Multi-task Learning for Hierarchically-Structured Images: Study on Echocardiogram View Classification

Jerome Charton[1], Hui Ren[1], Sekeun Kim[1], Carola Maraboto Gonzalez[1], Jay Khambhati[1], Justin Cheng[2], Jeena DeFrancesco[2], Anam Waheed[2], Sylwia Marciniak[2], Filipe Moura[2], Rhanderson Cardoso[2], Bruno Lima[2], Michael Picard[1], Xiang Li[1], and Quanzheng Li[1]

[1] Massachusetts General Hospital / Harvard Medical School
[2] Brigham and Women's Hospital

Abstract. Echocardiography is a crucial and widely adopted imaging modality for diagnosing and monitoring cardiovascular diseases. Deep learning has been proven effective in analyzing medical images but is limited in echocardiograms due to the complexity of image acquisition and interpretation. One crucial initial step to address this is automatically identifying the correct echocardiogram video views. Several studies have used deep learning and traditional image-processing techniques for this task. The authors propose an ablation study on a multi-task learning scheme with a hierarchically structured model output that arranges views in a tree structure. The proposed model, named "Multi-task Residual Neural Network (MTRNN) with masked loss", uses a conditional probabilistic training method and demonstrates superior performance for echocardiogram view classification. While the model has only been validated for the echocardiogram video classification task, it can be easily generalized to any medical image classification scenario with a hierarchical structure among the data labels.

Keywords: Echocardiagram · Multi-task learning · View classification.

1 Introduction

Echocardiography is a critical and widely adopted imaging modality for the screening, diagnosing, differential diagnosing, and follow-up of various cardiovascular diseases [13]. Deep learning has emerged as a powerful tool for analyzing medical images and has shown its potential to reduce the burden on cardiologists and radiologists [17]. However, applying deep learning methods for echocardiogram analysis is more challenging than other modalities due to the complexity of image interpretation and identification of the desired imaging view(s) and the focus in that view [8]. To address this issue, the first step toward comprehensive computer-assisted echocardiographic image analysis is to automatically identify the correct views for echocardiogram videos [16]. Recently, there have been multiple studies targeting this task, both using deep learning-based techniques [3,

9, 10, 12, 19] and traditional image processing-based techniques [1, 20] and many others.

On the other hand, from the perspectives of the sonographers, obtaining an echocardiogram generally involves several standard steps, such as localization, rotation, and tilting of the probe or transducer. These steps will naturally result in a hierarchical structure among echocardiogram views. We observe that these views could be arranged in a tree structure depending on the location (Apical, Parasternal, Others), the orientation of the probe (e.g., short axis views vs long axis views, orientation notch towards the right side or the up side of the body) and the focus in the view (e.g., short axis view at the level of the aortic valve vs apical long axis view with three chambers). Motivated by the imaging procedure and observations, we proposed a multi-task learning scheme with hierarchically-structured model outputs in this study. The proposed scheme simultaneously predicts the corresponding labels at each tree layer via different branches implemented as model heads. In addition, motivated by the multi-task learning schemes proposed by [4, 14], which utilizes model training with conditional probability and a masked loss function, we integrate the conditional probabilistic training into the branch-based model design. The final model, named "Multi-task Residual Neural Network (MTRNN) with masked loss", fully leverages the intrinsic tree structure of the relationship among video labels (views) and has demonstrated superior performance for the echocardiogram view classification task using an in-house dataset. Formulating the view classification task in a hierarchical multi-task learning framework can: 1) improve model generalizability by learning the related data labels simultaneously with shared representations learned across tasks, which can help capture standard features [15]; 2) reduce overfitting towards a single task by defining the loss function across multiple related tasks [22]; 3) improve model explainability as not only the leaf-level label (e.g., whether the given video belongs to A4C view) is predicted, but also along with the labels of each layer (e.g., whether the video belongs to apical view or parasternal views).

