DEX: Deep EXpectation of apparent age from a single image

Rasmus Rothe, Radu Timofte, Luc Van Gool
Computer Vision Lab, D-ITET, ETH Zurich, Switzerland
{rrothe,timofter,vangool}@vision.ee.ethz.ch

Abstract

In this paper we tackle the estimation of apparent age in still face images with deep learning. Our convolutional neural networks (CNNs) use the VGG-16 architecture [13] and are pretrained on ImageNet for image classification. In addition, due to the limited number of apparent age annotated images, we explore the benefit of finetuning over crawled Internet face images with available age. We crawled 0.5 million images of celebrities from IMDB and Wikipedia that we make public. This is the largest public dataset for age prediction to date. We pose the age regression problem as a deep classification problem followed by a softmax expected value refinement and show improvements over direct regression training of CNNs. Our proposed method, Deep EXpectation (DEX) of apparent age, first detects the face in the test image and then extracts the CNN predictions from an ensemble of 20 networks on the cropped face. The CNNs of DEX were finetuned on the crawled images and then on the provided images with apparent age annotations. DEX does not use explicit facial landmarks. Our DEX is the winner (1st place) of the ChaLearn LAP 2015 challenge on apparent age estimation with 115 registered teams, significantly outperforming the human reference.

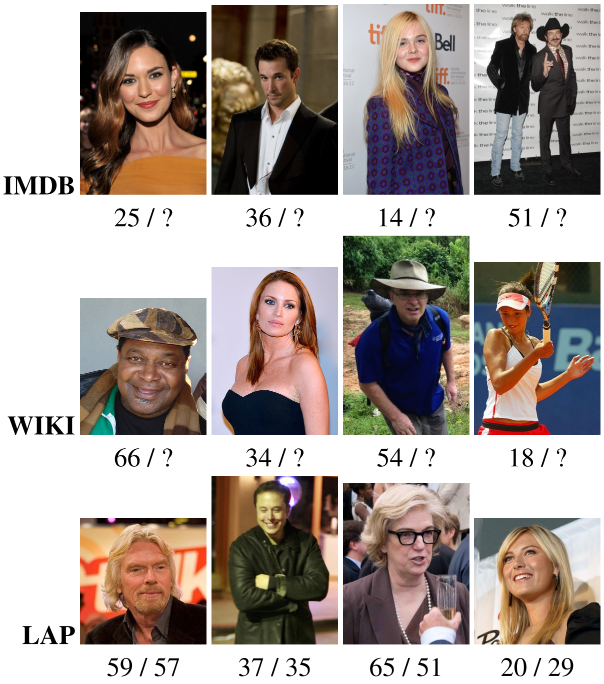


Figure 1. Real / Apparent (age)

1. Introduction

There are numerous studies [1, 3, 6] and several large datasets [1, 9, 11] on the (biological, real) age estimation based on a single face image. In contrast, the estimation of the apparent age, that is the age as perceived by other humans, is still at the beginning. The organizers of ChaLearn Looking At People 2015 [4] provided one of the largest datasets known to date of images with apparent age annotations (called here LAP dataset) and challenged the vision community.

The goal of this work is to study the apparent age estimation starting from single face images and by means of deep learning. Our choice is motivated by the recent advances in fields such as image classification [2, 8, 12] or object detection [5] fueled by deep learning.

Our convolutional neural networks (CNNs) use the VGG-16 architecture [13] and are pretrained on ImageNet [12] for image classification. In this way we benefit from the representation learned to discriminate object categories from images. As our experiments showed, this representation is not capable of good age estimation. Finetuning the CNN on training images with apparent age annotations is a necessary step to benefit from the representation power of the CNN. Due to the scarcity of face images with apparent age annotation, we explore the benefit of finetuning over crawled Internet face images with available (biological, real) age. The 524,230 face images crawled from IMDB and Wikipedia websites form our new dataset, the IMDB-WIKI dataset. Some images are shown in Fig. 1. We make our IMDB-WIKI dataset publicly available. It is the largest public dataset for biological age prediction.

