Accurate brain age prediction with lightweight deep neural networks

Han Peng[1,2,3*†], Weikang Gong[1*], Christian F. Beckmann[1,3], Andrea Vedaldi[2], Stephen M. Smith[1]

[1]Centre for Functional MRI of the Brain (FMRIB), University of Oxford, Oxford, OX3 9DU, United Kingdom
[2]Visual Geometry Group (VGG), University of Oxford, Oxford, OX2 6NN, United Kingdom
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, 6525 EN, The Netherlands

* These authors contributed equally
† Corresponding to Han Peng, han.peng@ndcn.ox.ac.uk

Highlights

- A lightweight deep learning model, Simple Fully Convolutional Network (SFCN), is presented, achieving state-of-the-art brain age prediction performance in UK Biobank MRI brain imaging data.
- Even with limited number of training subjects (e.g., 50), SFCN performs better than widely-used regression models.
- A semi-multimodal ensemble strategy is proposed and achieved first place in the PAC 2019 brain age prediction challenge.
- Linear regression can remove brain age predication bias (even on unlabelled data) while maintaining state-of-the-art performance.

Abstract

Deep learning has huge potential for accurate disease prediction with neuroimaging data, but the prediction performance is often limited by training-dataset size and compute memory requirements. To address this, we propose a deep convolutional neural network model, Simple Fully Convolutional Network (SFCN), for accurate prediction of brain age using T1-weighted structural MRI data. Compared with other popular deep network architectures, SFCN has fewer parameters, so is more compatible with small dataset size and 3D volume data. The network architecture was combined with several techniques for boosting performance, including data augmentation, pre-training, model regularization, model ensemble and prediction bias correction. We compared our overall SFCN approach with several widely-used machine learning models. It achieved state-of-the-art performance in UK Biobank data (N = 14,503), with mean absolute error (MAE) = 2.14y in brain age prediction and 99.3% in sex classification. SFCN also won (both parts of) the 2019 Predictive Analysis Challenge for brain age prediction, involving 79 competing teams (N = 2,638, MAE = 2.90y). We describe here the details of our approach, and its optimisation and validation. Our approach can easily be generalised to other tasks using different image modalities, and is released on GitHub.

1 Introduction

The emergence of machine learning techniques has made automatic disease prediction from medical imaging data possible. The recent development of deep learning pushes prediction accuracy beyond human performance in some scenarios, and is able to assist clinical diagnosis/treatment decisions (De Fauw et al., 2018; Kohl et al., 2018; LeCun et al., 2015). In neuroimaging, deep learning has had some successes, and yet faces several challenges (Arslan et al., 2018; Baumgartner et al., 2018; Cole et al., 2017; Kawahara et al., 2017). For example, 3D neuroimaging data requires much more GPU memory than most 2D images, meaning that models successful in 2D data (e.g., ImageNet classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014)) are infeasible in the 3D scenario. Further, deep networks usually require a large sample size for model fitting, but neuroimaging datasets often have relatively few samples compared to existing million-sample natural image datasets (Russakovsky et al., 2015), which could limit the ability to learn image features effectively, and result in overfitting problems. To design models to address these challenges for neuroimaging applications, we need to evaluate them with a publicly available and broadly accepted benchmarking platform.

Predicting chronological age based on structural brain magnetic resonance imaging (MRI) data not only provides a way for benchmarking, comparing and improving deep learning algorithms, but also receives attention for its potential clinical and biological relevance (Cole et al., 2018; Cole and Franke, 2017; Franke and Gaser, 2019; Kaufmann et al., 2019). The predicted age can be considered the "brain age", because it is derived purely from the brain imaging data. After estimating brain age, a further quantity of interest is the difference between the predicted age (brain age) and the actual age, sometimes referred to as the brain-age delta. Positive delta implies that a subject's brain looks older than their actual age, i.e., they are experiencing accelerated aging. For example, existing studies have observed that the brain-age delta is an effective biomarker that shows differences between different clinical groups (Kaufmann et al., 2019) and is predictive for mortality (Cole et al., 2018). Achieving accurate brain age prediction is an essential pre-requisite for optimising brain-age delta as a biomarker. To reach this goal, many studies have used different models, such as regularized linear regression, support vector machines and Gaussian process regression, for brain age prediction (Franke and Gaser, 2019). Some studies have used deep learning methods (Cole et al., 2017; Feng et al., 2019; Kolbeinsson et al., 2019). However, challenges exist for further improvement of prediction accuracy, especially on small datasets, and some studies have shown that deep learning performs no better than simpler machine learning models in typical neuroimaging datasets (He et al., 2019; Schulz et al., 2019). It has not yet, for example, been clearly established whether more complex deep learning models perform better than simpler models (for the task of brain age prediction using structural MRI data). In addition, predicted brain age is often systematically biased towards the group mean value, resulting in a correlation between the delta and the chronological age, which weakens the validity of the delta as a biomarker (Smith et al., 2019b). Therefore, it is both methodologically interesting and scientifically important to develop unbiased high-performance deep learning strategies for brain age prediction.