2 Methodology

2.1 Model Architecture

Hierarchically-structured image classification has been discussed through several research works but is still a largely overlooked topic. We can distinguish two approaches for leveraging the hierarchical structure among labels: revisiting the loss function or designing the hierarchical classification network. These two approaches were considered in [23] (also in [7]) and [4] (also in [2]). While [23] relay on a network architecture adapted to the hierarchical classification (BCNN: branch convolutional neural network for hierarchical classification) with a specific loss function designed for its network (Weighted Loss), [4] focuses on the definition of a loss function adapted to a hierarchical classification (Masked loss). This paper will compare these two loss functions upon several networks similar to the BCNN.

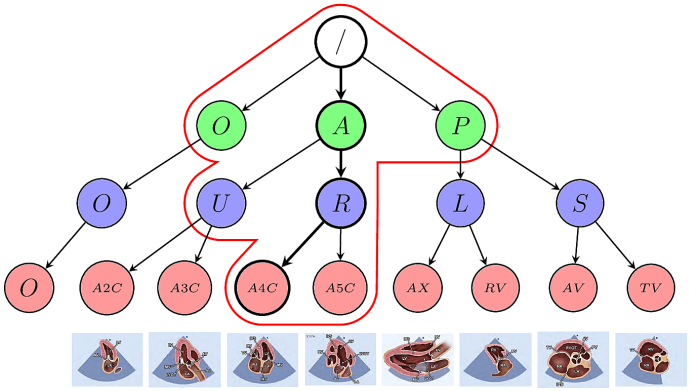

Fig. 1. Tree structure of the classification labels of this task. Echocardiography videos are classified into nine regular views, including Apical 2, 3, 4, 5 chambers (A2c, A3c, A4c, A5c, respectively), Parasternal long axis (AX), Parasternal long axis with right ventricle focus(RV), Parasternal short axis at the level of aortic valve (AV), Parasternal short axis at the level of tricuspid valve (TV), and Others (O), shown as leaf nodes on the hierarchical tree. Those views are associated at the first level by the orientation of the probe during image acquisition, including the Upper (U), Right (R), Long axis (L), Short axis (S), and Others (O). These five classes are then regrouped into three classes depending on the location of the probe Apical (A), Parasternal (P), and Others (O). Outlined nodes are A4c and its ancestors. The red contour draws the mask defined Eq. (2). Illustrations of the images for each view are visualized at the bottom, which were originally presented in [11].

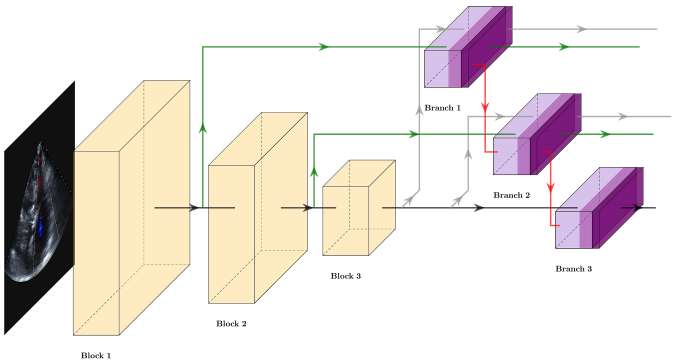

Fig. 2. Multi-task network architectures tested in this work on a VGG16 backbone. Black arrows: connections of the backbone network. BCNN (Branch Neural Network) [23] added two classifier heads (branches 1 and 2) with the green arrow connections, where the branches 1, 2, and 3 predict the green, purple, and red levels of the tree Fig. 1, respectively. We also tested the following combinations: black+green+red connections (R-BCNN: Residual Branch Neural Network); black+grey connections (MTNN: multi-task neural network); black+grey+red connections (MTRNN: multi-task residual neural network)

BCNN: branch convolutional neural network for hierarchical classification [23] proposed to use a convolutional neural network as a backbone and to adjunct two supplementary branches at different depths defining the BCNN (Fig. 2). Each branch is then used to predict a specific level of the classification tree. The loss is finally the weighted sum of the loss of each branch, called the Weighted Average. Those weights are progressively swayed from the higher level branch to the lower one. This approach suggests a breadth path of the classification graph (1).

$$L_{WeightedAvg} = \alpha \cdot L_1 + \beta \cdot L_2 + \gamma \cdot L_3, \tag{1}$$

where L_1, L_2, and L_3 are the cross-entropy losses of the branches 1, 2, and 3, respectively, and α, β, and γ are normalized weight that evolves with the epochs. In our experimentation, we used the backbone as proposed in [23] and preserved the original coefficients, but we increased the periodicity of their changes to match our convergence curves.

