MABAL: a Novel Deep-Learning Architecture for Machine-Assisted Bone Age Labeling

Simukayi Mutasa¹ · Peter D. Chang¹ · Carrie Ruzal-Shapiro¹ · Rama Ayyala¹

Published online: 5 February 2018
© Society for Imaging Informatics in Medicine 2018

Abstract

Bone age assessment (BAA) is a commonly performed diagnostic study in pediatric radiology to assess skeletal maturity. The most commonly utilized method for assessment of BAA is the Greulich and Pyle method (Pediatr Radiol 46.9:1269–1274, 2016; Arch Dis Child 81.2:172–173, 1999) atlas. The evaluation of BAA can be a tedious and time-consuming process for the radiologist. As such, several computer-assisted detection/diagnosis (CAD) methods have been proposed for automation of BAA. Classical CAD tools have traditionally relied on hard-coded algorithmic features for BAA which suffer from a variety of drawbacks. Recently, the advent and proliferation of convolutional neural networks (CNNs) has shown promise in a variety of medical imaging applications. There have been at least two published applications of using deep learning for evaluation of bone age (Med Image Anal 36:41–51, 2017; JDI 1–5, 2017). However, current implementations are limited by a combination of both architecture design and relatively small datasets. The purpose of this study is to demonstrate the benefits of a customized neural network algorithm carefully calibrated to the evaluation of bone age utilizing a relatively large institutional dataset. In doing so, this study will aim to show that advanced architectures can be successfully trained from scratch in the medical imaging domain and can generate results that outperform any existing proposed algorithm. The training data consisted of 10,289 images of different skeletal age examinations, 8909 from the hospital Picture Archiving and Communication System at our institution and 1383 from the public Digital Hand Atlas Database. The data was separated into four cohorts, one each for male and female children above the age of 8, and one each for male and female children below the age of 10. The testing set consisted of 20 radiographs of each 1-year-age cohort from 0 to 1 years to 14–15+ years, half male and half female. The testing set included left-hand radiographs done for bone age assessment, trauma evaluation without significant findings, and skeletal surveys. A 14 hidden layer-customized neural network was designed for this study. The network included several state of the art techniques including residual-style connections, inception layers, and spatial transformer layers. Data augmentation was applied to the network inputs to prevent overfitting. A linear regression output was utilized. Mean square error was used as the network loss function and mean absolute error (MAE) was utilized as the primary performance metric. MAE accuracies on the validation and test sets for young females were 0.654 and 0.561 respectively. For older females, validation and test accuracies were 0.662 and 0.497 respectively. For young males, validation and test accuracies were 0.649 and 0.585 respectively. Finally, for older males, validation and test set accuracies were 0.581 and 0.501 respectively. The female cohorts were trained for 900 epochs each and the male cohorts were trained for 600 epochs. An eightfold cross-validation set was employed for hyperparameter tuning. Test error was obtained after training on a full data set with the selected hyperparameters. Using our proposed customized neural network architecture on our large available data, we achieved an aggregate validation and test set mean absolute errors of 0.637 and 0.536 respectively. To date, this is the best published performance on utilizing deep learning for bone age assessment. Our results support our initial hypothesis that customized, purpose-built neural networks provide improved performance over networks derived from pre-trained imaging data sets. We build on that initial work by showing that the addition of state-of-the-art techniques such as residual connections and inception architecture further improves prediction accuracy. This is important because the current assumption for use of residual and/or inception architectures is that a large pre-trained network is required

✉ Simukayi Mutasa
s.mutasa@columbia.edu

¹ Columbia University Medical Center, PB 1-301, New York, NY 10032, USA

for successful implementation given the relatively small datasets in medical imaging. Instead we show that a small, customized architecture incorporating advanced CNN strategies can indeed be trained from scratch, yielding significant improvements in algorithm accuracy. It should be noted that for all four cohorts, testing error outperformed validation error. One reason for this is that our ground truth for our test set was obtained by averaging two pediatric radiologist reads compared to our training data for which only a single read was used. This suggests that despite relatively noisy training data, the algorithm could successfully model the variation between observers and generate estimates that are close to the expected ground truth.

