Learn Class Hierarchy using Convolutional Neural Networks

Riccardo La Grassa[a], Ignazio Gallo[a], Nicola Landro[a]

[a]*University of Insubria, Department of Theoretical and Applied Science, Varese, Italy*

ABSTRACT

A large amount of research on Convolutional Neural Networks has focused on flat Classification in the multi-class domain. In the real world, many problems are naturally expressed as problems of hierarchical classification, in which the classes to be predicted are organized in a hierarchy of classes. In this paper, we propose a new architecture for hierarchical classification of images, introducing a stack of deep linear layers with cross-entropy loss functions and center loss combined. The proposed architecture can extend any neural network model and simultaneously optimizes loss functions to discover local hierarchical class relationships and a loss function to discover global information from the whole class hierarchy while penalizing class hierarchy violations. We experimentally show that our hierarchical classifier presents advantages to the traditional classification approaches finding application in computer vision tasks.

1. Introduction

In recent years researchers have become increasingly interested in the multi-label and hierarchical learning approaches, finding many application to several domain, including classification Wehrmann et al. (2018); Cesa-Bianchi et al. (2006), image annotation Dimitrovski et al. (2011), bioinformatics Valentini (2009); Yan et al. (2019) Chen et al. (2018) Abacha et al. (2019). Nowadays, machine learning is commonly used to resolve complex problem into pattern recognition where an object is classified assigning a label in according with the model's rule used. However, classes are not always disjoint from others and objects within them can be related to others as a hierarchical structure Silla and Freitas (2011). Human beings perceive the world with different types of granularity and can translate information from coarse-grained to fine-grained and on the contrary, perceiving different levels of abstraction of the information acquired Hobbs (1990); McCalla et al. (1992). This concept is reflected in the taxonomy of the multi-label general approaches under the idea of structured output prediction Su et al. (2015).

In terms of neural models, the main difference between the prediction of structured output and flat multi-label classification lies in the level of neurons that contains the label prediction. In fact, in the presence of a structured output, the information is based on a different level of abstraction, while with the multi-label flat approach it is based on a single level.

Hierarchical multi-label classification (HMC) is a variant of the classification task where instances may belong to multiple classes at the same time and classes are organized in a hierarchy. In HMC approaches a relationship among classes and can be formalized by a tree or directed acyclic graph (DAG). Our approach to HMC exploits the annotation hierarchy by building a single neural network that can simultaneously predict all categorization of an input source exploiting multiple layers of a neural model. For example, considering the class label prediction for an image containing a tiger, the proposed system can simultaneously predict that a "tiger" has been found but at the same time the same object is also a "feline" and a "mammal".

In literature exists two main approaches to HMC problem, known as local and global Costa et al. (2007); Xu et al. (2019); Silla and Freitas (2011). In the global approach, the output of the final layer predicts the test instance in which only one classifier sees information globally without having local information. In the local approach, there is a set of trained classifiers that follows a top-down strategy, in particular, the training process is independently for each base classifier.

Different local approaches have been proposed in the literature, like Local classifier per Node (LCN) Valentini (2009), Local classifier per parent node (LCPN), Local classifier per level (LCL) Cerri et al. (2011). LCN strategy trains a local classifier for each node of a graph providing a local decision to make predictions. LCPN uses a multi-class classifier for each internal class to recognize classes from its sub-classes and LCL methods train a multi-class classifier per hierarchical level. In contrast with local (LCN, LCL, LCPN) and global approaches, we use a single trained model and a single back-propagation error with many different layers fully connected, responsible to synchronize with a concept linked to a given hierarchical structure.

A recent work Wehrmann et al. (2018) describes a novel method to solve HMC problem, that preserves local and global information simultaneously to discover the local hierarchical relationship among classes. Unlike this work, our architecture exploits recent neural network potentialities and facilitates the multi-class prediction for each deep layers to capture local con-

★★*This paper is under review to Applied Intelligence journal*

text following the hierarchical structure of the information. In our approach, we have a cascade of fully connected linear layers each one with softmax plus cross-entropy, where the output of a layer $l-1$ is the input of layer l; instead, in Wehrmann et al. (2018) the model has ReLu activation functions on two different layers fully connected with softmax and binary cross-entropy per block. Another difference with Wehrmann et al. (2018) is that the input of each layer l fuse with the input, instead, in our approach the input per layer is the output of the previous layer. The last difference is that our model uses local classification as final prediction in according to hierarchical multi-label classification task, instead of in HMCN-F the final layer is used as flat layer plus another layer that uses jointly local and global output information to obtain the final prediction.