Masking for evaluating the loss In opposition to [23], which proposed a network for progressively learning the hierarchical graph level by level for the root to the leaves, [4] proposes to focus on learning the differentiation between sibling nodes within the graph. For that, it proposed a masked evaluation of the cross-entropy loss. This mask filters out (turns to zero) the weights of the current prediction related to all non-adjacent nodes to the target path within the graph.

$$L_{Masked} = \sum_{n \in N} CE(y_n, \bar{y}_n) \cdot p(n), \tag{2}$$

where N, y_n, and $\bar{y}_n$ denote the set of graph nodes, the ground truth label of the node n, and its prediction. $p(n)$ is a binary function that returns 1 if of the parents of n is a ground truth label, 0 otherwise. For instance, if the target label is $A4C$, its ancestors are R, A, and / (outlined Fig. 1), then only the nodes in the red area of Figure 1 will be considered for calculating the loss.

Proposed model: multi-task residual neural network (MTRNN) These two methods have demonstrated that they outperformed the regular cross-entropy loss upon their associated networks. However, how much each strategy is more efficient than the others is undisclosed. Even though Weighted loss and Masked loss have orthogonal approaches, they are not exclusive. In this study, we propose an ablation study that combines those two loss functions and analyses their impact on the learning of several network architectures. As an initial network, we used the BCNN, to which we proposed a few modifications and built three additional variations of this network (Fig. 2). The first additional network, R-BCNN: residual branch neural network, preserves all the connections of the BCNN but additionally concatenates the output of the branch of the upper level of the tree successor branch (black, green, and red connections). In the second additional architecture, we propose to translate the connections of branches 1 and 2 to the

end of the features block of the VGG16 [18]. So we obtain a VGG16 with three classifier heads, named MTNN: multi-task neural network (black and grey connections). Finally, for the third additional network, we reuse the MTNN and add the red connections (black, grey, and red connections), defining the MTRNN: multi-task residual neural network. While the red connections aim to enforce the relationship between the adjacent nodes in the hierarchy, the grey connections propose a different depth for extracting features related to the higher classes. In our experimentation, we have tested combining and dissociating the two introduced methods for calculating the loss upon the four presented networks (BCNN, R-BCNN, MTNN, MTRNN).

2.2 Dataset and pre-processing

In this work, we studied a hierarchical classification task on a Doppler echocardiography videos dataset of Massachusetts General Hospital composed of 249 aortic stenosis patients and 8292 videos acquired with Philips devices, with video labels of nine views (A2C, A3C, A4C, A5C, PLAX, PLAX_RV, PSAX_AV, PSAX_TV, and OTHERS) annotated by three sonographers. Parasternal long axis view focused on the left or right ventricle (PLAX_LV, PLAX_RV), Parasternal short axis view focused on the aortic valve or tricuspid valve (PSAX_AV, PSAX_TV). In the pre-processing phase, videos are decomposed into frame images. Images are masked, cropped, and resized to 224 squared with a black filling, so the embedded metadata surrounding the record is removed. Only the imaging sector of the ultrasound probe remains (Fig. 3). The field of view was not part of the metadata of the DICOM files. It has been estimated by extracting the largest convex hull over the pixels with high variability across the video frames. Table 1 shows the preparation of our dataset within the 9-class classification. According to the steps of sonographers in obtaining different views of the echocardiogram, including localizing, rotating, and tilting the probe, we established the tree structure of these nine views based on the similarity in their imaging procedure and visual appearance, as shown in Fig. 1.

Table 1. Composition of the echocardiography dataset used in our classification task. This table indicates the number of videos per view and splits. The split was made such that each video is exclusive to a unique split.

