2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761420

TRIAGENET: A MULTI-AGENT DIAGNOSIS NETWORK FOR IMBALANCED DATA

Weixiang Chen, Jianjiang Feng, Jie Zhou*

Department of Automation, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China

ABSTRACT

Imbalanced and even long-tail distribution of different categories is a challenge for multi-class classification problem, especially for medical image diagnose whose data distribution is usually imbalanced. Toward this issue, we proposed an end-to-end multi-agent classification network called TriageNet, which is combined of multiple selectors and diagnostic agents. All categories are guided to different agents by selectors, and every agent is an expert in a specific group of categories. This process, which is similar to triage in hospitals, helps decrease the unbalance between categories for both selectors and agents. Experiments on an extremely imbalanced pneumonia CT dataset and a publicly available X-ray dataset Chexpert show that TriageNet is relatively robust to imbalanced data.

Index Terms— disease diagnosis, long-tail distribution, multiple experts, imbalanced data

1. INTRODUCTION

Computer assisted or aided diagnosis (CAD) is widely studied in recent years, and with the help of deep learning technology, many diagnostic algorithms [1, 2, 3, 4] have been proposed to deal with various diagnostic issues. However, most of disease of one organ have many types, but most existing diagnosis algorithms can only work on small groups of them, usually 2 to 5 classes. It is one of the challenges that CAD still faces, which limited the practical usages of these algorithms. The main causes of this situation are that collecting sufficient numbers of samples for multi-category deceases is difficult, and even when the data is collected, imbalanced distribution of them troubles the network training.

We suppose that creating methods with more universality is a trend and need for CAD, and some of publications have already worked on it. For example, Andre et al. [5] proposed a classification algorithm on skin cancer in which 2,032 different diseases were included. However, this algorithm can only perform two critical binary classifications rather than all diseases; Chexpert [6] is a multiple observation diagnosis dataset, on which many algorithms were developed [7, 8]. Since Chexpert is a multi-label dataset rather than a multi-category dataset, most algorithms approached multi-label classification problem as multiple binary classification tasks.

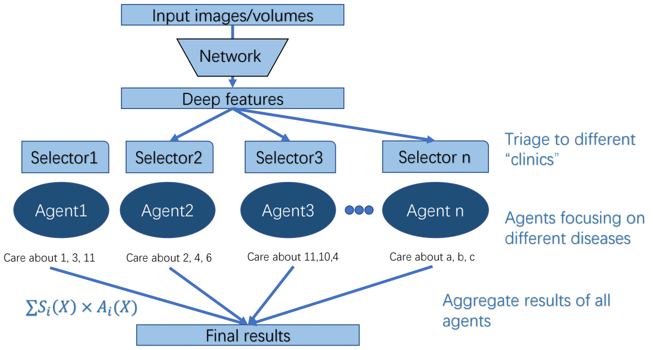


Fig. 1. Network diagram of TriageNet. TriageNet is combined by selectors and agents. Selectors work as triage worker to send cases to different agents. Agents work as clinics which are good at different categories of disease.

From the view point of network training, when the numbers of different categories are imbalanced, the training might be dominated by the categories with more samples. To deal with this issue, some former works used sampling [9, 10], and some others used weighted loss for different categories [11]. However, these methods cannot work well when the imbalanced is extreme, such as the so-called long-tail distribution. What's more, these designs cannot help in learning fine-level features between similar categories. Groupsoftmax [12] is a classification structure targeting the long-tail issue and it works well in natural image databases. It uses some abstract groups to cover the long-tail categories to balance training samples within and between groups. This method is quite close to our target but it still cannot help network to learn specific fine-level differences between confusing categories because the grouping is done based only on numbers. Multi-gate Mixture-of-Experts (MMoE) [13] is another network design towards in-group fine-level feature extraction. It uses multiple linear layers to represent different experts and these experts extract different features for fine-level differences within groups. However, the groups are not really defined in MMoE, instead it just processes features using different linear layers

*Corresponding author

and aggregates the results.