In this paper, a lightweight deep learning architecture, Simple Fully Convolutional Network (SFCN), is presented for brain age prediction. Its architecture is based on the fully convolutional network (FCN) (Zhang et al., 2018) and the VGG net (Simonyan and Zisserman, 2014) and takes 3D minimally-preprocessed T1 brain images (and/or preprocessed

segmentation outputs) as input. Using proper data augmentation and regularization techniques, the model achieved state-of-the-art mean absolute error (MAE) of 2.14 years in the UK Biobank dataset (14,503 subjects, of which 12,949 are used for training). This model performs better than several widely-used machine learning models in the literature. In addition, we propose a model ensemble strategy that averages the outputs of deep learning models based on different kinds of preprocessing applied to the T1 data, namely, white matter segmentation, grey matter segmentation, linearly-registered raw T1 and nonlinearly registered raw T1; this further boosts the accuracy of brain age prediction. Finally, we extended the bias correction techniques proposed by (Smith et al., 2019b) to greatly reduce the correlation between the brain-age delta and the chronological age, with very little compromise of performance, even when the true ages of test (validation) subjects are unknown. The ensemble SFCN came first in the Predictive Analysis Challenge 2019 in brain age prediction (MAE = 2.90 years) among the 79 participating teams[1]. With bias correction, our model achieved an MAE of 2.95 years, thereby ranking first also in the other part of the competition (most accurate age prediction while minimising bias). The trained model is available in the GitHub repository:
https://github.com/ha-ha-ha-han/UKBiobank_deep_pretrain/

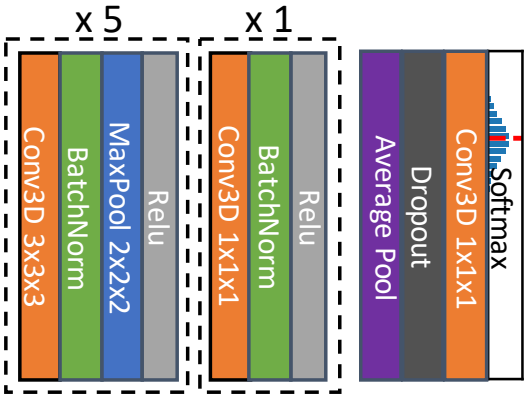

Figure 1. Illustration of the core network for the Simple Fully Convolutional Neural Network (SFCN) model. The model takes 3D brain image data and contains 7 blocks. Each of the first 5 consecutive blocks consists of a 3x3x3 3D convolution layer, a Batch Norm layer, a Max Pooling layer and a ReLU activation. The 6th block contains one 1x1x1 3D convolution layer, a Batch Norm layer and a ReLU activation. The 7th block contains an average pooling layer, a dropout layer, an 1x1x1 3D convolution layer and a softmax layer.

[1] PAC 2019 website: https://www.photon-ai.com/pac2019

2 Methods

2.1 Model: Simple Fully Convolutional Network (SFCN)

We use a convolutional neural network (CNN) architecture to estimate brain age using 3D T1 images. The architecture is based on VGGNet (Simonyan and Zisserman, 2014) and uses a fully convolutional structure (Zhang et al., 2018), but we keep the number of layers as small as possible to reduce the number of parameters to about 3 million, and therefore to reduce computational complexity and memory cost. We name this model structure "Simple Fully Convolutional Neural Network" (SFCN) to reflect its simplicity.

The model consists of seven blocks, as shown in Figure 1. Each of the first five blocks contains a 3-by-3-by-3 3D convolutional layer, a batch normalisation layer (Ioffe and Szegedy, 2015), a max pooling layer and a ReLU activation layer (LeCun et al., 2015). The 1mm-input-resolution 160x192x160 3D input image (with little or no brain tissue loss) goes through each block sequentially, with its feature map generated and spatial dimension reduced to 5x6x5 after the fifth block. The sixth block contains a 1x1x1 3D convolutional layer, a batch normalisation layer and a ReLU activation layer. The seventh block contains an average pooling layer, a dropout layer (only used for training, with 50% random dropout rate) (Srivastava et al., 2014), a fully connected layer and a softmax output layer. The channel numbers used in each convolution layer are [32, 64, 128, 256, 256, 64, 40]. The output layer contains 40 digits that represent the predicted probability that the subject's age falls into a one-year age interval between 42 to 82 (for UK Biobank) or a two-year age interval between 14 to 94 (for PAC). A weighted average of each age bin is calculated to make the final prediction:

$$pred = \sum_{c}^{40} x_c \cdot age_c$$

x_c stands for the probability predicted for the cth age bin and age_c stands for the bin centre for the age interval.

The internal process of the model can be interpreted as three stages: 1) The first five blocks extract feature maps from each input image. 2) The sixth block further increases the nonlinearity of the model by adding one extra nonlinear layer but without changing the output size of feature maps. 3) The seventh block maps the generated features to the predicted age probability distribution. The first two stages encode the input image to a feature vector, and the third stage can be viewed as a classifier based on the deep feature. At the first stage, the spatial information is maintained and takes most of the memory. To reduce the overall GPU memory consumption, we limited the channel numbers of the first layer to 32 and put only one convolutional layer in every block. To compare, a VGGnet usually has two convolutional layers inside a block and has 64 channels in the first layer (Simonyan and Zisserman, 2014). At the later stages (higher-level layers) of deep learning models, fully connected (FC) layers usually have the largest number of learnable parameters. For example, the penultimate layer of VGGnet (FC-4096) consists of about 16 million parameters. By removing most of the FC layers and keeping the number of channels small in the last two stages in SFCN, the number of learnable parameters is greatly reduced. Although reducing the number of FC layers can potentially reduce the nonlinearities learnt by the model, in most neuroimaging classification tasks, the number of classes is smaller than that for natural images. For example, there are only 40 "classes" (age bins) for brain age prediction, which is a very small number compared to 1000 classes in the ImageNet classification task. In this case,

the small parameter number and the lack of FC layers do not harm the testing (validation) performance.