Keywords Machine learning · Deep learning · Convolutional neural networks · Radiology · Pediatric radiology · Endocrinology

Introduction

Bone age assessment (BAA) is a commonly performed study in pediatric radiology to assess skeletal maturity. This is essential in the diagnostic workup for endocrine disorders and other disorders involving abnormal skeletal development. There are various methods available for evaluation of BAA, which all involve visual inspection of the developing ossification centers. A recent survey of the Society of Pediatric Radiology showed that the most commonly utilized method for assessment of BAA is the Greulich and Pyle method [1, 2]. The Greulich and Pyle method is an atlas of normal hand and wrists at various chronological ages. The patient's radiograph is compared to the age-appropriate image in the atlas, with the BAA determined by the shape and distribution of the patient's developing ossification centers. The predicted bone age is compared to the patient's chronological age, and these results are subsequently used by the clinical team to assess appropriate skeletal maturity.

The evaluation of BAA can be tedious for the radiologist, and not only provides very little in the way of intellectual pursuit, but is time-consuming. Compounding the problems with the classical method of evaluating BAA is the fact that there is a considerable amount of interobserver and intraobserver variation in the ultimate reads [3–5]. As a result, several computer-assisted detection/diagnosis (CAD) methods have been proposed for automation of BAA [6–8].

Classical CAD tools have traditionally relied on hard-coded algorithmic features for BAA. Although these techniques have shown promise in accurate BAA, they suffer from a variety of drawbacks: (1) high-order semantic features (“dog” and “cat” or “shape of ossification centers”) easily interpreted by humans are almost impossible to translate cleanly into rigid mathematical formulae; (2) hard-coded features are biased to focus on features seen useful by the human eye and not necessarily the features that may subtly change over time and would be better employed for the question at hand; (3) feature extraction is a tedious and computationally expensive process.

Recently, the advent and proliferation of convolutional neural networks (CNNs), a deep learning technique, has attracted interest in medical imaging analysis. Deep-learning

techniques overcome many of these limitations by allowing an algorithm to automatically learn the features useful for image interpretation without any direct human intervention during the training process. Early studies have already shown promise in a variety of medical imaging applications including evaluation of lung nodules, interstitial lung disease, breast cancer, cerebral microbleeds, brain tumors, and spinal metastasis [9–15].

Additionally, there have been at least two published applications of using deep learning for evaluation of bone age [6, 8]. However, current implementations are limited by a combination of both architecture design and relatively small datasets. These factors make it such that most deep-learning algorithms for medical imaging are unable to take full advantage of the latest state-of-the-art advances in network design (e.g., residual connections, inception architecture) or are only able to do so using large networks pre-trained on non-medical images. By contrast, the purpose of this study is to demonstrate the benefits of a customized neural network algorithm carefully calibrated to the evaluation of bone age utilizing a relatively large institutional dataset. In doing so, this study will aim to show that advanced architectures can be successfully trained from scratch in the medical imaging domain and can generate results that outperform any existing proposed algorithm.

Materials and Methods

Data Preparation

IRB approval was obtained for this retrospective study. A query of the Picture Archiving and Communication System (PACS) for bone age studies performed between December 2009 and October 2016 was performed. A total of 9188 radiographic images of left hands were included in the data set.

The data set was split into a training, validation, and testing set. The training set is the set of images that is fed into the network that it learns from. The validation set is a separate set of images that is utilized to optimize the neural network parameters. After training the network on the training set and fine tuning the network parameters based on its performance

on the validation set, a separate run is done utilizing the entirety of the training and validation data as training data. This second run is then tested on the testing set, which was separated from the rest of the data at the outset and represents images that the network has never seen before.

Images from the data set were excluded if they contained missing reports or if the report contained language that indicated an abnormality on the radiograph, not related to the patient's bone age, that prevented analysis of bone age. The skeletal age was automatically extracted from the radiology report. An additional 1383 images from a publicly available and double-annotated Digital Hand Atlas Database [16] were obtained to augment the training set. None of the publicly available data was utilized in the test or validation set. For the skeletal age on the public images, the average of two radiologist reads was utilized. In total, 10,289 images and skeletal ages were available for the network's data set, 8909 of which came from the hospital PACS and 1383 from the public database. Eightfold cross-validation was performed with 1105 of the internal PACS images set aside as a validation set utilized for tuning network hyperparameters.