Inspired by the triage system in real hospitals, we proposed our TriageNet which simulates a triage process for every sample. All categories are allocated by different diagnostic agents, which simulate clinics, so that different agents can only diagnose their allocated disease. A triage selector is also set for every agent, which helps to predict whether the cases are within the group or not. For every agent, the classifier is trained using only the cases within the group, resulting in an expert for the group, which inherits the essence of MMoE. The groups are defined randomly or based on prior knowledge of confusing diseases, which is different from Groupsoftmax. Besides, since agents learn fine-level differences between groups, we set groups with overlapping categories in order to find different fine-level features. The contributions of this work can be summarized as:

- proposed a model targeting imbalanced of data in CAD named TriageNet, whose architecture simulates the real hospital's triage.
- evaluate TriageNet on an extremely imbalanced pneumonia CT dataset and a publicly available X-ray dataset Chexpert. According to results, TriageNet is robust in data unbalance and long-tail distribution thanks to network triage of cases and aggregation of multiple agents.

2. METHOD

2.1. TriageNet Designs

In hospitals, most doctors are specialists in some specific fields so that almost all hospitals divided their doctors to different departments. Patients are assigned to proper departments according to general practitioners, experienced nurses or themselves, which is called triage. Our TriageNet simulates this system for better diagnostic performances, which works based on two assumptions:

- classifiers trained on specific part of data work better than those trained on the whole data, when the test data is only within that part.
- identification of groups of categories is easier than identification of all separate ones, whatever the principles of grouping.

Since multi-expert method is already successfully used in MMoE, the first assumption is reasonable. On the other hand, Groupsoftmax proved the second one should be feasible too.

Our TriageNet (Fig. 1) contains multiple different diagnostic agents which allocate different skilled categories, and multiple triage selectors which correspond to agents one by one and predict the possibility of cases sending to that agent. The training of TriageNet is end-to-end that both agents and selectors are optimized together. During inference, the final results can be computed by:

$$\begin{aligned} P(y=d|X) &= \frac{\sum_{i|d\in G_i} P(y\in G_i|X) * P(y=d|X, t\in G_i)}{\sum_{i|d\in G_i} P(y\in G_i|X)} \\ &= \frac{\sum_{i|d\in G_i} S_i(X) * A_i^{(d)}(X)}{\sum_{i|d\in G_i} S_i(X)} \end{aligned} \tag{1}$$

where $S_i(X)$ is the output of i-th selector, $A_i^{(d)}(X)$ is the output of i-th agent on category d, and G_i is the allocated categories of i-th agent.

2.2. Diagnostic Agents

The structure of diagnostic agents is shown on Fig. 2. For each agent, different features from backbone are selected, different classification jobs are defined and the data of different jobs are fed in training process. We denote the number of features as n_f, the number of agents as n_a, and the number of categories each agent as n_c. We can either set different n_a, n_f, n_c for every agent according to some prior of sense, or just fix them.

The inner structure of agents is a layer of fully-connected layer using the selected elements of features. The loss function for each agent is cross-entropy after a Softmax process on agent's outputs, which is defined as:

$$L_{agent} = \sum_{i=0}^{n_a} \hat{A}_i(X)\log(A_i(X))\mathrm{I}(y\in G_i) \tag{2}$$

where the function $\mathrm{I}(y\in G_i)$ remains zero unless y is within the range of G_i, $\hat{A}_i(X)$ is the groundtruth of agents' output which can be generated by y. Since the allocated categories have different indexes, for every agent the groundtruth of agent should be generated respectively.

2.3. Triage Selectors

The inner structure of triage selectors is a fully-connected layer and the groudtruth for them can be generated according to whether y is within the range of G_i. Different from diagnostic agents, selectors take all features from backbone for more accurate triages (Fig. 2). Therefore, the loss function of triage selectors can also be defined as cross-entropy:

$$L_{selector} = \sum_{i=0}^{n_a} \hat{S}_i(X)\log(S_i(X)) \tag{3}$$

The whole TriageNet is trained end-to-end so that the loss function is :

$$L = \lambda_1 L_{agent} + \lambda_2 L_{selector} + \lambda_3 R \tag{4}$$

where R is regularization item, and λ_x are weights to balance loss items.