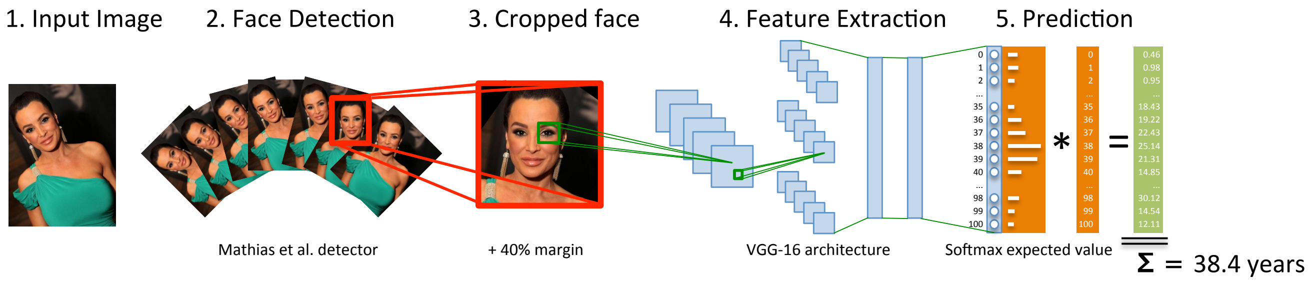


Figure 2. Pipeline of DEX method (with one CNN) for apparent age estimation.

Age estimation is a regression problem, as the age is from a continuous range of values. We go further than using the regression training of CNNs and train CNNs for classification where the age values are rounded into 101 year labels [0,1,..,100]. By posing the age regression as a deep classification problem followed by a softmax expected value refinement we improve significantly over direct regression training of CNNs.

Our proposed method, Deep EXpectation (DEX) of apparent age (see Fig. 2), first detects the face in the test image and then extracts the CNN predictions from an ensemble of 20 networks on the cropped face. DEX is pretrained on ImageNet, finetuned on our IMDB-WIKI dataset, and then on LAP images. Our DEX is the winner (1^{st} place) of the ChaLearn LAP 2015 challenge on apparent age prediction, outperforming the human reference.

Our main contributions are as follows:

1. the IMDB-WIKI dataset, the largest dataset for biological age prediction;

2. a novel regression formulation through a deep classification followed by expected value refinement;

3. DEX system, winner of the LAP 2015 challenge on apparent age estimation.

The remainder of the paper is as follows. In Section 2 we introduce our method (DEX) by describing the face detection and apparent age estimation components. In Section 3 we describe the datasets (including our new proposed IMDB-WIKI dataset for age estimation) and the experiments and discuss our method and its performance on the ChaLearn LAP 2015 challenge. Section 4 concludes the paper.

2. Proposed method (DEX)

Our proposed Deep Expectation (DEX) method follows the pipeline from Fig. 2. Next we provide details about each step and the final ensemble of CNNs.

2.1. Face Detection

For both training and testing images, we run the off-the-shelf face detector of Mathias *et al.* [10] to obtain the location of the face.

In order to align the faces we run the face detector not only on the original image but also on all rotated versions between $-60°$ and $60°$ in $5°$ steps. As a few of the training images were upside down or rotated by $90°$, we also run the detector at $-90°$, $90°$, and $180°$. Due to the limited computational resources we used only this discrete set of rotated images. We take the face with the strongest detection score and rotate it accordingly to a up-frontal position.

For very few images ($< 0.2\%$) the face detector is not able to find a face. In those cases we just take the entire image. On the final LAP test set this applies only to 1 image.

We then extend the face size and take 40% of its width to the left and right and 40% of its height above and below. Adding this context helps the prediction accuracy. If the face already covers most of the image, we just pad with the last pixel at the border. This ensures that the face is always at the same location of the image.

The resulting image is then squeezed to 256×256 pixels and used as an input to a deep convolutional network.