To compare SFCN with a popular CNN architecture, we implemented a 3D version of ResNet (He et al., 2016). The architecture of 3D-ResNet follows the literature but the convolution filters are changed to 3D. For the experiments, the SFCN and the ResNet share the same training parameters and both achieve successful performance in the training set. (Comparison against a broader set of alternative approaches is of course provided via the results from the PAC competition.)

SFCN contains only 3.0 million parameters, which is less than one tenth of the 33.2 million for 3D ResNet-18, 46.2 million for 3D ResNet50 and 133 million for 2D VGGnet (Simonyan and Zisserman, 2014).

2.2 Regression models: Elastic Net

We compared our deep learning model with simpler machine learning models using T1 MRI derived features as inputs (Schulz et al., 2019). We choose Elastic Net for our comparison, because it has been shown to be a high-performance and stable machine learning model for neuroimaging data (Jollans et al., 2019). Three forms of the T1 data from UKB were (separately) used for age prediction: (1) Voxel-level linearly registered "raw" T1 images; (2) Voxel-level grey matter partial-volume estimated by FSLVBM voxel-based morphometry; (3) T1-image derived region-level phenotypes (Miller et al., 2016). In the training set, we used principal component analysis (PCA) to reduce the data into an L-dimensional space (L=5000 for (1) and (2), no PCA for (3)), and then used Pearson correlation to select the top k features (from k=10 to all features) that correlate with age, and finally used elastic net regression (implemented in the glmnet package) to predict age (Friedman et al., 2010). All model parameters were optimised via internal cross-validation within the validation set. The selected best model was applied to the test set and performance reported. We also implemented a widely-used support vector machine for brain age prediction, but did not find better performance compared with the above model.

2.3 Bias correction

We used the linear bias correction method described in (Smith et al., 2019b) for bias correction for the delta. Such a bias correction is valuable for most brain-age prediction studies, as there is normally an underfitting of the prediction, due to problems such as regression dilution and non-Gaussian age distribution. Defining y to be chronological age and x the predicted age, we fitted a linear regression $x = ay + b$ to the left-out validation set (with labels). The corrected predicted age is estimated by

$$\hat{x} = (x - b)/a$$

This method requires (at the point of estimating a and b from x and y) that the chronological ages are known. For the label-missing (final evaluation) test set, we assumed that a and b are generalisable, and used the coefficients previously fitted in the left-out validation set to estimate the corrected brain-age delta.

3 Experiments

3.1 Datasets and preprocessing

3.1.1 UK Biobank

UK Biobank is collecting a large-cohort of brain imaging data from predominantly healthy participants (Miller et al., 2016). In this study, we used the T1 data from 14,503 subjects (mean age 52.7 years, standard deviation 7.5 years, range 44-80 years), of which 12,949 were used for training, 518 for validation and 1,036 for testing. The image preprocessing pipeline is described in (Alfaro-Almagro et al., 2018). The input data to the deep neural network model was brain extracted, bias corrected and linearly registered to MNI152 standard space (unless otherwise specified).

3.1.2 PAC 2019

As part of testing the performance of our method objectively, we participated in the Predictive Analytic Challenge (PAC) 2019. This competition was broken down into two parts: a) to achieve the lowest mean absolute error ($MAE = \frac{1}{N}\sum_{i=1}^{N} |pred_i - age_i|$) for brain age prediction; and b) to achieve the lowest MAE while keep the Spearman correlation between the brain-age delta and the chronological age under 0.1 (because in general, ideally delta would have no bias or age dependence). The dataset contains T1 structural MRI brain images from 2,638 subjects (mean age 35.9 years, standard deviation 16.2 years, range 17-90 years).[2] We used 2,199 subjects for training, and 439 subjects as a left-out validation set. In addition, there were 660 subjects whose labels were unknown to the challenge participants, forming a test set for benchmarking (i.e., the results on this test set determined the final challenge scores).

3.2 Training and testing

During the training process, we use an Stochastic Gradient Descent (SGD) optimiser (Sutskever et al., 2013) for the UKB dataset to minimise a Kullback–Leibler divergence loss function between the predicted probability and a Gaussian distribution (the mean is the true age, and the distribution sigma is 1 year for UKB) for each training subject. This soft-classification loss encourages the model to predict age as accurately as possible. To reduce over-fitting, two data augmentation methods are applied during the training phase. In every epoch, the training input is 1) randomly shifted by 0, 1 or 2 voxels along every axis; 2) has a probability of 50% to be mirrored about the sagittal plane.

The performance of the model can be evaluated by Mean Absolute Error (MAE) and Pearson correlation coefficient (r-value) in the validation and test sets.

All the models were trained with two NVIDIA P100 GPUs. The training time was approximately 0.5 hour to go through each of the 12,949 training subjects once (i.e., one training epoch). The L2 weight decay coefficient was 0.001. The batch size was 8. The learning rate for the SGD optimiser was initialized as 0.01, then multiplied by 0.3 every 30 epochs unless otherwise specified. The total epoch number is 120 for 12,949 training subjects. The epoch number is adjusted accordingly for the experiments with smaller training sets so that the training steps are roughly the same. The epoch with the best validation MAE is used for testing.

[2] This information is publicly available in the challenge website: https://www.photon-ai.com/pac2019

Model	Test MAE (years)	Train MAE (yrs)	Test-Train MAE gap (yrs)
SFCN without regularization	2.60±0.06	0.337±0.012	2.26
SFCN	**2.14±0.05**	**1.36±0.03**	**0.78**
3D ResNet18	2.50±0.06	0.40±0.04	2.10
3D ResNet50	2.32±0.05	0.88±0.05	1.44

Table 1. Performance of different deep learning models in UK Biobank data. The training set size is 12,949. The input data are T1 MRI images which are linearly registered to a standard space. After the training is done, the epoch with the best validation MAE is selected to be evaluated on the test set. The test results are bootstrapped 1000 times to compute the mean test MAE and the standard deviation. Epochs 95 to 110 are selected to compute the train MAE, standard deviation of the train MAE and the mean test-train MAE gap (the difference between the test MAE and the train MAE).