Random selecting processes is used in order to build various diagnostic agents (Fig. 2), which gives more possibility to overcome noises and meets the essence of multi-expert.

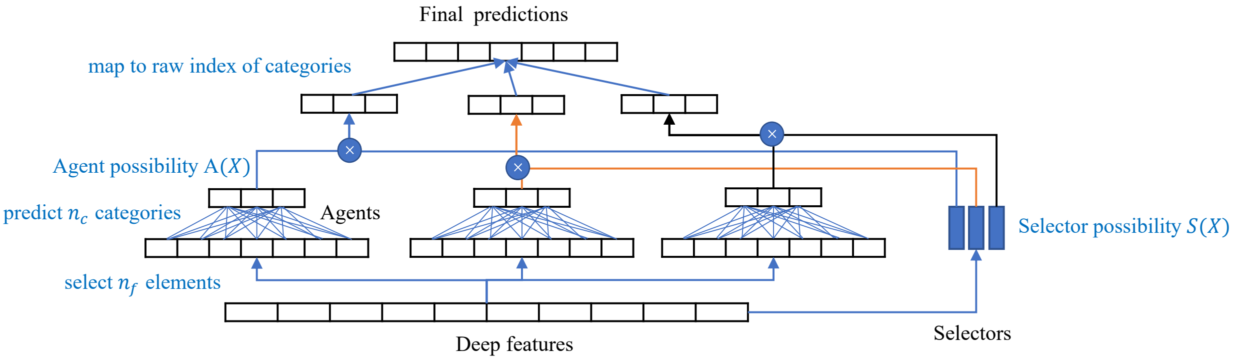


Fig. 2. Diagram for Agents, Selectors and the random sampling process. Deep features in the figure means features from backbone network, which are shared by all selectors and agents.

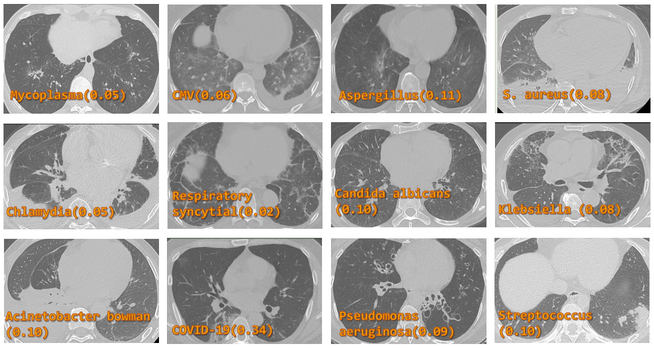


Fig. 3. Example of 12 categories in pneumonia dataset which are CTs shown in slices. The numbers after names are ratios of cases for the categories.

3. EXPERIMENTS

3.1. Dataset

We test our method on two datasets, and the first one is a pneumonia dataset which was collected from our cooperation hospitals. There are 2,353 CTs in the dataset which can be divided into 12 types of pneumonia based on their different pathogens (Fig. 3). According to the bio-taxonomy of the 12 pathogens, we defined 5 level-I categories (viral, fungal, bacterial, mycoplasma and chlamydial) in our dataset, while the raw detailed pathogens are used as level-II categories. Recognizing pathogens of pneumonia using images is believed a challenging task even for experienced radiologists [14, 15], and it suffers a severe unbalance in numbers of different categories.

The second one is a publicly available dataset called Chexpert [6] which contains 224,316 chest radiographs of 65,240 patients, and the labels are 14 common chest radiographic observations. Chexpert is unclearly labeled and the 14 observations are not mutual exclusive with each other. We test the U-Ignore results on Chexpert, because in this situation the unbalance of data is the most severe, and the number of cases is least.