2.2. Face Apparent Age Estimation

The apparent age prediction is obtained by applying a deep convolutional neural network to the detected face from the previous processing stage. Our method uses the VGG-16 architecture [13] which has shown impressive results on the ImageNet challenge [12].

2.2.1 Deep learning with CNNs

All our CNNs start from the VGG-16 architecture pretrained on the ImageNet dataset for image classification [13]. The CNNs are then finetuned on our IMDB-WIKI dataset. When training for regression the output layer is changed to have a single neuron for the regressed age. When training for classification, the output layer is adapted to 101 output neurons corresponding to natural numbers

Table 1. IMDB-WIKI dataset and its partitions sizes in number of images.

IMDB-WIKI	IMDB	Wikipedia	IMDB-WIKI used for CNN training
524,230	461,871	62,359	260,282 images

from 0 to 100, the year discretization used for age class labels.

2.2.2 Expected Value

Age estimation can be seen as a piece-wise regression or, alternatively, as a discrete classification with multiple discrete value labels. The larger the number of classes is, the smaller the discretization error gets for the regressed signal. In our case, it is a one dimensional regression problem with the age being sampled from a continuous signal ([0,100]).

We can improve the classification formulation for regressing the age by heavily increasing the number of classes and thus better approximating the signal and by combining the neuron outputs to recover the signal. Increasing the number of classes demands sufficient training samples per each class and increases the chance of overfitting the training age distribution and of having classes not trained properly due to a lack of samples or unbalance. After a number of preliminary experiments, we decided to work with 101 age classes. For improving the accuracy of the prediction, as shown in Fig. 2, we compute a softmax expected value, E, as follows:

$$E(O) = \sum_{i=0}^{100} y_i o_i \tag{1}$$

where $O = \{0, 1, \cdots, 100\}$ is the 101 dimensional output layer, representing softmax output probabilities $o_i \in O$, and y_i are the discrete years corresponding to each class i.

2.2.3 Ensemble of CNNs

After the finetuning on the IMDB-WIKI dataset, we further finetune the resulting network on 20 different splits of the ChaLearn LAP dataset [4]. In each split we use 90% of the images for training and 10% for validation. The splits are chosen randomly for each age separately, i.e. the age distribution in the training is always the same. We then train the 20 networks on an augmented version of the ChaLearn dataset, as we add 10 augmented versions of each image. Each augmentation randomly rotates the image by $-10°$ to $10°$, translates it by -10% to 10% of the size and scales it by 0.9 to 1.1 of the original size. We do the augmentation after splitting the data into the training and validation set to ensure that there is no overlap between the two sets. Each network is then trained and we pick the weights with the best performance on the validation set.

The final prediction is the average of the ensemble of 20 networks trained on slightly different splits of the data.

3. Experiments

In this section we first introduce the datasets and the evaluation protocols from our experiments. Then we provide implementation details for our DEX method, describe experimental setups and discuss results.

3.1. Datasets and evaluation protocol

3.1.1 IMDB-WIKI dataset for age prediction

For good performance, usually the large CNN architectures need large training datasets. Since the publicly available face image datasets are often of small to medium size, rarely exceeding tens of thousands of images, and often without age information we decided to collect a large dataset of celebrities. For this purpose, we took the list of the most popular 100,000 actors as listed on the IMDB website [1] and (automatically) crawled from their profiles birth dates, images, and annotations. We removed the images without timestamp (the date when the photo was taken), also the images with multiple high scored face detections (see Section 2.1). By assuming that the images with single faces are likely to show the actor and that the time stamp and birth date are correct, we were able to assign to each such image the biological (real) age. Of course, we can not vouch for the accuracy of the assigned age information. Besides wrong time stamps, many images are stills from movies, movies that can have extended production times. In total we obtained 461,871 face images for celebrities from IMDB.