Our work can be summarized in the following key contributions:

- We propose a new hierarchical deep loss approach (HDL) as an extension of convolutional neural networks to assign hierarchical multi-labels to images. Our extension can be adapted to a generic Convolutional Neural Network as final step.
- To prove the effectiveness of our hierarchical classification approach we conduct empirical studies on three different datasets. First, we created *Animals_Taxonomy8* dataset based on real animal images from Flickr on three groups of taxonomy (Class, Family, Species) with their relative label annotations. Second, we used a well-known biomedical dataset (*VQA-Med 2019*) contains radiology real images on different levels of hierarchy and third, we created *Geometry_shapes_annotations* that contains thousands of shapes images on three depth hierarchy levels. Further, all datasets have a different number of instances (2.8k,8k,40k) useful to prove the robustness of our approach.

2. The Proposed Approach

As mentioned above, our solution is an architectural extension that can be adapted to a generic neural network. In this paper, we used a standard Convolutional Neural Network, the ResNet18, as a base model to which we added our solution to solve a hierarchical images classification problem. As graphically represented in Figure 1, what we do is to extend the output layer with some fully connected layers equal to the number of layers available in the classes hierarchy tree of the problem to be solved, and to associate a loss function to each of these new layers added. In practice, we construct a mapping between the layers of a class hierarchy and the new layers of the neural network ($linear_1, \ldots, linear_N$ in Figure 1) so that the network can learn to discriminate between all class labels belonging to a given layer of the hierarchy. To minimize the intra-class variance and at the same time to keep the features among different classes separated we compute the Center Loss Wen et al. (2016) on each training mini-batch and update all class centers after each training epoch. More formally we compute the center loss $\mathcal{L}_C$ as follow:

$$\mathcal{L}_C = \sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{1}$$

where $c_{y_i} \in \mathbb{R}^d$ denotes the center for the class y_i in the features space of the deep model. In our experiments, we chose a Resnet-18 as a general model and apply Center Loss after the adaptive pooling layer. Finally, let l_1 be the linear layer at first level with dimension equal to the number of classes at first level of hierarchy, more formally:

$$l_G^1 = \phi(W_G^1 x + b_G^1) \tag{2}$$

where $W_G^1 \in \mathbb{R}^{|l_G^1|\times|d|}$, $b_G^1 \in \mathbb{R}^{|l_G^1|\times 1}$ is the bias vector with ϕ linear activation function and d be the number of features. Then, we add a linear layer l for each hierarchical level in a generic dataset and we perform the cross-entropy loss to maximize the inter-class variance. Precisely, we apply softmax function from logits of layer l_G^l and use cross-entropy loss as Eq.3

$$\mathcal{L}_{l_G^l} = -\sum_{i=1}^{m} log \frac{e^{W_{y_i}^T x_i + b_{yi}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{3}$$

Where, l is the layer l-th, m and n are the mini-batch size and number of classes respectively, $x_i \in \mathbb{R}^d$ denotes the ith deep feature, belonging to the y_ith class and b is the bias.

Finally, our total loss is:

$$\mathcal{L} = \lambda_0 \cdot \mathcal{L}_C + \lambda_1 \cdot \mathcal{L}_1 + \cdots + \lambda_N \cdot \mathcal{L}_N \tag{4}$$

Where $\lambda_{0,1,\ldots,N} = 1$, $\mathcal{L}_C$ it the centers loss value and $\mathcal{L}_{0,1,\ldots,N}$ is the cross-entropy loss value of the layer $\{1, ..., N\}$. The general formulation with N layer is defined as Eq.5

$$\mathcal{L} = \lambda_0 \cdot \mathcal{L}_C + \sum_{l=1}^{N} \lambda_l \cdot \mathcal{L}_l \tag{5}$$

3. Datasets

To evaluate the proposed method, we created our own datasets as there is no a standard benchmarked dataset on hierarchical multi-label images classification, available in the literature.