Split / View	A2C	A3C	A4C	A5C	PLAX	PLAX_RV	PSAX_AV	PSAX_TV	OTHERS
Training	209	446	702	374	775	201	475	133	2328
Validation	60	128	200	107	221	58	136	38	665
Testing	30	64	100	53	111	29	68	19	332

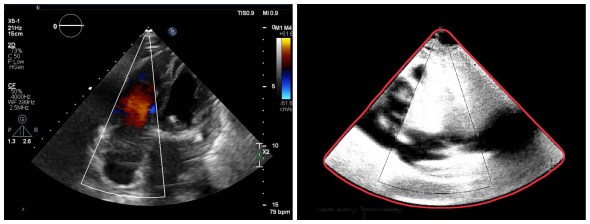

Fig. 3. ROI extraction from Doppler echocardiogram data. Left, original data. Right, the activity map was calculated across all the video frames with the largest convex hull extracted (red contour).

3 Results

3.1 Model implementation and hyper-parameter settings

The implementation of the neural networks was carried out over Pytorch and trained on an NVIDIA A100 GPU with 40GB of VRAM. For the training parameters, we used a batch size of 124 and a learning rate of 0.001 with a scheduler that reduces it by 10^{-1} every 30 epochs. Each network was training over 100 epochs. The source code of all the models tested in this work and the echocardiogram video processing pipeline will be shared with the general public via GitHub (URL anonymized).

Table 2. Average accuracy among the tested combinations by different networks and the strategies for defining the training loss. Avg, W. Avg, and Masked stand for average loss, weighted average loss, and masked loss calculation, respectively. The accuracy is evaluated for each video by voting over all the frames.

Structure/Method	Avg	Avg + Masked	W. Avg + Masked
VGG16	0.90	NA	NA
BCNN	0.92	0.98	0.97
R-BCNN	0.91	0.98	0.97
MTNN	0.92	0.98	0.98
MTRNN	0.92	0.98	0.97

3.2 Running example of the classification result

Figure 4 shows side by side the difference between the hierarchical multi-task architectures investigated in this work (BCNN, RBCNN, MTNN, and MTRNN) and the regular VGG-16, for the same input video. The additional labels predicted ("A" and "R") and the corresponding loss functions would be useful for improving model explainability and generalizability.

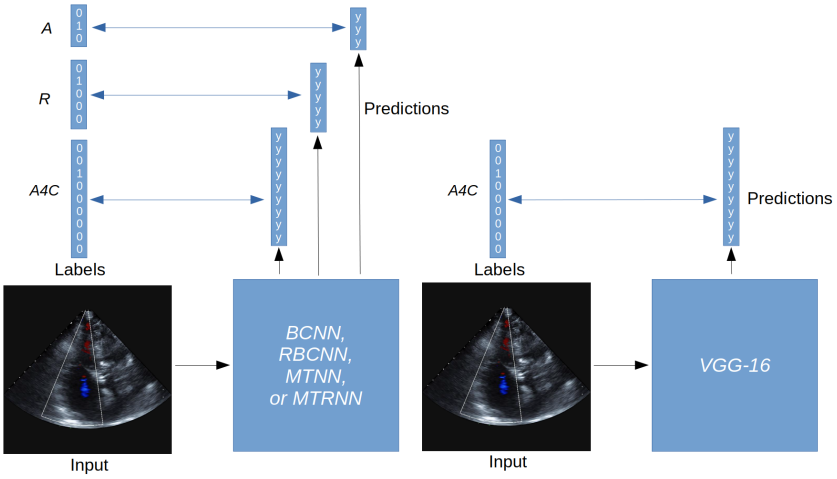

Fig. 4. The architectures observed in this study (left) take as an input an image and predict triple labels associated with a three-level hierarchical graph, where a regular VGG-16 (right) outputs only one prediction.

Table 3. Accuracy per class. The accuracy is evaluated for each video by voting over all the frames.

