SQET-MOE: TWO-STAGE BRAIN AGE ESTIMATION USING SQUEEZE-AND-EXCITATION TRANSFORMER AND MIXTURE-OF-EXPERTS

Rizuki Oura, Koichi Ito, and Takafumi Aoki

Graduate School of Information Sciences, Tohoku University,
6–6–05 Aramaki Aza Aoba, Aoba-ku, Sendai-shi, Miyagi 980–8579 Japan.

ABSTRACT

Brain age estimation is crucial for identifying brain disorders and developing biomarkers. While deep learning methods have been proposed for high-accuracy age estimation from T1-weighted images, further improvements are still needed. This paper proposes a novel brain age estimation method called SQET-MoE, utilizing a two-stage estimation framework. In the 1st stage, the Squeeze-and-Excitation Transformer (SQET), which fuses the benefits of CNNs and Transformers, performs coarse age estimation and feature extraction. The 2nd stage employs a Mixture-of-Experts (MoE) module, which uses SQET's output to estimate and correct the residual associated with systematic errors and complex non-linearity. This task division stabilizes training and maximizes estimation accuracy. Through a set of experiments using large-scale datasets, we demonstrate that the proposed SQET-MoE achieves the highest estimation accuracy compared to conventional methods.

***Index Terms*—** age estimation, brain MRI, 3DCNN, deep learning

1. INTRODUCTION

Statistical analysis of large-scale databases has shown that the human brain atrophies with normal aging [1]. Utilizing this tendency, the "brain age" estimated from MRI serves as an objective biomarker for assessing morphological changes. Evaluating the age gap between the estimated and chronological age helps quantify deviations from normal aging, supporting the diagnosis of diseases that alter brain morphology, such as Alzheimer's disease.

In age estimation based on brain morphology, brain MRI is widely used as the standard modality. While MRI can acquire various types of images by altering imaging conditions, T1-weighted images are usually utilized for age estimation due to its ability to easily capture the brain's anatomical structures. Major age estimation methods from T1-weighted images are based on Convolutional Neural Networks (CNNs) [2] [3–8]. Although CNNs are useful for extracting anatomical patterns through local receptive fields and hierarchical pooling, convolution used in CNNs is limited to local feature extraction, making it difficult to capture long-range dependencies between spatially distant regions. To address this issue, Transformer [9], which is fundamentally based on the self-attention mechanism, is employed as a feature extractor for brain age estimation. Transformer has achieved significant success in the fields of natural language processing and computer vision due to its ability to directly model long-range dependencies. Indeed, the introduction of Transformer to brain age estimation has been shown to contribute to the capture of wide-area dependency relationships [10, 11].

In this paper, to improve the accuracy of brain age estimation, we propose SQET-MoE, which adopts the two-stage-age-network (TSAN) framework [8]. In the first stage, we employ Squeeze-and-Excitation Transformer (SQET) [11] for coarse age estimation, leveraging its ability to simultaneously capture local anatomical patterns and global long-range dependencies. For the second stage, we introduce a Mixture-of-Experts (MoE) module to estimate the residual between the coarse prediction and the chronological age. By inputting the global features and the estimated age from SQET, the MoE utilizes a gating mechanism that dynamically assigns optimal weights to multiple experts for each sample. This enables sample-specific residual correction that effectively addresses systematic errors and non-linear biases across diverse cohorts. Experiments on large-scale public (IXI[1], ADNI [12]) and private datasets (Aoba-1 and Tsurugaya-1 datasets [13]) demonstrate that SQET-MoE significantly outperforms existing state-of-the-art methods.

2. METHOD

This section describes the details of the proposed brain age estimation method, SQET-MoE.

2.1. Overview

Fig. 1 illustrates the overview of SQET-MoE. The proposed method adopts a two-stage architecture inspired by TSAN [8]. In the 1st stage, volume data is input to SQET, which outputs