From Wikipedia [2] we crawled all profile images from pages of people and after filtering them according to the same criteria applied for the IMDB images, we ended up with 62,359 images. In Table 1 we summarize the IMDB-WIKI dataset that we make public. In total there are 524,230 face images with crawled age information. As some of the images (especially from IMDB) contain several people we only use the photos where the second strongest face detection is below a threshold. For the network to be equally discriminative for all ages, we equalize the age distribution, i.e. we randomly ignore some of the images of the most frequent ages. This leaves us with 260,282 training images for our CNNs.

3.1.2 LAP dataset for apparent age estimation

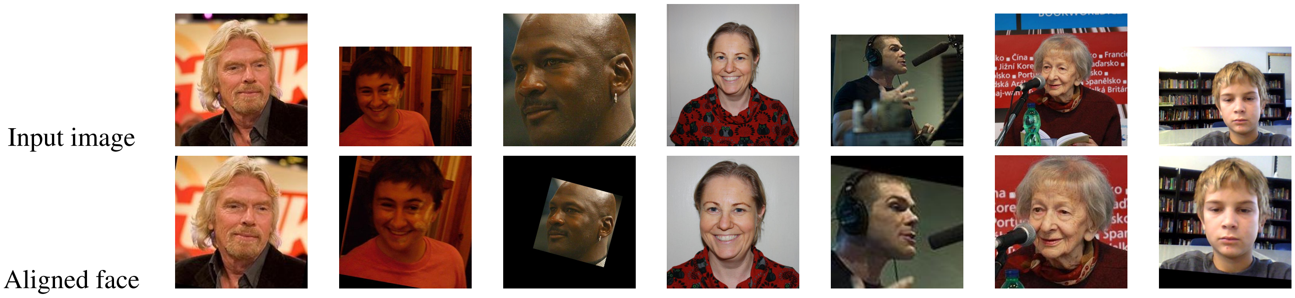
The ChaLearn LAP dataset [4] consists of 4699 face images collectively age labeled using two web-based applications.

[1] www.imdb.com

[2] en.wikipedia.org

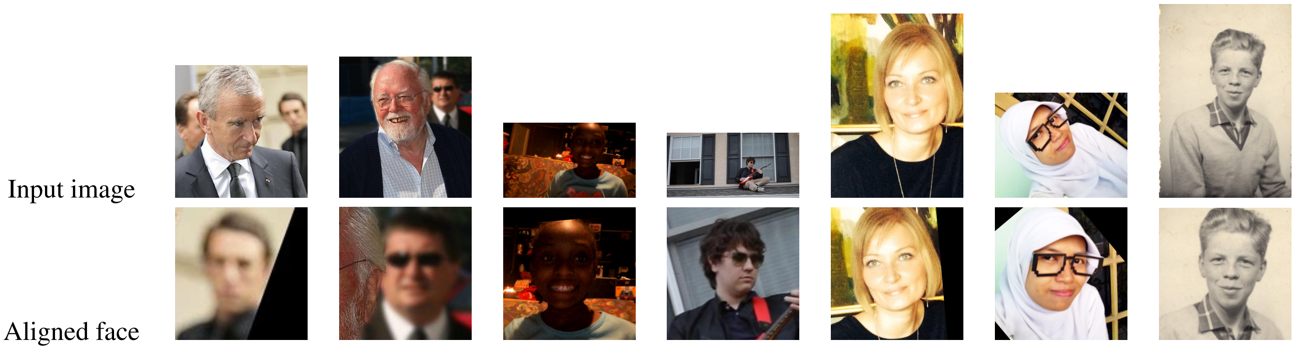
Table 2. Performance on validation set of ChaLearn LAP 2015 apparent age estimation challenge.

Network		Learning	MAE	ϵ-error
pretrain	finetune			
ImageNet [12]	LAP [4]	Regression	5.007	0.431
		Classification	7.216	0.549
		Classification + Expected Value	6.082	0.508
ImageNet [12] & **IMDB-WIKI (ours)**	LAP [4]	Regression	3.531	0.301
		Classification	3.349	0.291
		Classification + Expected Value	**3.221**	**0.278**



Apparent age	57	17	40	50	30	79	12
Predicted age	57.75	16.15	39.43	49.15	32.06	78.99	12.78

Figure 3. Examples of face images with good age estimation by DEX with a single CNN.