The medical Visual Question Answering task (VQA-Med 2019) Abacha et al. (2019) is focused on radiology images (example in Fig. 4) grouped in four main classes: Modality, Plane, Organ system, Abnormality. The original challenge is to classify an image from a question linked to it, indeed for each image in the training we have a paired question. Our focus is on the hierarchical multi-label classification of images, therefore, we will exclude our experiment from text classification task. We use all train size and use the validation set as a test set (because the test set is not labelled with all labels), respectively 2816/340 objects. In total, we consider three levels of hierarchy (Modality Class, Plane Class, Organ Class) with their relative different type of concepts. These classes have a size of 44, 15, 10 respectively per classes. In these experiments, our goal is to prove experimentally the effectiveness and robustness of our model to discriminate different concepts also in the case we have a few examples per classes in the train.

We have created a synthetic geometric shapes dataset which contains 2 different shapes (Triangle, Square, some image sample into fig. 2) at the first level of our hierarchy. Each shape has

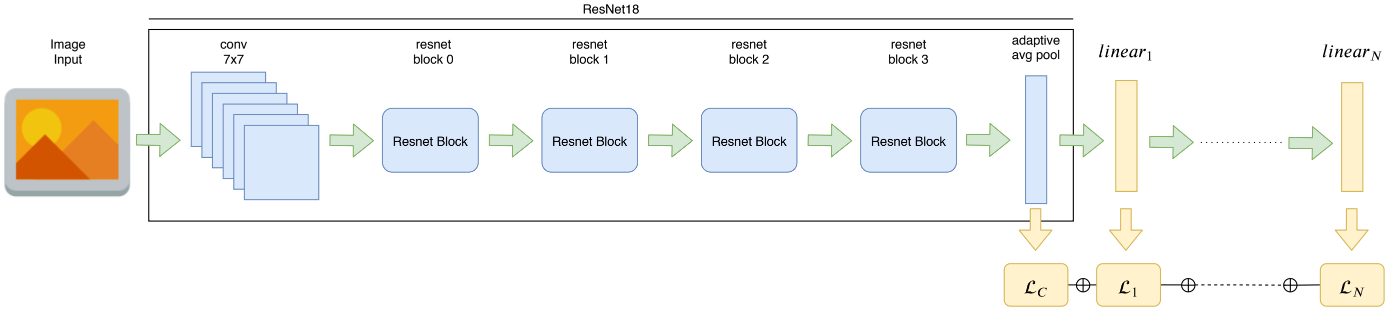
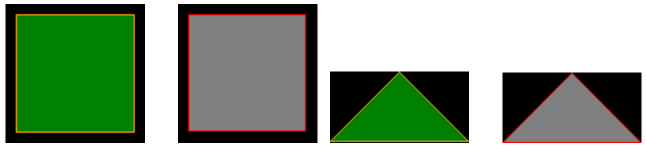


Fig. 1: The proposed architecture of HDL



(a) Green square with orange border (b) Gray square with red border (c) Green triangle with red border (d) Gray triangle with red border

Fig. 2: Some sample of geometric dataset.

6 different full colours and other 6 different colours for out-fill, the last two represents the second and third level of the hierarchy. The possible configuration is 72 so, we generate train/test with 20000 and 6000 objects respectively. The dimension of the images is 128x128x3. In these experiments, we want to answer the question ”Which kind of shape is this? What is the fill colour? and the out fill colour?.

This data set is created from Flickr animals images, the hierarchy represents a small taxonomy with class, family and species as in Fig. 6. The selected **class** is *mammalia* and *reptilia*. The second level of hierarchy is the **family**, in particular *felidae* and *ursidae* for mammalia and *crocodyle*, *iguanidae*, *emydidae* and *pythonidae* for reptilia. The last hierarchy level represent the **species**(example of images in fig. 5) as *malaysia tiger*, *felis catus* known as cat, *ailuropoda melanoleuca* known as giant panda, *ursus maritimus* known as polar bear, *python molurus* known as green python, *trachemys scripta* as small turtle, *iguana iguana* and *crocodylus niloticus* well known as nilus crocodile. A whole representation of the dataset is in Fig. 3.

4. Experiments

To evaluate the proposed method we develop four empirical studies.

1. In the first one, we use a well-known dataset (VQA-Med 2019) to test our approach with biomedical real images, also in the case we have few data available.
2. In the second we test the capability of abstraction of our approach on a synthetic dataset created in the context we have thousand of instances available.
3. In the third, we extract hierarchical structure on the real-images dataset contains images of three types of animal taxonomy levels (Class-Family-Species) and prove the robustness of our HDL in the case which images are hard to recognize and they contain noise.
4. In the four experiment, we compare our HDL with a ResNet18 proving the effectiveness of our approach.