Three hundred images were collected to construct the testing set. The testing set consisted of 10 images of boys and 10 images of girls for each 1-year-age cohort from newborn children to 16 years of age, the age where each gender reaches skeletal maturity. To construct the testing set without reducing the examples available in the training data, studies that were not identified by our automated extraction pipeline were used after manual inspection. The origin of these studies included the following:

1. Studies that were stored in PACS as secondary captures
2. Studies with a typographic error in a critical section of the report, preventing correct labeling
3. Studies performed between the date of the original database query and construction of the testing set

Finally, due to the lack of adequate numbers of children from age 0 to 6 years old in the available testing pool, normal images of left hands obtained from children undergoing hand imaging as part of a trauma workup or skeletal surveys were utilized to fill in the test set. On manual inspection, there were no qualitative differences between the radiographs performed for bone age assessment and those performed for other reasons once all external labeling and images containing medical instrumentation were removed.

Data Augmentation and Segregation

Raw images were fed into the neural network without preprocessing. To perform data augmentation, queued images were randomly flipped vertically and/or horizontally, rotated by a random angle between $+0.78$ and -0.78 rad, and randomly

cropped to a box 80% of their initial size. The pixel values of the data set were divided by the standard deviation of the data set and subtracted by the mean of the data set.

There is no known logical correlation between the rate of skeletal maturity for boys and the rate of skeletal maturity for girls. Additionally, skeletal maturity during puberty was thought to be a point of difficulty for neural networks to overcome. To combat these factors, four different networks with the same architecture but different age- and gender-separated data sets and hyperparameters were designed. The first and second networks were trained on female and male children between 0 and 10 years old. The third and fourth networks were trained on female and male children between 8 years old and skeletal maturity. Of note, there was a 2-year overlap for each age cohort to account for children with a chronological age on the boundary of an age cohort but with a bone age that is advanced or delayed for their age. The final training sets included 4093 girls for the older girl network, 2309 girls for the younger girl network, 3259 boys for the older boy network, and 1346 boys for the younger boy network.

A customized CNN architecture was designed and trained from random weight initializations for evaluation of bone age. Originally described by LeCun et al. [17] in 1998, CNNs are implemented by applying a series of convolution matrices to a vectorized input image that iteratively separates the input to a target vector space. Building on this original design, our goal was to create a network architecture that would balance the use of state-of-the-art strategies while keeping the overall number of trainable parameters as small as possible given the relatively small datasets seen in medical imaging. After a hyperparameter search of possible architectures, our final trained network incorporates a combination of residual connections, inception architecture, and an explicit spatial transformation module (Table 1).

After an initial standard convolutional layer, a series of residual layers are utilized in the first portion of the network (Fig. 1). Originally described by He et al. [18, 19], residual neural networks can stabilize gradients during back propagation, leading to improved optimization and facilitating greater network depth. Beginning with the eighth hidden layer, residual inception style layers are utilized (Fig. 2). The inception layer architecture, initially described in 2015 by Szegedy et al. [20], introduced a computationally efficient method of allowing a network to selectively determine the optimal filter architectures to an input feature map, leading to improved learning rates. Compared to a base architecture composed entirely of residual layers, the addition of inception layers overall improves performance (Table 2). Furthermore, inserting inception layers deeper in the network has a twofold advantage over early inception layer use: (1) the benefits of evaluating input feature maps at higher resolutions are more pronounced with the more complex feature maps found deeper in the network; (2) additionally, the gains in parameter reduction when