Data Augmentation and Regularisation	Test MAE (yrs)	Train MAE (yrs)	Test-Train MAE gap (yrs)
None	3.67±0.09	0.305±0.006	3.29
DP	3.59±0.08	0.92±0.03	2.75
VS	3.05±0.07	0.59±0.01	2.55
MR	3.13±0.08	0.47±0.01	2.78
DP + VS + MR	**2.82±0.07**	**1.51±0.03**	**1.21**
DP + VS + MR + 1FC	3.59±0.08	1.50±0.06	2.04

Table 2. Performance of SFCN with different regularisation and data augmentation methods. DP=Dropout, VS=Voxel Shifting, MR=Mirroring, 1FC = SFCN with one extra fully connected layer. The training set size is 2,072. The input data are T1 MRI images which are linearly registered to a standard space. After the training is done, the epoch with the best validation MAE is selected to be evaluated on the test set. The test results are bootstrapped 1000 times to compute the mean test MAE and the standard deviation. Epochs 185 to 200 are selected to compute the train MAE, standard deviation of the train MAE and the mean validation-train MAE gap (the difference between the test MAE and the train MAE).

For the ensemble strategy, we randomly initialised and trained 20 models; 5 (identical network structure but randomly-initialised parameters) models were trained on each of the four input data types: linearly registered GM and WM, non-linearly registered T1 and linearly registered T1. The ensemble experiments use 2,590 subjects for training to reduce the overall computation time. The prediction is made by averaging the results of all 20 models.

3.3 Sex classification

To show the generalisability of SFCN to other tasks, we also tested the performance for sex classification. The architecture of the model and the training setting remains mostly the same as for age prediction, with differences now described. Training was performed with two NVIDIA GTX1080ti GPUs and the channel numbers are reduced to [28, 58, 128, 256, 256, 64] to accommodate to the GPU memory. The number of classes is two and the loss function is the cross entropy. The training set size is 5,180 and the test set is the same as for age prediction. The input brain is linearly registered to standard space so that the overall brain size is the same for all subjects of both sexes.

4 Results

4.1 The performance of SFCN in UK Biobank data

Table 1 shows the performance of the SFCN in the UKB dataset with 12,949 training subjects. SFCN with data augmentation and dropout achieved an MAE of 2.14 years, which is 0.46 years better than that without these regularizations.

With the same regularization techniques, we then compared SFCN with other popular CNN architectures, namely the 3D version of ResNet-18 and ResNet-50 (He et al., 2016). MAE is 2.50 years for ResNet-18, and 2.32 years for ResNet-50. SFCN performs 0.36 years and 0.18 years better, respectively.

We attribute this performance boost of SFCN to the lightweight model structure. As shown in Table 1, the SFCN with regularizations achieved the best test performance but by far the worst in the training set, suggesting a significant difference (between models) in levels of over-fitting. The gap (a measure of over-fitting) between the test MAE and the train MAE is 0.78 years for the SFCN, which is the smallest among all the models (ResNet18: 2.10 years, ResNet50: 1.44 years).

The SFCN model trained with a dropout layer and data augmentation achieves the best MAE. To study the effect of the regularisation techniques, we trained models with one of the three techniques, namely, dropout, voxel shifting and mirroring, and show the test results in Table 2, with 2,072 training subjects (hence the worse overall results compared with the above). When applied to each of these 3 techniques individually during training, each of the regularisation methods reduces over-fitting and improves the test MAE by about 0.1-0.5 years. Combining all the three methods together, the model achieves the best test performance given this number of training subjects (MAE = 2.82 years), showing a large improvement of 0.85 years compared with the unregularized model. Finally, we added one fully connected layer with 64 channels (together with batch normalisation) before the final layer. While giving similar training MAE, the added layer reduces the generalisability to the test set (test MAE=3.59y). These results clearly show that the lightweight model structure and the regularisation techniques improve the model performance and can be used for future reference to design deep learning strategy in neuroimage datasets.

Our presented strategy achieves state-of-the-art results in brain age prediction. Table 3 shows a summary of previously reported brain age prediction MAE results (Kolbeinsson et al., 2019;

Training set size	Model	Performance
		MAE (yrs)
2590	**SFCN**	**2.76±0.06**
2679	Linear regression Ning et al. 2019 bioRxiv	3.5
5180	**SFCN**	**2.28±0.05**
5700	3D-ResNet + Tensor Regression Kolbeinsson et al. 2019 arxiv	2.58
12949	**SFCN**	**2.14±0.05**
17100	SVD + Linear regression Smith et al. 2019 NeuroImage	3.6

Table 3. A summary of the reported UK Biobank study in brain age prediction.

Ning et al., 2018; Smith et al., 2019b). To eliminate the effect of sample size differences (i.e., to make these comparisons as meaningful as possible), we trained SFCN with comparable training set sizes as the previous studies, and compared performance with those. With about 2600 training subjects, SFCN achieves an MAE of 2.76 years, while linear regression achieves MAE 3.5 years (Ning et al., 2018). With about 5000 training subjects, SFCN achieves 2.28 years MAE while 3D-ResNet with tensor regression achieves 2.58 years (Kolbeinsson et al., 2019). For the larger training set with more than 10,000 subjects, linear regression with multi-modality IDPs (including fMRI and DTI features) achieves an MAE of 2.9 years (Smith et al., 2019b, 2019a), whereas SFCN obtains the best MAE in UK Biobank with an MAE of 2.14 years.