Apparent age	57	62	11	20	40	23	15
Predicted age	27.50	43.23	26.35	34.07	26.63	35.81	27.25

Figure 4. Examples of face images were DEX fails the age estimation. DEX uses a single CNN.

Each label is the averaged opinion of at least 10 independent users. Therefore, a standard deviation σ is also provided for each age label. The LAP dataset is split into 2476 images for training, 1136 images for validation and 1087 images for testing. The age distribution is the same in all the three sets of the LAP dataset. LAP covers the 20-40 years interval best, while for the [0,15] and [65,100] intervals it suffers from small number of samples per year.

3.1.3 Evaluation protocol

In our paper the results are evaluated either by using the standard MAE measure or the ϵ-error as defined for the ChaLearn LAP challenge.

MAE. The standard mean absolute error (MAE) is computed as the average of absolute errors between the estimated age and the ground truth age. Note that the error does not capture the uncertainty in the ground truth labeled age. The ϵ-error covers such aspect.

ϵ-error. LAP dataset images are annotated with the average and the standard deviation σ of the age votes casted by multiple users. The LAP challenge evaluation employs fitting a normal distribution with the mean μ and standard deviation σ of the votes for each image:

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

For a set of images the ϵ-error is the average of the above introduced errors at image level. ϵ can be maximum 1 (the worst) and minimum 0 (the best).

3.2. Implementation details

The pipeline is written in Matlab. The CNNs are trained on Nvidia Tesla K40C GPUs using the Caffe framework [7]. The face detection was run in parallel over a Sun Grid Engine which was essential for the IMDB and Wikipedia images.

Training the network on the IMDB and Wikipedia images took around 5 days. Finetuning a single network on the ChaLearn dataset takes about 3h. Testing the face detection at each rotation takes around 1s. The feature extraction per image and network takes 200ms.

The source codes and IMDB-WIKI dataset are publicly available at:
http://www.vision.ee.ethz.ch/~timofter

3.3. Validation results

During experiments we noticed that the softmax expected value on the network trained for classification works better than a) training a regression, b) learning a regression (i.e. SVR) on top of the CNN features of the previous layer, or c) just taking the age of the neuron with the highest probability. In Table 2 we report the MAE and ϵ-error for different setups and a single CNN. We notice the large improvement (2 up to 4 years reduction in MAE) brought by the additional training on the IMDB-WIKI face images. This matches our expectation since the network learns a powerful representation for age estimation which is relevant to the apparent age estimation target on the LAP dataset. Training the network directly for regression leads to 0.301 ϵ-error (3.531 MAE) on the validation set of the LAP dataset. By changing to the classification formulation with 101 output neurons $\{0, 1, \cdots, 100\}$ corresponding to the rounded years we improve to 0.291ϵ-error (3.349 MAE). With our softmax expected value refinement we get the best results on the LAP validation set, 0.278ϵ-error and 3.221 MAE.

Qualitative results (such as in Fig. 3) showed that our proposed solution is able to predict the apparent age of faces in the wild as well as people can. This is partly enabled by learning from our large IMDB-WIKI dataset depicting faces in the wild.

In Fig. 4 we show a number of face images where our DEX method with a single CNN fails. The main causes are: 1) the failure of the detection stage – either no face is detected or the wrong face (a background face) is selected; 2) extreme conditions and/or corruptions, such as dark images, glasses, old photographs.