Table 1 The network architecture

Input layer	Input layer dimensions	Filter type	Filter size	Output layer
Input	$256 \times 256 \times 1$	Convolutional	$3 \times 3 \times 32$	Hidden layer 1
Hidden layer 1	$256 \times 256 \times 16$	Residual	$3 \times 3 \times 64$	Hidden layer 2/3
Hidden layer 2/3	$128 \times 128 \times 64$	Residual	$3 \times 3 \times 128$	Hidden layer 4/5
Hidden layer 4/5	$64 \times 64 \times 128$	Residual	$3 \times 3 \times 256$	Hidden layer 5/6
Hidden layer 5/6	$32 \times 32 \times 256$	Residual	$3 \times 3 \times 512$	Hidden layer 6/7
Hidden layer 6/7	$16 \times 16 \times 512$	Wide residual	$3 \times 3 \times 1024$	Hidden layer 8/9
Hidden layer 8/9	$8 \times 8 \times 1024$	Inception	$\times 1024$	Hidden layer 10
Hidden layer 10	$8 \times 8 \times 1024$	Affine transform	N/A	Hidden layer 11
Hidden layer 11	$8 \times 8 \times 1024$	Inception	$\times 2048$	Hidden layer 12
Hidden layer 12	$4 \times 4 \times 2048$	Fully connected	$\times 128$	Hidden layer 13
Hidden layer 13	1×128	Linear	128×32	Hidden layer 14
Hidden layer 14	1×32	Regression	32×1	Bone age prediction

Dimensions of all of the intermediate layers of the convolutional neural network. The first column contains the input layer names. The second column displays the size of the input feature map. The middle column describes the type of filter applied followed by a column describing the filter size if applicable. “Wide residual” filter types correspond to a residual filter with an initial inception layer. The final column displays the name of the output layer, which serves as the input for the next layer. Residual layers contain two feature maps per layer

using inception layers are more pronounced. These observations were confirmed in our experimental results which showed improved performance when inception layers were introduced at layer 8 as opposed to layer 3 in the network architecture (Table 2).

Subsequently, a spatial transformer module was inserted after the 10th hidden layer. Initially applied to convolutional neural networks by Jaderberg et al. [21], spatial transformer layers allow a network to explicitly learn affine transformation parameters that regularize global spatial variations in feature space. This is important for the network to be robust against large changes in imaging technique or positioning that, upon visual inspection, were commonly seen in both the internal data and public Digital Hand Atlas Database. Spampinato et al. [6] showed that inserting the spatial transformer module deeper in the network leads to improved performance, a result which was replicated in our experimentation (Table 2).

A fully connected layer was implemented after the 12th hidden layer and a final linear regression output layer was inserted as the last layer.

In addition to the proposed customized network and its variations described above, several additional network architectures were tested (Table 2). In keeping with the results from Spampinato et al. [6], these included an initial VGG-16 network pre-trained on the ImageNet database as well as a similar network trained from randomly initialized weights. Additionally, a customized six-layer CNN architecture was reproduced from the analysis in Spampinato et al. [6].

Training was implemented using the Adam optimizer [22] combined with the Nesterov accelerated gradient [23] described by Dozat [24]. Parameters are initialized using the heuristic described by Xavier et al. [25]. L2 regularization is implemented to prevent overfitting of data by limiting the squared magnitude of the kernel weights. Dropout was also employed to prevent overfitting by limiting unit co-adaptation

Fig. 1 Inception layer model utilized in this network

Table 2 Representative experiments

Experiment name	Best MAE (years)
Residual with late inception	0.637
Residual with early inception	0.665
Base residual network	0.712
BoneNet	0.730
Residual with late inception—not age specific	0.742
Residual with late inception—no linear transform	0.775
Randomly initialized VGG	0.796
Pre-trained VGG	0.830

Multiple deep-learning architectures were evaluated, a representative summary of which are shown here. These include several variations of the proposed customized network: a base residual network (without inception layers), a residual network with early inception layer introduction (layer 3), a residual network with late inception layer introduction (layer 8). In addition, a late residual inception network was tested without the use of the spatial transformer. Finally, several previously published architectures were evaluated: a VGG network pre-trained on Imagenet, a randomly initialized VGG network, a customized six-layer network described by Spampinato et al. [6] (BoneNet). Each network architecture tested was evaluated with and without utilizing age-specific networks

[26]. Batch normalization was utilized to improve network training speed and regularization performance by reducing internal covariate shift [27].

Software code for this study was written utilizing the Python TensorFlow v1.1 library. Experiments and network training was performed on an Ubuntu 16.04 workstation with an NVIDIA TITAN X Pascal GPU. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

Mean squared error was utilized as the primary objective function of the network in order to penalize large deviations from the ground truth. Mean absolute error was employed as the primary performance metric [28–30].

