

AutoMO-Mixer: An automated multi-objective Mixer model for balanced, safe and robust prediction in medicine

Xi Chen¹, Jiahuan Lv¹, Dehua Feng¹, Xuanqin Mou¹, Ling Bai¹, Shu Zhang¹,
and Zhiguo Zhou² *

¹ School of Information and Communication Engineering, Xi'an Jiaotong University,
Xi'an, China xi_chen@mail.xjtu.edu.cn

² Department of Biostatistics and Data Science, University of Kansas Medical
Center, Kansas City, KS
zgzhou2013@gmail.com

Abstract. Accurately identifying patient's status through medical images plays an important role in diagnosis and treatment. Artificial intelligence (AI), especially the deep learning, has achieved great success in many fields. However, more reliable AI model is needed in image guided diagnosis and therapy. To achieve this goal, developing a balanced, safe and robust model with a unified framework is desirable. In this study, a new unified model termed as automated multi-objective Mixer (AutoMO-Mixer) model was developed, which utilized a recent developed multiple layer perceptron Mixer (MLP-Mixer) as base. To build a balanced model, sensitivity and specificity were considered as the objective functions simultaneously in training stage. Meanwhile, a new evidential reasoning based on entropy was developed to achieve a safe and robust model in testing stage. The experiment on an optical coherence tomography dataset demonstrated that AutoMO-Mixer can obtain safer, more balanced, and robust results compared with MLP-Mixer and other available models.

Keywords: Image guided diagnosis and therapy · reliable artificial intelligence · balance · safe · robustness.

1 Introduction

With the development of modern medicine, medical image has become an essential tool to carry out personalized and accurate diagnosis. Due to the strong ability to analyze image, deep learning has been widely used in medical image analysis and has achieved great success [1,2] in the past years. However, many current available models can also lead to unreliable predictions. For example, the car's perception system misclassified the white part of the trailer into the sky, resulting in a fatal accident [3]. As such, different from other application

* corresponding author

fields such as face recognition, nature image classification, model reliability is more important in medicine as it is related to human life and health. On the other hand, we not only need to obtain the accurate prediction results, but also need to know whether the outcome is reliable or not. To realize this abstract goal by considering the clinical needs, we believe that building a unified model to achieve balance, safe and robust is desirable.

Currently, most prediction models use a single objective (e.g., accuracy, AUC) [4,5] function in the model training. However, the imbalanced sensitivity and specificity may result in higher rate of missed diagnosis [6]. Therefore, a multi-objective model which considers sensitivity and specificity simultaneously is needed. So far, there have been some studies on multi-objective optimization [7,8].

Furthermore, since most models are data-driven based strategy, it is hard to evaluate whether the prediction outcome for an unseen sample is reliable or not. A possible solution is evaluating the model output by introducing a “third party” to independently estimate the model reliability or uncertainty. There have been several studies on uncertainty estimation for deep learning. [9] proposed a framework based on test-time data augmentation to quantify the diagnostic uncertainty in deep neural networks. [10] used the prediction of the augmented images to obtain entropy to estimate uncertainty.

Meanwhile, it is found that the model built based on the dataset collected from one institution always obtain bad performance when the testing dataset is from another institution [11,12,13], demonstrating the poor robustness. On the other hand, a reliable model should always work well across the multiple institutions. Several studies have investigated this issue. Adversarial attack is one of the most serious factors that cause models not to be robust [14]. Some attackers perturbed test reports to obtain medical compensation [15]. Adversarial examples lead to wrong decisions that can cause dangerous effects on the patient’s life [16]. [17,18] evaluated the robustness of the model with adversarial attacks.