3.2. Hyper-parameters and Training Details

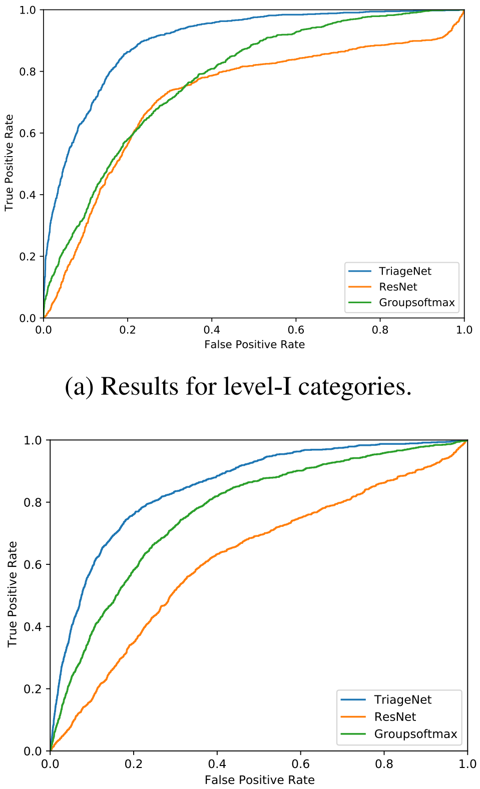
The hyper-parameters can influence results a lot, so we choose them based on experiments and memory of GPUs. For pneumonia dataset, we set $n_f = 1024$, $n_c = 3$ for its best performances in training set, and used $n_t = 50$ since it fills up one piece of GPU, which is NVIDIA RTX3090. Among all agents, half of them were used to predict level-I label, and the others were for level-II. While for Chexpert, we set $n_f = 512$, $n_c = 4$, and $n_t = 50$. Since Chexpert is in fact a multi-label task rather than a multi-category task, we removed the selectors when dealing with it. As a substitute, we regard the exponent of minus outputs' entropy as $S_i(X)$ for agents during inference (as (5)). All the training was performed for 20 epochs with 10^{-4} learning rate.

$$S_i(X) = \mathrm{e}^{\sum_{d \in G_i} A_i^{(d)}(X) \times \log A_i^{(d)}(X)} \tag{5}$$

The feature extraction backbones are not the main topic of this work, so that we just use fixed backbone in our experiments which is ResNet [16] for pneumonia dataset and DenseNet [17] for Chexpert. Because pneumonia dataset contains CTs, we use backbone network to extract features of all slices and aggregate all slices of the same volumes into one feature using maxpooling.

3.3. Results on Pneumonia Dataset

To compare our method against the state of the art, we set up some baselines. **ResNet** is one baseline results that used ResNet features and two layers of fully-connected to predict level-I or level-II labels. **Groupsoftmax** is another method to



(a) Results for level-I categories.

(b) Results for level-II categories.

Fig. 4. ROC curves on pneumonia dataset.

Table 1. Performances on pneumonia dataset

Methods	level-I		level-II	
	AUC(%)	Acc.(%)	AUC(%)	Acc.(%)
ResNet	71.88±1.07	54.70±1.21	61.90±0.78	24.48±1.00
Groupsoftmax	77.59±0.59	69.68±1.05	76.84±0.51	48.07±0.95
TriageNet	**82.93±0.39**	**71.39±1.10**	**85.08±0.43**	**52.48±1.05**

overcome long-tail distribution [12]. In our experiments, we grouped long-tailed level-II labels into four groups and non-long-tails into one. Our method used the hyper-parameters described in 3.2.

The results (Fig. 4 and Table 1) shows that our method outperformed all other methods and gave accuracies of 71.39±1.10% and 52.48±1.05% respectively for level-I and level-II. Since this task is quite challenging, the accuracies are acceptable and much better than the baselines. Without any improvement, the naive ResNet can only get 54.70±1.21% and 24.48±1.00% accuracies respectively.