Results

MAE accuracies on the validation and test sets for young females were 0.654 and 0.561 respectively. For older females,

validation and test accuracies were 0.662 and 0.497 respectively. For young males, validation and test accuracies were 0.649 and 0.585 respectively. Finally, for older males, validation and test set accuracies were 0.581 and 0.501 respectively. The female cohorts were trained for 900 epochs each and the male cohorts were trained for 600 epochs before convergence. An eightfold cross-validation set was employed for hyperparameter tuning. Test error was obtained after training on a full data set with the selected hyperparameters.

Discussion

Using our proposed customized neural network architecture on our large available data, we achieved an aggregate validation and test set mean absolute errors of 0.637 and 0.536 respectively. To date, this is the best published performance on utilizing deep learning for bone age assessment.

Our results support the findings by others demonstrating the applicability of deep-learning techniques for automation of bone age assessment [6, 8]. Importantly, our results reflect the analysis from Spampinato et al. [6] in that customized, purpose-built neural networks provide improved performance over networks derived from pre-trained imaging data sets. We build on that initial work by showing that the addition of state-of-the-art techniques such as residual connections and inception architecture further improves prediction accuracy. This is important because the current assumption for use of residual and/or inception architectures is that a large pre-trained network is required for successful implementation given the relatively small datasets in medical imaging. Instead we show that a small, customized architecture incorporating advanced CNN strategies can indeed be trained from scratch, yielding significant improvements in algorithm accuracy.

Compared to Lee et al., we opted for end-to-end learning with minimal preprocessing of bone age radiographs instead of initial segmentation. The success of such an approach highlights the flexibility and power of neural networks, which often can learn complex high-order imaging patterns without explicit feature generation by humans. Furthermore, minimizing image preprocessing not only increases algorithm

Fig. 2 Residual layer architecture utilized in this network

prediction speed but also facilitates generalizability by limiting the number of requisite manual steps. Finally, as a marginal benefit, retaining random background noise present in radiographic films adds stochasticity and helps to stabilize the algorithm training process.

Like Spampinato et al., we chose to design the CNN with a regression loss instead of a classification loss as proposed by Lee et al. This is important because bone age is a continuous variable and preserving a continuous output allows the algorithm to retain a hierarchical model of possible outputs. In other words, given a ground truth of 5 years, a prediction of 4 years 6 months will be penalized less than a prediction of 10 years, whereas both such estimates are equally incorrect in a classification scheme.

Furthermore, our results support the findings by Sun et al. [31] and the consensus in the computer vision community in that the performance of neural networks on various tasks increases logarithmically with increases in training data volume. This is important as training data is often a significant limiting factor in developing CNN algorithms for medical imaging applications. While tedious to obtain and curate properly, the relatively large dataset in our study was almost certainly contributed significantly to our improved results.

It should be noted that for all four cohorts, testing error outperformed validation error. One reason for this is that our ground truth for our test set was obtained by averaging two pediatric radiologist reads compared to our training data for which only a single read was used. This suggests that despite relatively noisy training data, the algorithm could successfully model the variation between observers and generate estimates that are close to the expected ground truth.

Despite promising results, one potential limitation is generalizability beyond our single institution given that the clear majority of training data originated from a single academic center. To some extent, this is countered by the relatively long 7-year extent of radiographs included in this study obtained from a heterogeneous combination of different machines and acquisition parameters, as well as the addition of a small cohort of outside examinations. Another potential limitation is the exclusion of significantly abnormal findings unrelated to bone age assessment in our test set such as bone fractures, congenital malformations, or systemic diseases with osseous manifestations. Finally, as with most other CNNs for medical imaging at this time, the algorithm is trained to evaluate a single diagnosis, and as such cannot provide information regarding ancillary findings. This highlights the continued need for review of the input radiographs by a trained radiologist to rule out alternative pathology unrelated to bone age that may be present.

Ongoing work in progress includes applying this network architecture to different types of normal images of left hands including studies done for trauma, skeletal

surveys, and evaluation of non-osseous disease with the goal of creating a new atlas for skeletal maturity. Additionally, methods to integrate this network into the clinical workflow are being explored.