In summary, there have been several studies on building balanced, safe and robust model independently, but there is no unified framework that can achieve three goals simultaneously. As such, a new automated multi-objective Mixer (AutoMO-Mixer) model based on multiple layer perceptron Mixer (MLP-Mixer) is developed in this study to build a more reliable model. In AutoMO-Mixer, both sensitivity and specificity were considered as the objective functions simultaneously and a Pareto-optimal model set can be obtained through the multi-objective optimization [20] in training stage. In testing stage, the Pareto-optimal models with balanced sensitivity and specificity were chosen so as to improve model balance. To obtain safer and more robust model, evidential reasoning based on entropy (ERE) approach was developed to fuse the outputs of Pareto-optimal models to obtain the final outcome. The experimental studies on optical coherence tomography (OCT) dataset demonstrated that AutoMO-Mixer can outperform MLP-Mixer and other deep learning models, and more balanced, robust and safer results can be achieved as well.

2 Method

2.1 Overview

The framework of AutoMO-Mixer is shown in Fig. 1, which consists of training and testing stages. To build a balanced model, both sensitivity and specificity are considered as objective functions simultaneously in training stage, and a Pareto-optimal model set is generated then. To build a safer and more robust model, ERE strategy is developed to fuse the probability outputs of multiple Pareto-optimal models in testing stage.

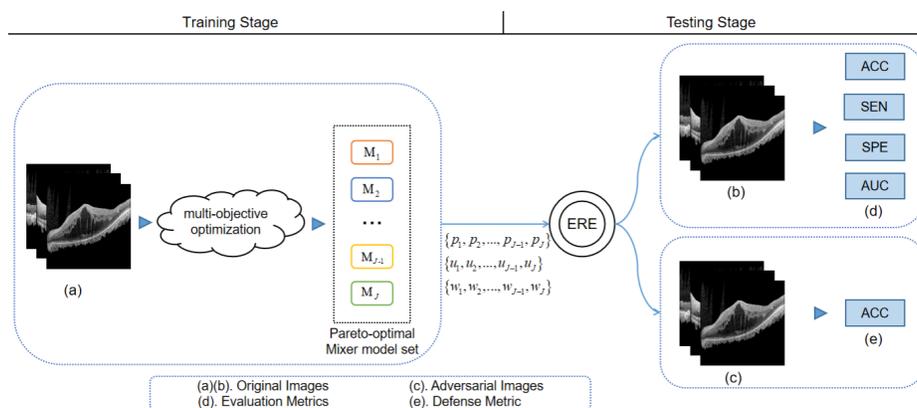


Fig. 1. The framework of AutoMO-Mixer model.

2.2 MLP-Mixer

Since the computational complexity is increased sharply when there are more parameters in multi-objective learning, it is better to have fewer parameters in model training. The recently proposed MLP-Mixer [19] model is a full MLP architecture. Compared with CNN, the convolutional layer is removed from MLP-Mixer, leading to decreasing the scale of the architecture parameters sharply. On the other hand, MLP-Mixer can achieve similar performance to CNN [19]. Therefore, it is a better choice in multi-objective learning.

2.3 Training stage

In training stage, sensitivity denoted by f_{spe} and specificity denoted by f_{sen} are considered as objective functions simultaneously, they are:

$$f_{sen} = \frac{TP}{TP + FN} \quad (1)$$

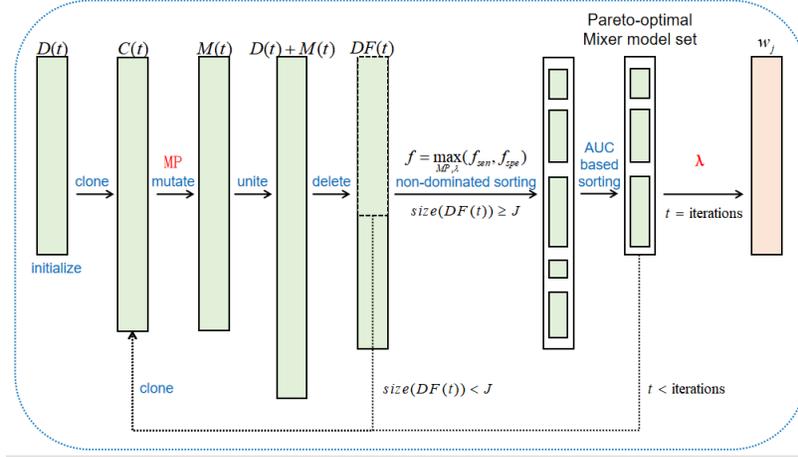


Fig. 2. The illustration of training stage.