Besides the state-of-the-art brain age MAE performance among all the reported studies in UK Biobank, our model and strategy also achieved 99.3% accuracy for sex classification (0.7% error rate) based on T1 images, which is a considerate improvement compared to the previously reported results (classification accuracies varying from 69% to 93%, with or without head size regressed out) (Chekroud et al., 2016; Giudice et al., 2016; Joel et al., 2016, 2015; Rosenblatta, 2016). This result suggests that SFCN is generalisable to other tasks for neuroimaging research.

4.2 Comparing the learning curves of SFCN with simpler regression models

There are controversies regarding whether a deep learning model can perform better than linear models to predict phenotypic and behavioural variables using neuroimaging data (He et al., 2018, 2019; Schulz et al., 2019). For the task brain age prediction using T1 structural MRI data, two questions remain to be answered: 1. Do DL models surpass the performance of simpler regression models? 2. How many training samples do DL or simpler regression models need for good performance?

We compared our deep learning model and a well-tuned regression model, elastic net, for brain age prediction. We also explored the effect of the training dataset size (from 50 to

12,949 subjects) on the performance of the two models. As summarised in Figure 2, we find that the SFCN outperforms elastic net regardless of the training set size. Even with as few as 50 training subjects, the DL model achieves better performance.

MAE decreases with the increase of training set size, and approximately follows a linear-log relationship for all methods. If the training set size doubles, the MAE decreases by about 0.3 to 0.4 years. In our experiment setup, there is no conclusive signature of performance saturation for the large dataset size, although the last few data points do deviate from the simple linear-log relationship. With the increasing size of UK Biobank and other datasets, we can expect even better performance in future studies.

4.3 Semi-multimodal model ensemble improves the performance with limited number of training subjects

In previous sections, we trained our SFCN model using only one modality, namely raw T1 data linearly registered to the MNI space (Lin). To test whether adding other modalities (here "modalities" refers to different kinds of preprocessing of the T1 data) can further boost performance, we trained SFCN using three other modalities derived from T1 image data: raw

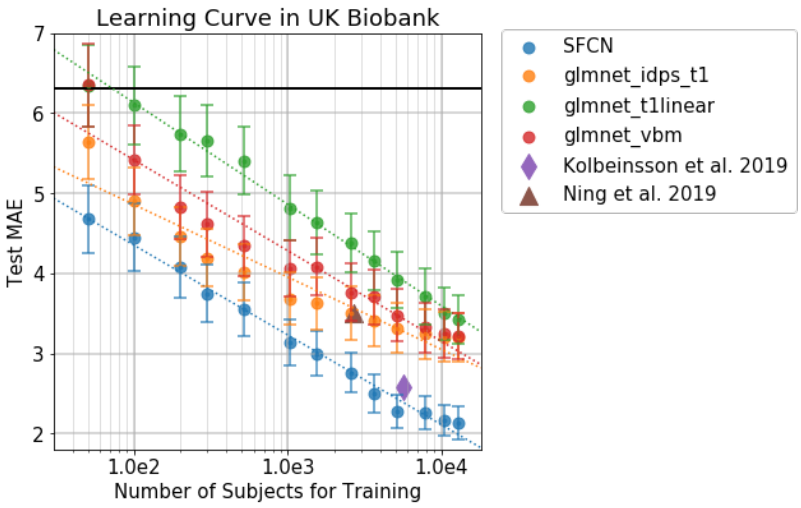

Figure 2. Learning curve for SFCN in UK Biobank data. The methods include SFCN, Elastic nets (glmnet) with different input features, and existing studies. The error bars show the standard deviation by 1000 bootstrap samples. The dashed lines show the log-linear relationship between the training set size and the testing MAE. This shows that as the dataset size doubles, the MAE decrease by around 0.3 to 0.4 years for the linear and the deep learning models. The horizontal dark solid line indicates the MAE when using the population mean age as the predicted age for every testing subject.

T1 data nonlinearly registered to the MNI space (NonLin), segmented grey matter (GM) and white matter (WM) volumes.

For each of the above 4 modalities, we trained 5 models using different random parameter initializations. To prove the effectiveness of the ensemble strategy without greatly increasing

the computing time, we choose a training dataset size of 2,590 subjects for all the models used in this section.

Models trained with different modalities achieve comparable performance with small differences in MAE. NonLin achieved the best MAE (2.73 years), while the Lin and GM achieved comparable MAE of 2.80 years. These modalities are all better than the MAE for WM (2.86 years), as shown in Table 4.

Modality	**Performance**			
	Single Model		**Ensemble**	
	MAE (yrs)	**r value**	**MAE (yrs)**	**r value**
Raw, linearly registered	2.80±0.03	0.883±0.003	2.71±0.03	0.892±0.002
Raw, non-linearly registered	2.73±0.02	0.890±0.002	2.62±0.03	0.900±0.002
Grey matter	2.80±0.04	0.881±0.003	2.72±0.02	0.888±0.002
White matter	2.86±0.04	0.878±0.003	2.78±0.02	0.887±0.002
All models	2.80±0.06	0.883±0.005	**2.58±0.01**	**0.904±0.001**

Table 4. Performance of models trained/tested with different modalities in the test set of the UK Biobank dataset. 5 models were trained for each modality and used to predict brain age individually. The mean and the standard deviation of the single model performances were computed within each modality. For the ensemble performance, 5 models are randomly selected (with duplications allowed) and the predictions were averaged to give the ensemble prediction. This process is repeated for 1000 times to compute the mean test MAE and the standard deviation. For the final ensemble with all modalities, 5 models are randomly selected (with duplications allowed) within each modality and the 20 selected models were used to make the final prediction. This process is repeated for 1000 times to compute the mean test MAE and the standard deviation.