[1]https://brain-development.org/ixi-dataset/

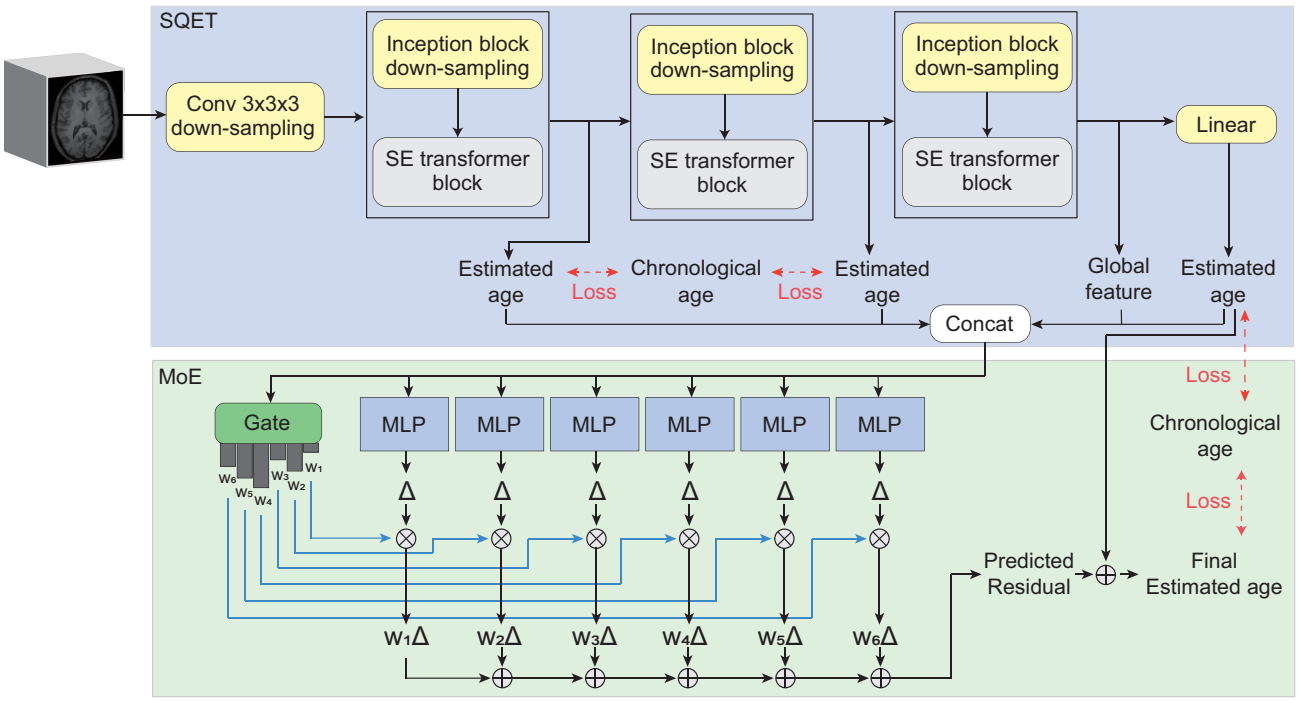

Fig. 1. Overview of SQET-MoE, which consists of SQET for coarse age estimation in the first stage and MoE for residual correction in the second stage.

a global feature vector and a coarse estimated age. In addition, an intermediate estimated age is output from each intermediate layer of SQET by auxiliary regression. In the 2nd stage, the global feature vector, the estimated age from the last layer and intermediate layers obtained from the 1st stage are concatenated and input into MoE. MoE is then used to estimate the residual Δ between the chronological age and the estimated age in the 1st stage. Finally, the final estimated age is output by summing the estimated residual and the estimated age in the 1st stage. The following subsections describe the details of the 1st and 2nd stages, as well as the loss function used for training SQET-MoE.

2.2. 1st Stage: SQET

SQET has a 3D hierarchical architecture that combines a down-sampling block, a lightweight 3D Inception block, and an SE-Transformer block. The down-sampling block consists of a 3D convolutional layer, a batch normalization layer, and an activation function. The Inception block extracts local features, and the SE-Transformer block performs re-calibration based on wide-area context. In the SE-Transformer block, given the input $X \in \mathbb{R}^{D \times H \times W \times C}$ and a patch size P, the input feature map is first compressed into $X_s \in \mathbb{R}^{\frac{D}{P} \times \frac{H}{P} \times \frac{W}{P} \times C}$. Specifically, 3D average pooling with a kernel size P and stride P is used for this compression. Subsequently, local-to-local attention information X_a is obtained using a standard self-attention mechanism and a Sigmoid activation function. X_a is then combined with the original input X using an element-wise Hadamard product with broadcasting. The down-sampling block, Inception block, and SE-Transformer block are defined as one stage, and this process is repeated three times. Finally, the global feature vector and the estimated age are obtained after global average pooling and a fully-connected layer. Auxiliary regression is applied to the intermediate feature maps obtained at each stage to output the corresponding intermediate estimated age.

2.3. 2nd Stage: MoE

In MoE, the input vector is a concatenation of the global feature vector and the estimated age output by SQET, and all intermediate estimated ages. MoE estimates the residual Δ between the chronological age and the estimated age in the 1st stage. By including the intermediate estimated ages, MoE considers age-specific trends and layer-wise difference information, enabling stable and high-precision residual correction. The gating mechanism in MoE consists of LayerNorm (LN) and Linear layers, calculating the logit for each expert, and obtaining the mixed weights w via a Softmax function with a temperature parameter. The expert network uses a small MLP composed of LN, Linear, GELU, Dropout, and Linear layers. Each expert i estimates an individual residual Δ_i, and the final residual is estimated by a weighted sum using the weights obtained from the gate. The final estimated age is then obtained by summing the estimated residual and

Table 1. Specification of the T1-weighted MRI datasets used in the experiments.