$$f_{spe} = \frac{TN}{TN + FP} \quad (2)$$

where TP and TN represent the number of true positives and true negatives, FP and FN are the number of false positives and false negatives, respectively.

Assume $M = \{m_1, \dots, m_q\}$ denotes the MLP-Mixer model, where q represents the number of model parameters. To obtain the balanced models, we aim to maximize f_{sen}, f_{spe} simultaneously, and an iterative multi-objective immune algorithm (IMIA) [20] is used. IMIA consists of six steps: initialization, cloning, mutation, deletion, update, and termination. First, the initial model set denoted by $D(t) = \{M_1, \dots, M_N\}$ is generated, where $M_i = \{m_{i1}, \dots, m_{iq}\}, i = 1, 2, \dots, N$. Then the models with higher f_{sen}, f_{spe} will be replicated using the proportional cloning method. In the third step, a probability of mutation is randomly generated for each model, and the model performs mutation when its probability is larger than the mutation probability (MP). After the mutation, the new models are generated. If some models have same sensitivity and specificity, only one model is remained. Then the model set size is kept through AUC based non-dominated sorting strategy. The training process will not stop until the maximum number of iterations is reached. Finally, the Pareto-optimal Mixer model set is generated, where the model set size is J . Since the two hyperparameters MP and λ may affect the model performance, Bayesian optimization [21] is used to optimize the hyperparameters. The illustration of the training phase is shown in Fig. 2.

2.4 Testing stage

In testing stage, the probability outputs of Pareto-optimal models are fused through the evidential reasoning [22,23] based on entropy approach. The workflow is shown in Fig. 3.

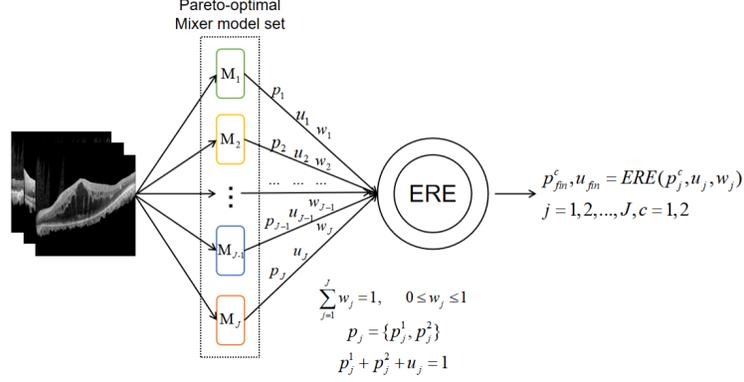


Fig. 3. The illustration of testing stage.

Weight calculation Since the performance of different Pareto-optimal models cannot be the same, the weight for each model should be estimated, which is denoted by w_j . As the balanced model between sensitivity and specificity is desired, the ratio between them is considered in the weight calculation, that is $\frac{f_{sen}^j}{f_{spe}^j}$ or $\frac{f_{spe}^j}{f_{sen}^j}$. When the ratio is less than 0.5 or greater than 1, the model is considered as extreme imbalance, setting w_j as 0. Meanwhile, AUC is a good measure for model reliability, it is also considered. The expression of w_j is as follows:

$$w_j = \begin{cases} \lambda \frac{f_{sen}^j}{f_{spe}^j} + (1 - \lambda)AUC_j, & \text{when } 0.5 \leq \frac{f_{sen}^j}{f_{spe}^j} \leq 1 \\ \lambda \frac{f_{spe}^j}{f_{sen}^j} + (1 - \lambda)AUC_j, & \text{when } 0.5 \leq \frac{f_{spe}^j}{f_{sen}^j} \leq 1, j = 1, 2, \dots, J \\ 0 & \text{Other situations} \end{cases} \quad (3)$$

where λ indicates the importance of balance, and $1 - \lambda$ indicates the importance of AUC. After calculating the w_j for each model, the weights are normalized.