First experiment In this experiment, we test our model in the case we have few instances and with a high complexity of images. Our hypothesis is that the performance in terms of accuracy in a layer is higher when the number of different concepts to distinguish is inferior to a layer with many concepts to recognize. As we show in 1 at the first row (VQA-Med 2019), we have accuracies of 38.05, 74.04, 66.66 for the size of layers 44, 15, 10 respectively. We can observe that the accuracy of the first layer is lower of 1.94 times than the second layer and to 1.75 times than the third layer, this proves that our model offers better scalability when we have few concepts per layer to learn. Similar results can be found in *Animals_Taxonomy8*, where the higher accuracy of the third layer at the third row of Table 1 than others, is due to the fact we have only two concept (mammals or reptiles) to distinguish than the second layer (8 concepts) Figs. 7 and 8.

Second experiment In a second experiment, we use a synthetic dataset with simple geometric shapes and several instances 7.10 times greater than VQA-Med 2019. Our intuition is that attribute more samples per classes can improve the training of our model and subsequently, to obtain better performance in terms of accuracy than the first experiments. To prove this conjecture, we train with 20K instances our HDL and test it with 6k instances. The results in Table. 1 at the second row per tables, confirm our expectations. The higher number of instances jointly with the simplicity of images allows the model to reach high accuracy starting from the first ten epochs. Furthermore, we conduct three different runs with a learning rate of 0.005, 0.001, 0.01 using batch-size of 64.

Third experiment In these experiments, we test our model using more instances than the first experiment and with images of animal (*Animals_Taxonomy8*) that contains noise. In particular, our model offers good performance also in the case the images are not simple as in the second experiments and when they contain noise or offers little comprehensibility, indeed many images are not clear, like for example a snake completely hidden by forest or a bar sign with a panda logo. However, as we show in 1, the accuracy of the third layer, responsible to recognize mammals or reptile is very high. We conclude that

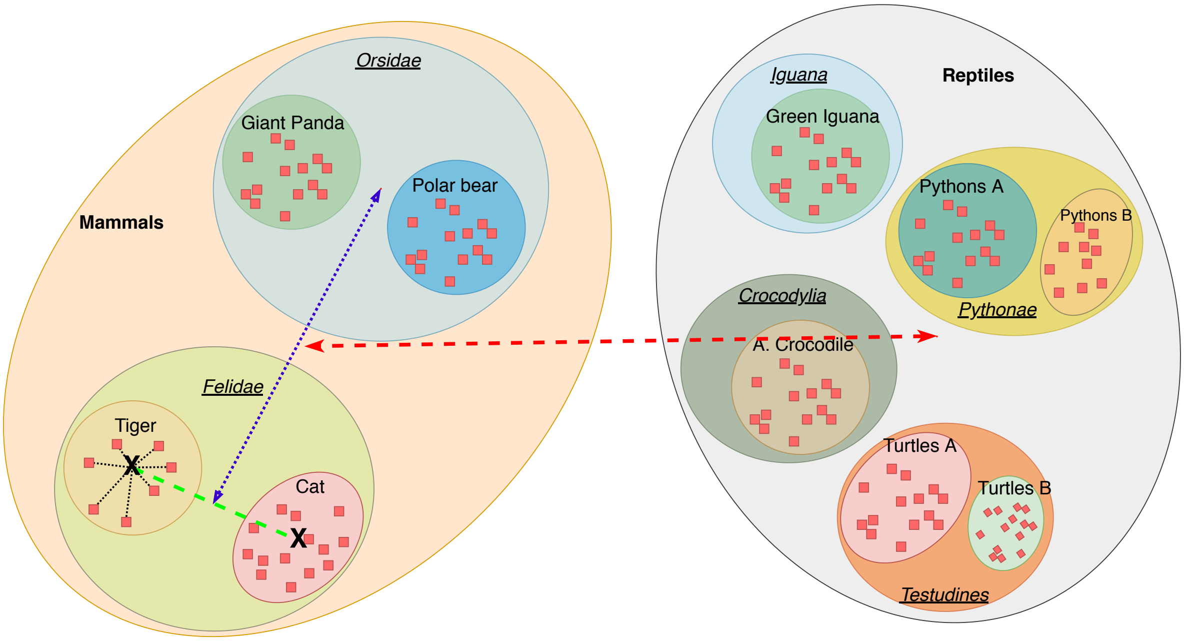
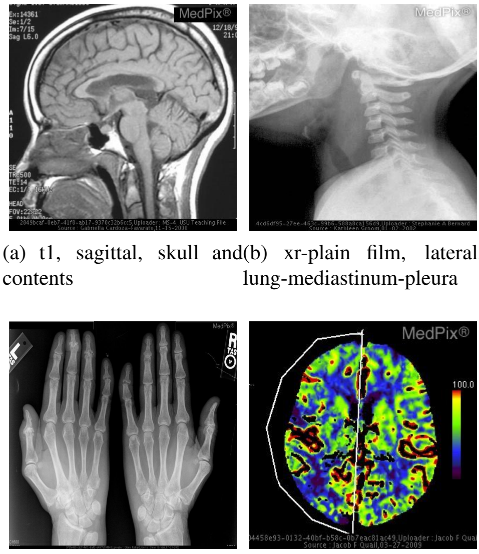