Dataset	$N_{samples}$	Age range	Mean age [y/o]
IXI/ADNI	1,175	20.0-–92.8	62.3
A1/T1	1,214	13.1—82.3	48.7

the estimated age in the 1st stage.

2.4. Loss Function

The loss function used for training SQET-MoE is the Smooth L1 loss applied to the final estimated age, the age estimated at the final layer of SQET, and the ages estimated at the intermediate layers, all with respect to the chronological age. We denote the loss for the final estimated age as L_{main}, the loss for the final layer of SQET as L_{coarse}, and the losses for the intermediate layers of SQET as L_{aux1} and L_{aux2}, respectively. The total loss function L used for training is defined by

$$L = L_{main} + w_{coarse}L_{coarse} + w_{aux}(L_{aux1} + L_{aux2}), \quad (1)$$

where w_{coarse} and w_{aux} are weights for the respective loss terms. In this paper, we set $w_{coarse} = 0.3$ and $w_{aux} = 0.15$. The use of this loss function facilitates the stabilization of the initial training and promotes effective gradient propagation to the deeper layers.

3. EXPERIMENTS

This section describes the performance evaluation of SQET-MoE using T1-weighted image datasets.

3.1. Datasets and Preprocessing

We use the IXI/ADNI combined dataset and the Aoba-1/Tsurugaya-1 (A1/T1) dataset [13] in the experiments. The IXI/ADNI dataset is created by integrating T1-weighted images from the IXI dataset[1] and the ADNI dataset [12]. The details of each dataset are presented in Table 1. Each dataset is split into 70% for the training dataset, 15% for the validation dataset, and the remaining 15% for the test dataset. For preprocessing, we apply registration and normalization to the MNI152NLin2009cAsym standard space using SPM12[2]. The input image for each subject scales the voxel values to the range $[0, 1]$ by min-max normalization, and is then resized to $128 \times 128 \times 128$ voxels.

3.2. Implementation Details

The AdamW optimizer [14] is used for the simultaneous optimization of all parameters. The training process supports mixed precision (AMP) and gradient accumulation. The initial learning rate is set to 3×10^{-4}, and a WarmupCosine scheduler is applied, which decays the learning rate toward 0 after a linear warmup phase. The number of epochs is set to 100. For data augmentation during training, RandomResizedCrop3D (scale 0.8-1.0) is applied to the input images, where a random region is cropped and resized to $128 \times 128 \times 128$ voxels. In addition, a lightweight 3D data augmentation method using axis flipping is applied.

[2]https://www.fil.ion.ucl.ac.uk/spm/

Table 2. Ablation study on the number of MLP experts.

Dataset	# of experts	MAE [y/o]	r	$CS@5$ [%]
IXI/ADNI	1	3.407	0.950	70.31
	2	3.418	0.944	68.99
	4	3.364	0.940	70.49
	6	**3.109**	**0.979**	**81.80**
	8	3.440	0.948	68.10
	10	3.524	0.928	65.01

Table 3. Experimental results for comparing brain age estimation by conventional methods and SQET-MoE.

Dataset	Method	MAE [y/o]	r	$CS@5$ [%]
IXI / ADNI	Ueda3DNet [5]	3.620	0.941	72.75
	ScaledDenseNet [8]	3.817	0.930	71.98
	TSAN [8]	3.422	0.938	75.89
	GLT [10]	5.454	0.920	51.23
	SQET [11]	4.322	0.939	67.33
	SQET-MoE	**3.109**	**0.979**	**81.80**
A1/T1	Ueda3DNet [5]	3.351	0.945	70.12
	ScaledDenseNet [8]	3.604	0.942	68.33
	TSAN [8]	3.210	0.953	78.81
	GLT [10]	4.831	0.929	65.22
	SQET [11]	3.862	0.965	72.04
	SQET-MoE	**2.967**	**0.988**	**82.31**

3.3. Evaluation Metrics

Performance is evaluated using the Mean Absolute Error (MAE), the Pearson correlation coefficient (r), and the Cumulative Score (CS). MAE quantifies the average prediction error, while r measures the linear correlation between estimated and chronological ages. CS is the percentage of samples within an error threshold α, defined as $CS(\alpha) = N_{e \leq \alpha}/N \times 100\%$, where $N_{e \leq \alpha}$ is the number of samples with errors within α. We set $\alpha = 5$ (CS@5) following the protocol in [11].