Even though different modalities may result in similar MAEs, the trained models (and deltas) may contain distinct information. This is shown in the correlation matrix of deltas predicted by each of the 20 models in the test sets in Figure 3A (these correlations are between any two estimates of the $N_{subjects} \times 1$ vector of deltas). Models with the same modalities show higher correlation for the brain-age delta prediction.

To better utilise the information contained within different modalities, we used all four modalities to form an ensemble. For every subject, the 20 models predicted 20 brain ages. The final prediction for the subject was made using the mean of all the predicted ages. This strategy achieved an MAE of 2.58 years, which is 0.22 years better than single model prediction, with 2,590 training subjects.

The success of the ensemble strategy is not only owing to the large number of models, but also to the independent information gathered from different modalities. To illustrate this, we combined every pair of models and plotted the MAE improvement after ensemble averaging against the delta correlation coefficient in Figure 3B. The result clearly shows that the less correlated two models are, the better performance the ensemble will produce. It is also

shown that models trained from different modalities tend to be less correlated. Therefore, combining models from different modalities with complementary information gives the greatest performance enhancement.

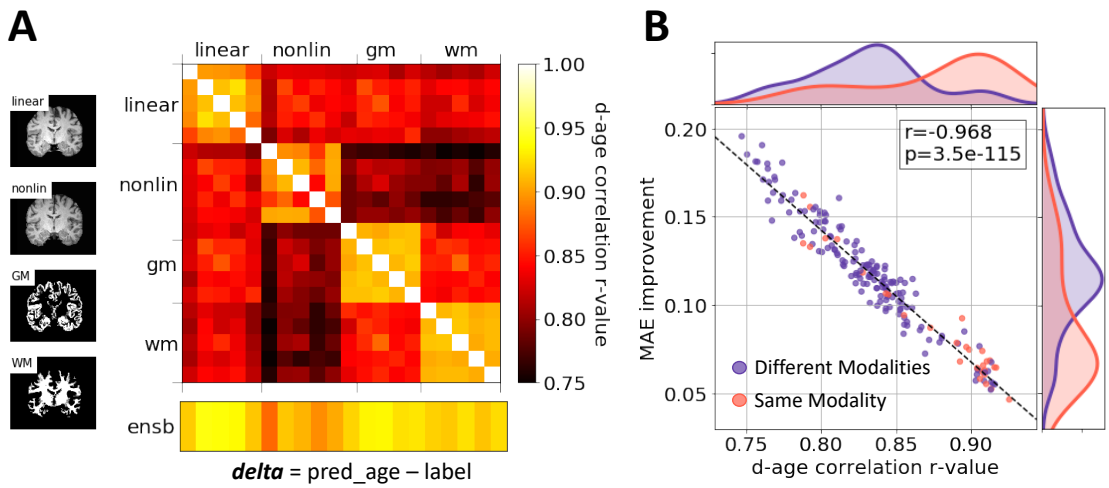

Figure 3. Model ensemble. A) Correlations of brain-age delta predictions between models trained and tested with different modalities. The color coding shows the correlation r-value. For each modality, the training subjects are split into 5 folds, and each model is trained with one-fold being left-out. The delta estimation is made in the common validation set. Any two models trained with the same modality show stronger correlation (between their respective delta estimates) than the models trained with different modalities. The bottom row shows the correlation between individual models and the ensemble prediction. B) Scatter plot: ensemble performance improvement versus delta correlation of any two models. Purple dots represent for ensembles with different modalities. Red dots are for ensembles with the same modality. Normalized histograms of performance improvement and delta correlation of the two-model combinations are plotted alongside. The combination of two models with smaller correlation shows better improvement for the ensemble performance.

4.4 Bias correction

The next challenge is bias correction. We illustrate the age prediction results of SFCN trained with 12,949 subjects in Figure 4. The predictions tend to bias towards the mean age of the cohort, which means that younger subjects will be predicted to be older and vice versa. This is due to regression dilution (MacMahon et al., 1990) and other factors (Smith et al., 2019b), and results in a high correlation between brain-age delta (prediction – age) and chronological age (Spearman's r=-0.39). We followed (Smith et al., 2019b) to regress age out of the delta. In the PAC 2019 competition framework, we do not know the label of the test set. In this case, we regressed out age in the 518-subject validation set and then used the estimated bias correction regression coefficients for bias removal in the test set. This process does not require any knowledge of the age labels in the test set. This reduced the bias Spearman's r-value from -0.37 to 0.03, with an increase of just 0.15 years, in the MAE for the validation set. The generalised strategy (for unlabelled data) reduced the r-value from -0.39 to 0.01, with a small increase (0.11 years) in the MAE for the test set.