Fig. 3: Hierarchy of categories used in *Animals_Taxonomy8*



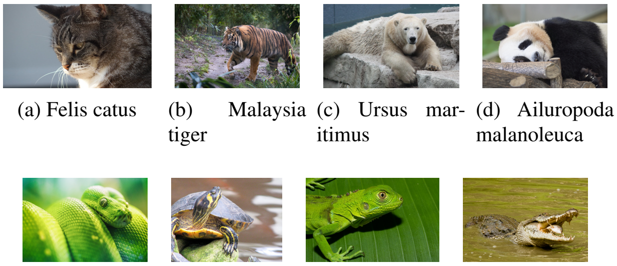
(a) t1, sagittal, skull and contents

(b) xr-plain film, lateral, lung-mediastinum-pleura

(c) xr-plain film, ap, musculoskeletal

(d) ct w/contrast (iv), axial, skull and contents

Fig. 4: Four images extracted from VQA-Med 2019 Abacha et al. (2019) dataset (synpic371, synpic10103, synpic16486, synpic48315). The labels separated by comma belong to the sub-categories of the three main classes following the order: Modality, Plain, Organ.



(a) Felis catus

(b) Malaysia tiger

(c) Ursus maritimus

(d) Ailuropoda malanoleuca

(e) Python molurus

(f) Trachemys scripta

(g) Green iguana

(h) Crocodylus niloticus

Fig. 5: Example of images extracted from *Animals_Taxonomy8*

considering the poor understanding of images, noise and hard images to recognize, experimental results prove the robustness of our model.

Fourth experiment HDL is designed to maximize the learning capacity and to extract the hierarchical structure from the labelled data. Our intuition is that our model, lead to different losses at any level, with the power to reduce intra-variance and to maximize inter-variance, can obtain better accuracy than a classical convolutional neural network. To prove this, we conduct six different experiments using a classic ResNet18 and our HDL on *Animals_Taxonomy8* using two learning rate and a batch size of 64. The results in Fig. 2 and 9,10 clearly confirm our expectations. In all cases, the accuracy is higher than a

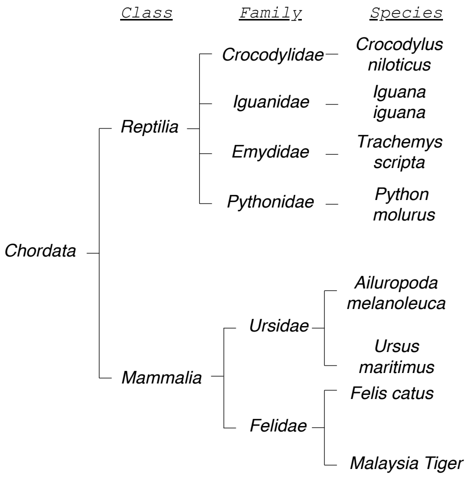


Fig. 6: *Animals_Taxonomy8* dataset classification hierarchy.

classical ResNet18, this experiment proves the effectiveness of our proposed model.