Average									
Structure	OTHERS	A2C	A3C	A4C	A5C	PLAX	PLAX_RV	PSAX_AV	PSAX_TV
VGG-16	0.86	0.87	0.92	0.93	0.91	0.96	0.90	0.94	0.95
BCNN	0.89	0.87	0.95	0.95	0.96	0.97	0.90	0.94	0.89
R-BCNN	0.88	0.90	0.94	0.95	0.92	0.98	0.83	0.91	0.79
MTNN	0.89	0.90	0.94	0.95	0.92	0.98	0.90	0.94	0.84
MTRNN	0.89	0.90	0.95	0.95	0.94	0.97	0.97	0.93	0.89
Average + Masked									
Structure	OTHERS	A2C	A3C	A4C	A5C	PLAX	PLAX_RV	PSAX_AV	PSAX_TV
BCNN	0.87	0.90	0.98	0.96	0.96	0.99	0.93	0.93	0.84
R-BCNN	0.88	0.97	0.89	0.92	0.94	0.95	0.86	0.90	0.68
MTNN	0.87	0.83	0.95	0.95	0.94	0.98	0.97	0.94	0.84
MTRNN	0.85	0.97	0.95	0.96	0.94	0.97	0.93	0.94	0.79
Weighted Average + Masked									
Structure	OTHERS	A2C	A3C	A4C	A5C	PLAX	PLAX_RV	PSAX_AV	PSAX_TV
BCNN	0.97	0.93	1.00	0.94	0.94	1.00	1.00	0.99	1.00
R-BCNN	0.97	0.93	1.00	0.95	0.96	1.00	1.00	0.99	1.00
MTNN	0.96	0.93	1.00	0.96	0.98	1.00	1.00	1.00	1.00
MTRNN	0.97	0.90	1.00	0.95	0.96	1.00	1.00	0.99	1.00

3.3 Performance comparison

In the experimentation, the four architectures presented above (i.e., BCNN, R-BCNN, MTNN, and MTRNN) have been tested over the same backbone, a VGG-16. Each network has been trained with and without both strategies in-

troduced in Section 2.1, Weighted Average and Masked. We established three different parametrisations for the training of each network. Table 2 shows the overall accuracy obtained for each combination of the experimental workbench and a regular VGG-16 as a baseline. All enhanced architectures outperformed the baseline VGG16 in any configuration based on overall accuracy. In addition, the Weighted Average method not only increases the training time significantly due to its periodic weights but also deteriorates the performances of each network, even on its original network. Using a regular cross-entropy loss (weighted or not), the BCNN comes on top. However, when the masking method is used, the MTNN gets an edge over the other networks. It was noticed that the masking method introduced a higher confusion between the A2c and OTHERS views upon all the networks compared in this article. Lowering into the details, Table 3 shows the accuracy of the tested configuration over each targeted class. At this scale, the negative impact of the Weighted Average is nuanced since it increases the accuracy on OTHER, A3c, PLAX, PLAX_AR, and PSAX_TV.

4 Conclusion and Discussion

This work proposed a framework for more effective learning on data with hierarchically organised labels: multi-task residual neural network (MTRNN) with masked loss. MTRNN integrated two schemes for multi-tasking learning and can perform better for an echocardiogram view classification task. While the proposed model is only validated for a specific task in this work, We envision that it can be easily adapted to other medical image analysis scenarios where the data labels are hierarchically organised, such as thoracic disease diagnosis by chest x-ray images. Furthermore, most object detection tasks in medical imaging can be formulated as a hierarchical multi-task learning problem, as a series of multi-scale regions inherently define the target. Examples of such tasks include but are not limited to skin lesion classification tasks from dermoscopic images [5] and gastrointestinal disease detection using colonoscopy imaging [6]. Furthermore, the proposed scheme can be integrated with the Knowledge Graph in the medical domain, providing knowledge-based guidance from integrated heterogeneous data resources [21]. Thus, we can achieve a more formalised modelling of interrelationships between imaging and its labels.