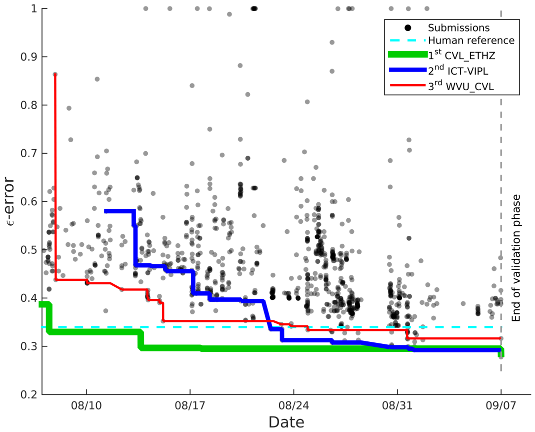


Figure 5. One month validation entries for LAP challenge. For the top 3 teams we plot the best scores curves. **CVL_ETHZ is ours.**

3.4. Looking at people (LAP) challenge

The ChaLearn Looking at people (LAP) challenge [4] on apparent age estimation consisted of two phases: development (validation) and test.

3.4.1 Development phase

For the development phase the training and validation images of the LAP datasets were released. Whereas the training images had the apparent age labels, the validation labels were kept unknown until the beginning of the second phase. The teams submitted their results on the validation images to the server for getting their performance scores. The evolution of the scoreboard for the validation images is depicted in Fig. 5. In order to plot the above mentioned scoreboard we crawled the scores from the competition website. We can easily notice that the quality of the results improve over time on average.

3.4.2 Test phase

For the test phase the validation labels were released and the access to the test images was granted but without test labels. The teams were invited to submit their results on the test images to the competition server. The scores remained unknown until the organizers announced the final ranking after the test phase. Our results were obtained using DEX with the full ensemble of 20 CNNs, classification prediction and expected value refinement.

3.4.3 Final ranking

The final ranking of the ChaLearn LAP challenge [4] on apparent age estimation matches the scoreboard evolution during the online validation phase (see Table 3). The best 4

Table 3. ChaLearn LAP 2015 final ranking on the test set. 115 registered participants. AgeSeer did not provide codes. The human reference result is the one reported by the organizers.

Rank	Team	ϵ error
1	CVL_ETHZ (ours)	0.264975
2	ICT-VIPL	0.270685
~~3~~	~~AgeSeer~~	~~0.287266~~
3	WVU_CVL	0.294835
4	SEU-NJU	0.305763
	human reference	*0.34*
5	UMD	0.373352
6	Enjuto	0.374390
7	Sungbin Choi	0.420554
8	Lab219A	0.499181
9	Bogazici	0.524055
10	Notts CVLab	0.594248

methods drop below 0.34 ϵ-error, the human reference performance as reported by the organizers during the development phase.

Noteworthy is that we are the only team from the top 6 that did not use facial landmarks. This said, we believe that the performance of our DEX method can be improved by using landmarks.

4. Conclusions

We tackled the estimation of apparent age in still face images. Our proposed Deep EXpectation (DEX) method uses convolutional neural networks (CNNs) with VGG-16 architecture pretrained on ImageNet. In addition, we crawled Internet face images with available age to create the largest such public dataset known to date and to pretrain our CNNs. Further, our CNNs are finetuned on apparent age labeled face images. We posed the age regression problem as a deep classification problem followed by a softmax expected value refinement and show improvements over direct regression training of CNNs. DEX ensembles the prediction of 20 networks on the cropped face image. DEX does not explicitly employ facial landmarks. Our proposed method won (1^{st} place) the ChaLearn LAP 2015 challenge [4] on apparent age estimation, significantly outperforming the human reference.

Acknowledgments. This work was supported by the KTI-SUPSI (#2-69650-14) project.

References

[1] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014. 1

[2] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *CVPR*, 2012. 1

[3] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 1

[4] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, and I. Guyon. Chalearn 2015 apparent age and cultural event recognition: datasets and results. In *ICCV, ChaLearn Looking at People workshop*, 2015. 1, 3, 4, 5, 6

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[6] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *ICB*, 2013. 1

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014. 5

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[9] A. Lanitis. The fg-net aging database. *www-prima.inrialpes.fr/FGnet/html/benchmarks.html*, 2002. 1

[10] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*. 2014. 2

[11] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, 2006. 1

[12] 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. *IJCV*, pages 1–42, 2014. 1, 2, 4

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 2