3.4. Ablation Study

To investigate the impact of the number of MLP experts in MoE, we conducted an ablation study by varying the expert count on the IXI/ADNI dataset. In this experiment, common conditions were fixed, including data splitting, preprocessing,

training settings, gate structure, MLP shape per expert, and loss weighting. Note that setting the number of experts to one is equivalent to residual estimation by a single MLP. The results are shown in Table 2. The highest accuracy was observed with six experts. Conversely, accuracy decreased when the number of experts was too small or too large. With few experts, the model could not sufficiently represent the complex non-linearity of the residual, leading to performance saturation. Furthermore, with a large number of experts, the mixing via temperature-controlled Softmax caused the gate weights to disperse across multiple experts. This resulted in smaller gradients for each expert, making training convergence difficult and lowering accuracy. We conclude that approximately six experts offer the optimal balance between non-linear expressivity and training stability.

3.5. Comparison with Conventional Methods

To demonstrate the effectiveness of SQET-MoE, we compared it against 3D-CNN based methods (Ueda3DNet [5], ScaledDenseNet [8], TSAN [8]) and Transformer-based methods (GLT [10], SQET [11]). All models were trained on the same dataset with optimized hyperparameters. The results are shown in Table 3. First, SQET outperformed the other comparative methods. This is attributed to its design fusing CNN and Transformer benefits, effectively utilizing both fine-grained local and spatially distant features. Notably, the proposed SQET-MoE achieved the highest accuracy among all methods. Integrating MoE for residual correction consistently improved performance. This improvement stems from effective task division: SQET performs coarse estimation, while MoE compensates for systematic errors and complex non-linearities. Additionally, by optimizing expert weights based on input data, MoE effectively accounts for characteristics of unlabeled potential subgroups, such as age or cohort differences.

4. CONCLUSION

In this paper, we proposed SQET-MoE to achieve high-accuracy brain age estimation. The proposed method consists of a two-stage estimation framework: coarse age estimation by SQET, which fuses the benefits of CNNs and Transformers, and residual correction for systematic errors by MoE. Through a set of experiments, we demonstrated that SQET-MoE achieved the highest estimation accuracy compared to other methods.

5. REFERENCES

[1] Y. Taki, B. Thyreau, S. Kinomura, K. Sato, R. Goto, R. Kawashima, and H. Fukuda, "Correlations among brain gray matter volumes, age, gender, and hemisphere in healthy individuals," *PLoS ONE*, vol. 6, no. 7, pp. e22734–1–e22734–13, Apr. 2011.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, The MIT Press, 2016.

[3] T. W. Huang, H. T. Chen, R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki, "Age estimation from brain MRI images using deep learning," *Proc. Int'l Symp. Biomed. Imaging*, pp. 849–852, Apr. 2017.

[4] J. H. Cole, R. P. K. Poudel, D. Tsagkrasoulis, M. W. A. Caan, C. Steves, T. D. Spector, and G. Montana, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *NeuroImage*, vol. 163, pp. 115–124, Dec. 2017.

[5] M. Ueda, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki, "An age estimation method using 3D-CNN from brain MRI images," *Proc. Int'l Symp. Biomed. Imaging*, pp. 380–383, Apr. 2019.

[6] H. Sajedi and N. Pardakhti, "Age prediction based on brain MRI image: A survey," *Image & Signal Processing*, vol. 43, no. 279, pp. 1–30, July 2019.

[7] M. Tanveer, M. A. Ganaie, I. Beheshti, T. Goel, N. Ahmad, K.-T. Lai, K. Huang, Y.-D. Zhang, J. Del Ser, and C.-T. Lin, "Deep learning for brain age estimation: A systematic review," *Information Fusion*, vol. 96, pp. 130–143, Aug. 2023.

[8] J. Cheng, Z. Liu, H. Guan, Z. Wu, H. Zhu, J. Jiang, W. Wen, D. Tao, and T. Liu, "Brain age estimation from MRI using cascade networks with ranking loss," *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3400–3412, June 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inform. Process. Syst.*, pp. 1–11, Dec. 2017.

[10] S. He, PE. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE Trans. Med. Imaging*, vol. 41, no. 1, pp. 213–224, Jan. 2022.

[11] Y. Hu, H. Wang, and B. Li, "SQET: Squeeze and excitation transformer for high-accuracy brain age estimation," *Proc. IEEE Int'l Conf. Bioinformatics and Biomedicine*, pp. 1554–1557, Dec. 2022.

[12] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods.," *J. Magn. Reson. Imaging*, pp. 685–691, Feb. 2008.

[13] K. Sato, H. Fukuda, and R. Kawashima, "Neuroanatomical database of normal Japanese brains," *Neural Networks*, vol. 16, no. 9, pp. 1301–1310, Nov. 2003.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Int. Conf. Learn. Represent.*, pp. 1–8, May 2019.