Uncertainty estimation Test-time data augmentation (TTA) [9,10] is used to perform useful estimates of model uncertainty. The test image is fed into model $M_j, j = 1, 2, \dots, J$ to generate the probability output $p_j^c, c = 1, 2$, where $p_j^1 + p_j^2 = 1$. The original test image is enhanced T times to generate prediction $p_{j,t}^c, t = 1, 2, \dots, T$. The mean class probability \bar{p}_j^c and the uncertainty u_j are:

$$\bar{p}_j^c = \frac{1}{T} \sum_{t=1}^T p_{j,t}^c, c = 1, 2 \quad (4)$$

$$u_j = - \sum_{c=1}^2 \bar{p}_j^c \log(\bar{p}_j^c) \quad (5)$$

To satisfy the conditions of the ERE strategy, p_j^c and u_j are normalized so that $p_j^1 + p_j^2 + u_j = 1$.

ERE strategy Assume that the output probability for each model is denoted by $p_j = \{p_j^1, p_j^2\}$, $p_j^1 + p_j^2 \leq 1$, $j = 1, 2, \dots, J$. If $p_j^1 + p_j^2 < 1$, it shows that the j th model has uncertainty u_j on its output. Then the final output probability p_{fin}^c , $c = 1, 2$ and uncertainty u_{fin} are obtained through the ERE fusion strategy. that is:

$$p_{fin}^c, u_{fin} = ERE(p_j^c, u_j, w_j), j = 1, 2, \dots, J, c = 1, 2 \quad (6)$$

where ERE is:

$$p_{fin}^c = \frac{\mu \times [\prod_{j=1}^J (w_j p_j^c + 1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^J (1 - w_j(p_j^1 + p_j^2))]}{1 - \mu \times [\prod_{j=1}^J (1 - w_j)]}, c = 1, 2 \quad (7)$$

$$u_{fin} = \frac{\mu \times [\prod_{j=1}^J (1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^J (1 - w_j)]}{1 - \mu \times [\prod_{j=1}^J (1 - w_j)]} \quad (8)$$

The normalized factor μ is:

$$\mu = [\sum_{c=1}^2 \prod_{j=1}^J (w_j p_j^c + 1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^J (1 - w_j(p_j^1 + p_j^2))]^{-1} \quad (9)$$

2.5 Robustness evaluation

In this study, fast gradient sign method (FGSM) [24] is used to disturb the original samples, which is a white box attack with full information of the models. Adversarial samples are generated by the following formula:

$$x^a = x + \delta \quad (10)$$

where x^a represents the adversarial sample, x represents the original sample. δ represents the perturbation. The degree of perturbation is controlled by ε . In our study, ACC is used to evaluate robustness [18].

3 Experiments

3.1 Experimental setup

The dataset used in this study was collected from the Second Affiliated Hospital of Xi'an Jiaotong University (Xi'an, China), including 228 patients with Choroidal neovascularization (CNV) and cystoid macular edema (CME) between October 2017 and October 2019. First, OCT images of each patient were acquired via the Heidelberg Retina Tomograph-IV (Heidelberg Engineering, Heidelberg, Germany). These patients were then injected with anti-vascular endothelial

Table 1. The range of values for MLP-Mixer network structure parameters.

parameters	range of values
Number of layers	[2, 3, 4]
Hidden size C	256*[1, 1.2, 1.4, 1.6]
MLP dimension Ds	196*[2, 3, 4, 5]
MLP dimension Dc	256*[2, 4, 6, 8, 10]