We can tell from Fig. 5 that TriageNet worked significantly better on most categories and the averaged accuracy. Especially for some categories with lower numbers, such as CMV, Respiratory syncytial, TriageNet outperformed baseline significantly.

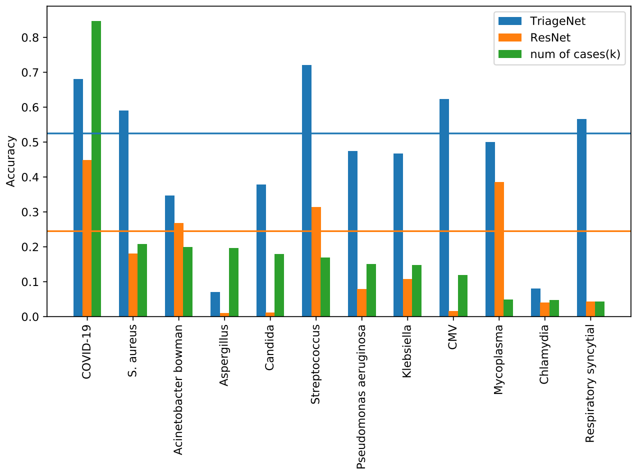


Fig. 5. Accuracies for categories of pneumonia dataset. The lines are averaged accuracies of two methods across all categories.

Table 2. Performances on Chexpert dataset

Methods	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
DenseNet	76.48	**82.14**	86.96	**92.81**	85.96
TriageNet	**79.62**	82.10	**87.70**	91.47	**87.48**

3.4. Results on Chexpert Dataset

For Chexpert, a result is reported in the original paper[6], but the results of 14 observations did not come from single model. Therefore, we reran the model of original paper in a multi-label training process as our baseline. Results (shown in Table 2) show that in the five labels that Chexpert reported, our TriageNet slightly outperformed DenseNet baseline. Only results for Edema are worse than that the baseline, which is the label with least positive samples. It shows that our TriageNet seems not that robust for multi-label tasks when the number of cases is extremely insufficient. However, the other labels still get better results which probably benefits from the multi-expert design.

4. CONCLUSIONS

The design of TriageNet simulates the triage in real hospitals which divides a complex diagnose task into easier sub-tasks. The diagnosis task for every agent is easier because of the less imbalanced data and fewer diagnostic choices when inference. Experiments on two datasets show TriageNet helps in dealing with imbalanced data which is very common in medical image diagnosis field.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study got ethical approval of ChiCTR2000038609 for using pneumonia dataset. The Chexpert dataset is publicly available.

6. ACKNOWLEDGMENTS

This study was supported by the National Natural Science Fuoundation of China (grant 82071921).

7. REFERENCES

[1] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[3] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al., "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[4] Eric J Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[5] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 590–597.

[7] Imane Allaouzi and Mohamed Ben ahmed, "A novel approach for multi-label chest x-ray classification of common thorax diseases," *IEEE Access*, vol. 7, pp. 64279–64288, 2019.

[8] Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, and Ha Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.

[9] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Class rectification hard mining for imbalanced deep learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1851–1860.

[10] Tuanfei Zhu, Yaping Lin, and Yonghe Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, 2017.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[12] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10991–11000.

[13] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1930–1939.

[14] Grant Mackenzie, "The definition and classification of pneumonia," *Pneumonia*, vol. 8, no. 1, pp. 1–5, 2016.

[15] Atsushi Nambu, Katsura Ozawa, Noriko Kobayashi, and Masao Tago, "Imaging of community-acquired pneumonia: roles of imaging examinations, imaging diagnosis of specific pathogens and discrimination from noninfectious diseases," *World Journal of Radiology*, vol. 6, no. 10, pp. 779, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Rattern Recognition*, 2016, pp. 770–778.

[17] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Rattern Recognition*, 2018, pp. 2752–2761.