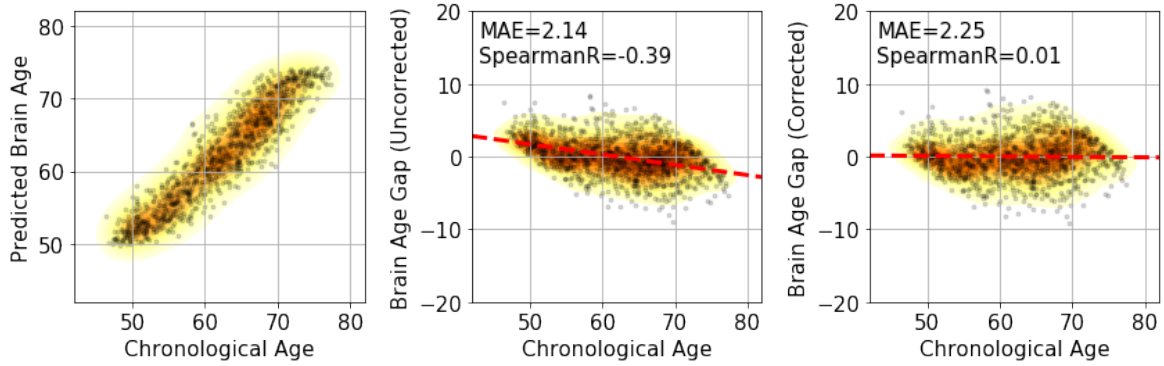

Figure 4. Bias correction. (Left panel) Results of brain age prediction for the UK Biobank test set, SFCN trained with the full training set. (Middle) Results of delta without correction. (Right) Results of delta with correction.

Finally, we tested our methods of SFCN, data augmentation, ensemble and bias correction in the PAC 2019 brain age prediction challenge and achieved first places in both goals of the challenge: 1. to achieve the smallest MAE and 2. To achieve the smallest MAE with bias Spearman's r-value under 0.1. For the first objective, we achieved MAE=2.90 years, which was 0.18/0.42 years better than the second/third places. For the second objective, we achieved MAE=2.95 years, and this result was 0.85/0.97 years better than the second/third places, i.e., a significant improvement (over the other best approaches) of almost one year. These results are available in the challenge website[3].

5 Discussion

To conclude, we proposed SFCN, a lightweight deep neural network architecture, which achieved state-of-the-art brain age prediction using T1-weighted structural MRI images. We investigated different approaches for boosting the performance of the deep learning model, and tested three factors that are valuable for improving the performance of a single deep learning model in a neuroimaging dataset: 1) the lightweight model structure, 2) data augmentation and regularisation techniques (e.g., dropout, voxel shifting and mirroring), 3) large dataset size. For semi-multimodal data (i.e., data from a single modality but which has been through several distinct processing steps), we presented an ensemble strategy that improved single modality results by utilising the (somewhat) independent information from different modalities. Finally, we showed that regressing the true age out of brain-age delta (predicted age minus actual age) can effectively correct bias, and the fitted slope and intercept can be directly transferred to the unknown test set.

[3] Link to the results: https://www.photon-ai.com/pac2019#results Team: BrainAgeDifference

Dataset	Performance		Performance with Bias Correction	
	MAE (years)	Spearman Correlation delta vs age	MAE (years)	Spearman Correlation delta vs age
UK Biobank validation set	2.10	-0.37	2.25	0.03
UK Biobank test set	2.14	-0.39	2.25	0.01
PAC 2019 Brain Age Prediction Challenge1	2.90	-0.39	2.95	-0.03

Table 5. Performance of the ensemble model with and without bias correction. The UK Biobank validation set is used to estimate the slope and intercept from a linear fitting, which is then used to generalized in the unseen UK Biobank test set. This strategy, together with SFCN and the ensemble, was used to take first place in the PAC 2019 Brain Age Prediction Challenge.

We have demonstrated that with a well-trained model, deep learning method can achieve better performance than the simpler regression models tested. While a few studies have shown the opposite conclusion (that linear regression outperforms deep learning in neuroimaging data (Schulz et al., 2019), we argue that DL method is a large family of algorithms and techniques, some more suitable than others for neuroimaging. Different choices of model architectures and training strategies can result in very different results. In our study, we successfully demonstrated the effectiveness of SFCN, as an example of a DL method, in UK Biobank, one of the largest neuroimaging datasets. We also show that even in a training set size as small as 50 subjects, SFCN can outperform simpler regression models. These results are successful explorations of the application of DL in the neuroimaging data.

6 Acknowledgements

We are grateful for funding from the Wellcome Trust. This project is supported by the DeepMedicine project in the Oxford Martin School and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777394 (for AIMS-2-TRIALS) which receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI. This research has been conducted in part using the UK Biobank Resource under Application 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Arslan, S., Ktena, S.I., Glocker, B., Rueckert, D., 2018. Graph Saliency Maps through Spectral Convolutional Networks: Application to Sex Classification with Brain Connectivity, in: Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities. pp. 3–13. https://doi.org/https://doi.org/10.1007/978-3-030-00689-1_1

Baumgartner, C.F., Koch, L.M., Tezcan, K.C., Ang, J.X., 2018. Visual Feature Attribution Using Wasserstein GANs, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8309–8319. https://doi.org/10.1109/CVPR.2018.00867

Chekroud, A.M., Ward, E.J., Rosenberg, M.D., Holmes, A.J., 2016. Patterns in the human brain mosaic discriminate males from females. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1523888113

Cole, J.H., Franke, K., 2017. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends Neurosci. 40, 681–690. https://doi.org/10.1016/j.tins.2017.10.001

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2017. NeuroImage Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage 163, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059

Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2018. Brain age predicts mortality. Mol. Psychiatry 23, 1385–1392. https://doi.org/10.1038/mp.2017.62

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24, 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Feng, X., Lipton, Z.C., Yang, J., Small, S.A., Provenzano, F.A., 2019. Estimating brain age based on a healthy population with deep learning and structural MRI. arXiv Prepr. arXiv1907.00943.