Table 2. The evaluation results on OCT dataset.

models	SEN	SPE	AUC	ACC	$\frac{\min(\text{SEN}, \text{SPE})}{\max(\text{SEN}, \text{SPE})}$
MLP-Mixer	0.611±0.052	0.703±0.077	0.709±0.041	0.671±0.038	0.869
ResNet-18	0.728±0.075	0.706±0.071	0.791±0.046	0.714±0.052	0.970
AutoMO-Mixer	0.778±0.000	0.779±0.000	0.844±0.000	0.779±0.000	0.999

growth factor (anti-VEGF) and the evaluations were made after 21 days. Among them, anti-VEGF was effective for 171 patients, and the remaining 57 patients had no sign of effectiveness. The study was approved by the Research Ethics Committee, and each patient provided written informed consent. In our study, we built a binary classifier to determine whether anti-VEGF would be effective for patients using OCT images. In the training stage, there were 135 effective cases and 44 ineffective cases. In the testing stage, there were 34 and 12, respectively, in these two classes.

Before being fed into the model, all the images were resized into 224 x 224. MP and λ were set to 0.5 and 0.8, respectively. The MLP-Mixer contains four parameters, these settings are shown in Table 1.

As AutoMO-Mixer was built based on MLP-Mixer and ResNet-18 is a classical deep learning model, they were used in comparative study. The four parameters in MLP-Mixer network were set to 5, 256, 392, 1024, respectively, and transfer learning was used on ResNet-18 as pre-training. Sensitivity (SEN), specificity (SPE), area Under Curve (AUC), and accuracy (ACC) were used for evaluation. All the experiments were performed five times, and mean and standard deviation were evaluated.

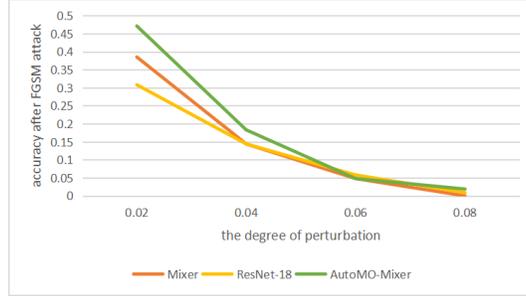
3.2 Results

The evaluation results on MLP-Mixer, ResNet-18 and AutoMO-Mixer are shown in Table 2. In this study, $\frac{\min(\text{SEN}, \text{SPE})}{\max(\text{SEN}, \text{SPE})}$ was used to assess the balance of the model. It can be seen that AutoMO-Mixer model is the most balanced. In addition, both the AUC and ACC of the AutoMO-Mixer are better than the other two models.

Safety evaluation In this study, the uncertainty estimation was used to measure model safety. If the performance of the model can improve as the number of test samples with high uncertainty decreases, it is indicated that the model

Table 3. Model performance of the test cohorts stratified by the uncertainty.

uncertainty	SEN	SPE	AUC	ACC
0.4245	0.778	0.779	0.844	0.779
0.4206	0.783	0.796	0.860	0.792
0.4165	0.818	0.829	0.823	0.827
0.4045	1.000	0.895	1.000	0.920

**Fig. 4.** Comparison of the robustness between the AutoMO-Mixer, ResNet-18 and AutoMO-Mixer models.

is safe. The entire test samples were arranged from smallest to largest in order of uncertainty, with the maximum uncertainty being 0.4245, the upper quartile being 0.4206, the median being 0.4165, and the lower quartile being 0.4045. Samples with less uncertainty than them were grouped into four cohorts, and the evaluation results are shown in Table 3. It can be seen that the lower the cutoff uncertainty is, the better the model’s performance is, indicating our model can assess whether the prediction is safe based on uncertainty.

Robustness After the original samples were attacked by FGSM, indistinguishable adversarial samples were generated. We measured the accuracy of adversarial samples in each model in Fig. 4. It is obvious that except slightly less when $\varepsilon=0.06$, the robustness of AutoMO-Mixer is better than the other as a whole.