References

1. Balaji, G., Subashini, T., Chidambaram, N.: Automatic classification of cardiac views in echocardiogram using histogram and statistical features. Procedia Computer Science **46**, 1569–1576 (2015)
2. Bannur, S., Oktay, O., Bernhardt, M.B., Schwaighofer, A., Jena, R., Nushi, B., Wadhwani, S.S., Nori, A., Natarajan, K., Ashraf, S., Alvarez-Valle, J., de Castro, D.C.: Hierarchical analysis of visual covid-19 features from chest radiographs. ArXiv **abs/2107.06618** (2021)

3. Charton, J., Ren, H., Khambhati, J., DeFrancesco, J., Cheng, J., Waheed, A.A., Marciniak, S., Moura, F., Cardoso, R., Lima, B.B., Steen, E., Samset, E., Picard, M.H., Li, X., Li, Q.: View classification of color doppler echocardiography via automatic alignment between doppler and b-mode imaging. In: Aylward, S., Noble, J.A., Hu, Y., Lee, S.L., Baum, Z., Min, Z. (eds.) Simplifying Medical Ultrasound. pp. 64–71. Springer International Publishing, Cham (2022)
4. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hiearchical multi-label classification applied to chest x-ray abnormality taxonomies. CoRR **abs/2009.05609** (2020), https://arxiv.org/abs/2009.05609
5. Hsu, B.W.Y., Tseng, V.S.: Hierarchy-aware contrastive learning with late fusion for skin lesion classification. Computer Methods and Programs in Biomedicine **216**, 106666 (2022)
6. Khaleel, M., Tavanapong, W., Wong, J., Oh, J., De Groen, P.: Hierarchical visual concept interpretation for medical image classification. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). pp. 25–30. IEEE (2021)
7. Khaleel, M., Tavanapong, W., Wong, J., Oh, J., de Groen, P.: Hierarchical visual concept interpretation for medical image classification. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). pp. 25–30 (2021). https://doi.org/10.1109/CBMS52027.2021.00012
8. Kusunose, K.: Steps to use artificial intelligence in echocardiography. Journal of echocardiography **19**(1), 21–27 (2021)
9. Liao, Z., Jafari, M.H., Girgis, H., Gin, K., Rohling, R., Abolmaesumi, P., Tsang, T.: Echocardiography view classification using quality transfer star generative adversarial networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 687–695. Springer (2019)
10. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R.: Fast and accurate view classification of echocardiograms using deep learning. NPJ digital medicine **1**(1), 6 (2018)
11. Mitchell, C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M.C., Horton, K., Ogunyankin, K.O., Palma, R.A., Velazquez, E.J.: Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. Journal of the American Society of Echocardiography **32**(1), 1–64 (2019)
12. Østvik, A., Smistad, E., Aase, S.A., Haugen, B.O., Lovstakken, L.: Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. Ultrasound in medicine & biology **45**(2), 374–384 (2019)
13. Otto, C.M.: Textbook of clinical echocardiography. Elsevier Health Sciences (2013)
14. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
15. Sanh, V., Wolf, T., Ruder, S.: A hierarchical multi-task approach for learning embeddings from semantic tasks. Proceedings of the AAAI Conference on Artificial Intelligence **33**(1), 6949–6956 (2019)
16. Seetharam, K., Raina, S., Sengupta, P.P.: The role of artificial intelligence in echocardiography. Current Cardiology Reports **22**, 1–8 (2020)
17. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

19. Vaseli, H., Liao, Z., Abdi, A.H., Girgis, H., Behnami, D., Luong, C., Dezaki, F.T., Dhungel, N., Rohling, R., Gin, K., et al.: Designing lightweight deep learning models for echocardiography view classification. In: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 10951, pp. 93–99. SPIE (2019)
20. Wu, H., Bowers, D.M., Huynh, T.T., Souvenir, R.: Echocardiogram view classification using low-level features. In: 2013 IEEE 10th International Symposium on Biomedical Imaging. pp. 752–755. IEEE (2013)
21. Zhang, Y., Sheng, M., Zhou, R., Wang, Y., Han, G., Zhang, H., Xing, C., Dong, J.: Hkgb: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. Information Processing & Management **57**(6), 102324 (2020)
22. Zhao, J., Peng, Y., He, X.: Attribute hierarchy based multi-task learning for fine-grained image classification. Neurocomputing **395**, 150–159 (2020)
23. Zhu, X., Bain, M.: B-cnn: Branch convolutional neural network for hierarchical classification. ArXiv (2017). https://doi.org/10.48550/ARXIV.1709.09890