Franke, K., Gaser, C., 2019. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? Front. Neurol. 10. https://doi.org/10.3389/fneur.2019.00789

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i01

Giudice, M. Del, Lippa, R.A., Puts, D.A., Bailey, D.H., Bailey, J.M., Schmitt, D.P., 2016. Joel et al.'s method systematically fails to detect large, consistent sex differences. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1525534113

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in:

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778.
He, T., Kong, R., Holmes, A., Nguyen, M., Sabuncu, M., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2018. Do Deep Neural Networks Outperform Kernel Regression for Functional Connectivity Prediction of Behavior? bioRxiv 473603. https://doi.org/10.1101/473603
He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2019. Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics. bioRxiv 473603. https://doi.org/10.1101/473603
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: 32nd International Conference on Machine Learning, ICML 2015. pp. 448–456.
Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., Shefi, N., Pool, J., Urchs, S., Margulies, D.S., Liem, F., Hänggi, J., Jäncke, L., Assaf, Y., 2015. Sex beyond the genitalia: The human brain mosaic. Proc. Natl. Acad. Sci. U. S. A. 112, 15468–15473. https://doi.org/10.1073/pnas.1509654112
Joel, D., Persico, A., Hänggi, J., Pool, J., Berman, Z., 2016. Do brains of females and males belong to two distinct populations? Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1600792113
Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., Martinot, J.L., Paus, T., Smolka, M.N., Walter, H., Schumann, G., Garavan, H., Whelan, R., 2019. Quantifying performance of machine learning methods for neuroimaging data. Neuroimage 199, 351–365. https://doi.org/10.1016/j.neuroimage.2019.05.082
Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M.K., Bøen, E., Borgwardt, S., Brandt, C.L., Buitelaar, J., Celius, E.G., Cervenka, S., Conzelmann, A., Córdova-Palomera, A., Dale, A.M., de Quervain, D.J.F., Carlo, P., Djurovic, S., Dørum, E.S., Eisenacher, S., Elvsåshagen, T., Espeseth, T., Fatouros-Bergman, H., Flyckt, L., Franke, B., Frei, O., Haatveit, B., Håberg, A.K., Harbo, H.F., Hartman, C.A., Heslenfeld, D., Hoekstra, P.J., Høgestøl, E.A., Jernigan, T.L., Jonassen, R., Jönsson, E.G., Kirsch, P., Kłoszewska, I., Kolskår, K.K., Landrø, N.I., Hellard, S., Lesch, K.-P., Lovestone, S., Lundervold, A., Lundervold, A.J., Maglanoc, L.A., Malt, U.F., Mecocci, P., Melle, I., Meyer-Lindenberg, A., Moberget, T., Norbom, L.B., Nordvik, J.E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pauli, P., Pergola, G., Persson, K., Richard, G., Rokicki, J., Sanders, A.-M., Selbæk, G., Shadrin, A.A., Smeland, O.B., Soininen, H., Sowa, P., Steen, V.M., Tsolaki, M., Ulrichsen, K.M., Vellas, B., Wang, L., Westman, E., Ziegler, G.C., Zink, M., Andreassen, O.A., Westlye, L.T., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat. Neurosci. 22, 1617–1623. https://doi.org/10.1038/s41593-019-0471-7
Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146, 1038–1049. https://doi.org/10.1016/j.neuroimage.2016.09.046
Kohl, S.A.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Ali Eslami, S.M., Rezende, D.J., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images, in: Advances in Neural Information Processing Systems. pp. 6965–6975.
Kolbeinsson, A., Kossaifi, J., Panagakis, Y., Bulat, A., Anandkumar, A., Tzoulaki, I., Matthews,

P., 2019. Robust Deep Networks with Randomized Tensor Regression Layers. arXiv Prepr. arXiv1902.10758.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1097–1105.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539
MacMahon, S., Peto, R., Collins, R., Godwin, J., MacMahon, S., Cutler, J., Sorlie, P., Abbott, R., Collins, R., Neaton, J., Abbott, R., Dyer, A., Stamler, J., 1990. Blood pressure, stroke, and coronary heart disease: Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 335, 765–774. https://doi.org/10.1016/0140-6736(90)90878-9
Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536. https://doi.org/10.1038/nn.4393
Ning, K., Zhao, L., Matloff, W., Sun, F., Toga, A.W., 2018. Association of brain age with smoking , alcohol consumption , and genetic variants. bioRxiv. https://doi.org/10.1101/469924
Rosenblatta, J.D., 2016. Multivariate revisit to “sex beyond the genitalia.” Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1523961113
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y
Schulz, M., Yeo, B.T.T., Vogelstein, J.T., Mourao-, J., Kather, J.N., Kording, K., Richards, B., Bzdok, D., 2019. Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. bioRxiv 757054.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Prepr. arXiv1409.1556.
Smith, S.M., Elliott, L.T., Alfaro-Almagro, F., McCarthy, P., Nichols, T.E., Douaud, G., Miller, K.L., 2019a. Brain aging comprises multiple modes of structural and functional change with distinct genetic and biophysical associations. bioRxiv 802686. https://doi.org/10.1101/802686
Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., Miller, K.L., 2019b. Estimation of brain age delta from brain imaging. Neuroimage 200, 528–539. https://doi.org/10.1016/j.neuroimage.2019.06.017
Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research.
Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning, in: 30th International Conference on Machine Learning, ICML 2013. pp. 2176–2184.
Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T., 2018. Fully Convolutional Adaptation Networks for Semantic Segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 6810–6818.

https://doi.org/10.1109/CVPR.2018.00712