References

1. Breen MA et al.: Bone age assessment practices in infants and older children among Society for Pediatric Radiology members. *Pediatr Radiol* 46(9):1269–1274, 2016
2. Bull RK et al.: Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. *Arch Dis Child* 81(2):172–173, 1999
3. Thodberg HH, Sävdahl L: Validation and reference values of automated bone age determination for four ethnicities. *Acad Radiol* 17(11):1425–1432, 2010
4. Ontell FK et al.: Bone age in children of diverse ethnicity. *AJR. Am J Roentgenol* 167(6):1395–1398, 1996
5. Berst MJ et al.: Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. *Am J Roentgenol* 176(2):507–510, 2001
6. Spampinato C et al.: Deep learning for automated skeletal bone age assessment in X-ray images. *Medical image analysis* 36:41–51, 2017
7. Thodberg HH et al.: The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging* 28(1):52–66, 2009
8. Lee H, et al: Fully Automated Deep Learning System for Bone Age Assessment. *J Digit Imaging* (2017): 1–15
9. Shen W, Zhou M, Yang F, Yang C, Tian J: Multi-scale Convolutional Neural Networks for Lung Nodule Classification. *Inf Process Med Imaging* 24:588–599, 2015
10. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S: Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans Med Imaging* [Internet]. 2016 May [cited 2017 Jul 2];35(5):1207–16. Available from: <http://ieeexplore.ieee.org/document/7422082/>
11. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al: Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal* [Internet]. Elsevier; 2017 Jan 1 [cited 2017 Sep 18]; 35:303–12. Available from: <http://www.sciencedirect.com/science/article/pii/S1361841516301244>
12. Dou Q, Chen H, Yu L, Zhao L, Qin J, Wang D, et al: Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. *IEEE Trans Med Imaging* [Internet]. 2016 May [cited 2017 Jul 2];35(5):1182–95. Available from: <http://ieeexplore.ieee.org/document/7403984/>
13. Pereira S, Pinto A, Alves V, Silva CA: Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans Med Imaging* [Internet]. 2016 May [cited 2017 Jul 2];35(5):1240–51. Available from: <http://ieeexplore.ieee.org/document/7426413/>
14. Chang PD: Fully Convolutional Deep Residual Neural Networks for Brain Tumor Segmentation. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H, editors. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised* [Internet]. Cham: Springer International Publishing; 2016. p. 108–18. Available from: https://doi.org/10.1007/978-3-319-55524-9_11

15. Wang J, Fang Z, Lang N, Yuan H, Su MY, Baldi P: A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. *Comput Biol Med* 84:137–146, 2017
16. Cao F et al.: Digital hand atlas and web-based bone age assessment: system design and implementation. *Comput Med Imaging Graph* 24(5):297–307, 2000
17. LeCun Y et al.: Gradient-based learning applied to document recognition. *Proceed IEEE* 86(11):2278–2324, 1998
18. He K, et al: Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016
19. Szegedy C, et al: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI*. 2017
20. Szegedy C, et al: Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015
21. Jaderberg M, Simonyan K, Zisserman A.: Spatial transformer networks. *Adv Neural Inf Process Syst* 2015
22. Kingma D, Ba J: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014
23. Nesterov Y: Gradient methods for minimizing composite objective function. 2007
24. Dozat T: Incorporating nesterov momentum into Adam. 2016
25. Glorot X, Bengio Y: Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010
26. Srivastava N et al.: Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15(1):1929–1958, 2014
27. Ioffe S, Szegedy C: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*. 2015
28. Maclaurin D, Duvenaud D, Adams R: Gradient-based hyperparameter optimization through reversible learning. *International Conference on Machine Learning*. 2015
29. LeCun YA, et al: Efficient backprop. *Neural networks: Tricks of the trade*. Springer Berlin Heidelberg, 2012. 9–48
30. Bengio Y: Practical recommendations for gradient-based training of deep architectures. *Neural networks: Tricks of the trade*. Berlin Heidelberg: Springer, 2012, pp. 437–478
31. Sun C, et al: Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968* 2017