4 Conclusions

In this study, a new model termed as AutoMO-Mixer was developed for image guided diagnosis and therapy. In AutoMO-Mixer, sensitivity and specificity were considered as the objective functions simultaneously and a Pareto-optimal Mixer model set can be obtained in training stage. In testing stage, ERE was used to obtain safer and more robust results. The experimental results on OCT dataset showed that AutoMO-Mixer outperformed MLP-Mixer and ResNet-18 in balance, safe and robustness.

References

1. Zhang, Y., An, M.: Deep learning-and transfer learning-based super resolution reconstruction from single medical image. *Journal of healthcare engineering* **2017** (2017)
2. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Annual review of biomedical engineering* **19**, 221–248 (2017)
3. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* **30** (2017)
4. Huynh, E., Coroller, T.P., Narayan, V., Agrawal, V., Hou, Y., Romano, J., Franco, I., Mak, R.H., Aerts, H.J.: Ct-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiotherapy and Oncology* **120**(2), 258–266 (2016)
5. Vallières, M., Freeman, C.R., Skamene, S.R., El Naqa, I.: A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine & Biology* **60**(14), 5471 (2015)
6. Błaszczyszński, J., Deckert, M., Stefanowski, J., Wilk, S.: Integrating selective pre-processing of imbalanced data with ivotes ensemble. In: *International conference on rough sets and current trends in computing*. pp. 148–157. Springer (2010)
7. Chen, H., Deng, T., Du, T., Chen, B., Skibniewski, M.J., Zhang, L.: An rf and lssvm–nsga-ii method for the multi-objective optimization of high-performance concrete durability. *Cement and Concrete Composites* p. 104446 (2022)
8. Bagheri-Esfeh, H., Dehghan, M.R.: Multi-objective optimization of setpoint temperature of thermostats in residential buildings. *Energy and Buildings* p. 111955 (2022)
9. Ayhan, M.S., Berens, P.: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks (2018)
10. Dohopolski, M., Chen, L., Sher, D., Wang, J.: Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty. *Physics in Medicine & Biology* **65**(22), 225002 (2020)
11. Uwimana, A., Senanayake, R.: Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. *arXiv preprint arXiv:2107.04882* (2021)
12. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017)
13. Ge, Z., Wang, X.: Evaluation of various open-set medical imaging tasks with deep neural networks. *arXiv preprint arXiv:2110.10888* (2021)
14. Apostolidis, K.D., Papakostas, G.A.: A survey on adversarial deep learning robustness in medical image analysis. *Electronics* **10**(17), 2132 (2021)
15. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 493–501. Springer (2018)
16. Mangaokar, N., Pu, J., Bhattacharya, P., Reddy, C.K., Viswanath, B.: Jekyll: Attacking medical image diagnostics using deep generative models. In: *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. pp. 139–157. IEEE (2020)

17. Xu, M., Zhang, T., Li, Z., Liu, M., Zhang, D.: Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Medical Image Analysis* **69**, 101977 (2021)
18. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6023–6032 (2019)
19. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems* **34** (2021)
20. Zhou, Z., Folkert, M., Iyengar, P., Westover, K., Zhang, Y., Choy, H., Timmerman, R., Jiang, S., Wang, J.: Multi-objective radiomics model for predicting distant failure in lung sbirt. *Physics in Medicine & Biology* **62**(11), 4460 (2017)
21. Pelikan, M.: Bayesian optimization algorithm. In: *Hierarchical Bayesian optimization algorithm*, pp. 31–48. Springer (2005)
22. Yang, J.B., Xu, D.L.: On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **32**(3), 289–304 (2002)
23. Wang, Y.M., Yang, J.B., Xu, D.L.: Environmental impact assessment using the evidential reasoning approach. *European Journal of Operational Research* **174**(3), 1885–1913 (2006)
24. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014)