M. Tax. Attended end-to-end architecture for age estimation from facial expression videos. *arXiv preprint arXiv:1711.08690*, 2017.

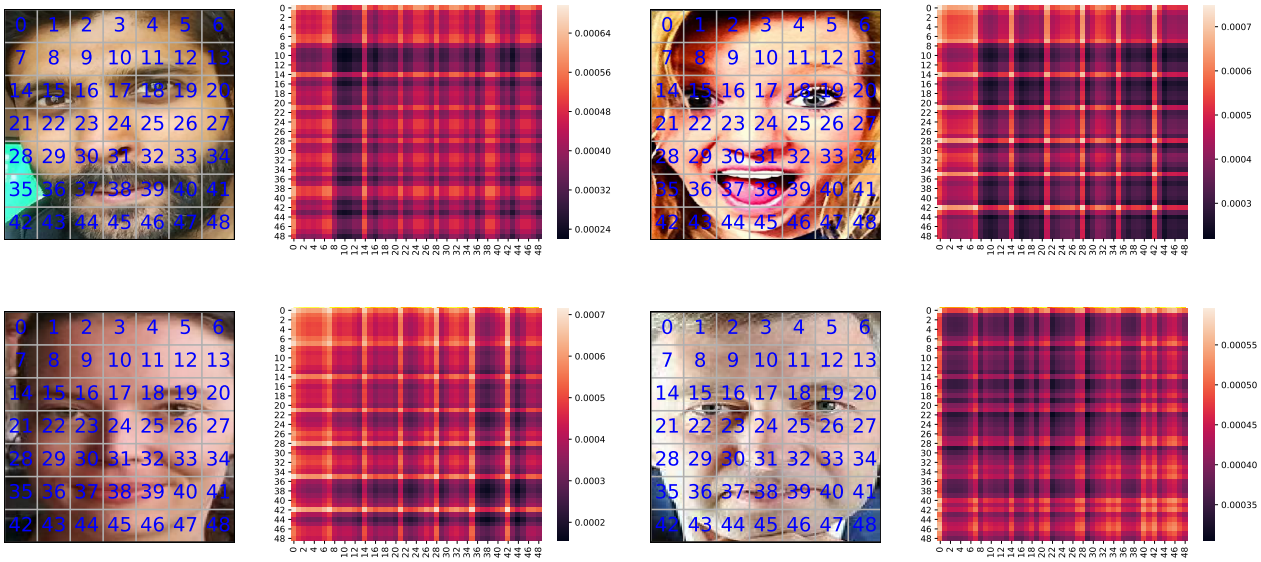

Figure 8. Activation scores of paired grid cells in the relational network for some example images taken from the ChaLearn LAP 2015 dataset. The brightness of the heatmap corresponds to the interaction of a grid cell when paired with another grid cell. The brighter the color, the higher the interaction. A partial interpretation of the heatmaps are as follows: (**Top Left**) The lower forehead (grid 8 to 12) of this image has very little interaction with the rest of the face. (**Top Right**) In this image, the upper forhead area (grid 0 to 6) has high interactions with the nose area (grid 21 to 34). (**Bottom Left**) The lips and chin area (grid 35 to 48) in this image have little interaction with the rest of the face. (**Bottom Right**) The chin area (grid 42 to 48) in this image has high interaction with the rest of face, especially those grids that are closer to it.

[11] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.

[12] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[13] 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*, pages 10–15, 2015.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge.

[15] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4974–4983, 2017.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. Yamamoto and Y. Niwa. Human and object interaction processing (hoip) project. pages 379–384, 01 2002.

[18] X. Yang, J. Liu, Y. Ma, and J. Xue. Facial age estimation from web photos using multiple-instance learning. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[20] Y. Zhang, L. Liu, C. Li, and C. C. Loy. Quantifying facial age by posterior of age comparisons. In *British Machine Vision Conference (BMVC)*, 2017.