4.1. Experiments settings

We build our hierarchical multi-label classifier model as an extension on a Resnet-18, but is it possible to apply to any Convolutional Neural networks. We implement our extension in Python using Pytorch framework. Fig.1 shows the architecture used for experiments. The size of the input images is re-scaled to 64x64x3 for Geometry dataset and 256x256x3 for *VQA-Med 2019* and *Animals_Taxonomy8* datasets. We do not apply any preprocessing of images as data augmentation, rotation or normalization. The kernel size of the first convolutional layers is 7x7 with a stride of 2 pixels, followed by a normalization of layer and a non-linear layer with ReLu activation function. A max-pooling operation over 3x3 regions and a stride of 2 pixels. Then, we have four blocks of Convolution, with 64, 128, 256, 512 numbers of plans respectively and apply an adaptive average pooling over 1x1 region. Finally, we add three fully connected linear layer, where each layer corresponding to the total number of concepts in our hierarchical dataset. In the forward process, we take the output after the adaptive average pooling and apply Center loss function and for each linear layers we apply softmax function and then cross-entropy loss. The total loss will be the sum of the local loss per layers. Our network was trained with Adam optimizer Kingma and Ba (2014). The batch-size used, learning rate, epochs are described jointly with the results for each dataset.

5. Results and Discussion

This study is placed in the sub-category of multi-label classification called Structure output learning. In according with experimental results at Tables 1, 2, we achieved good results on three different datasets finding the way to exploit the dependency among classes and make accurate predictions, reducing the misclassification than a classic ResNet18. The main reason we have created these datasets is to prove our proposal in the field of computer vision and with more than 2 levels of

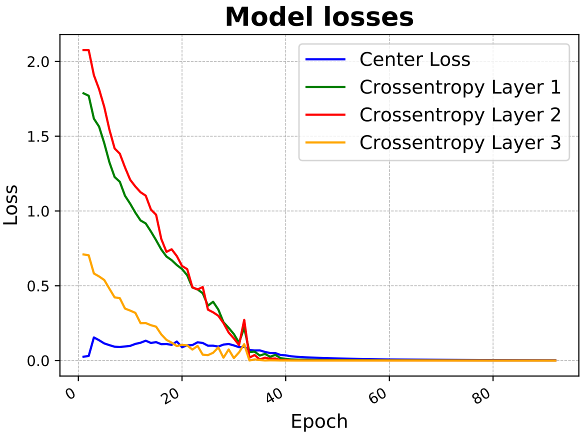


Fig. 7: Training losses *Animals_Taxonomy8* with *lr* = 0.01

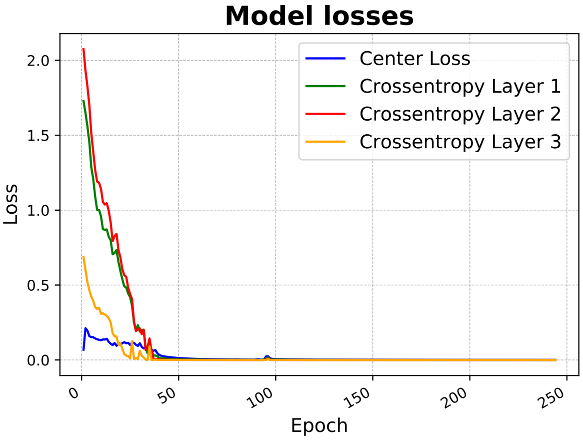


Fig. 8: Training losses *Animals_Taxonomy8* with *lr* = 0.005. We emphasize the different descent of losses. This is due to the number of concepts to distinguish from each layer. Each line represents the loss for each layer. For this dataset, we design our model with a shape of 6,8,2 to distinguish Family, Species and Classes respectively. As we show also in 1, the line yellow that represents linear layer 3 with 2 concepts (Mammals or Reptiles) has more descent power, indicating that our model quickly learns a few concepts rather than many as red line or green.

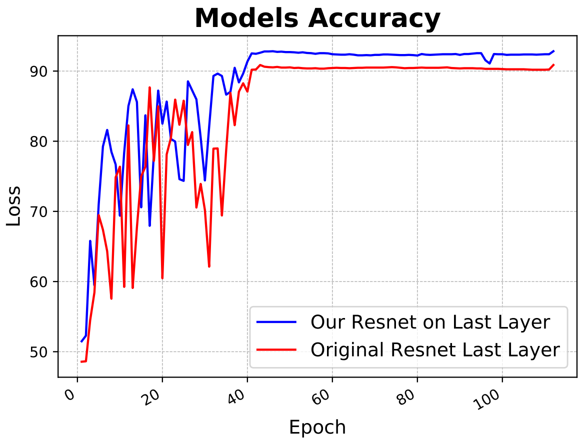


Fig. 9: HDL vs original Resnet18 with lr = 0.005

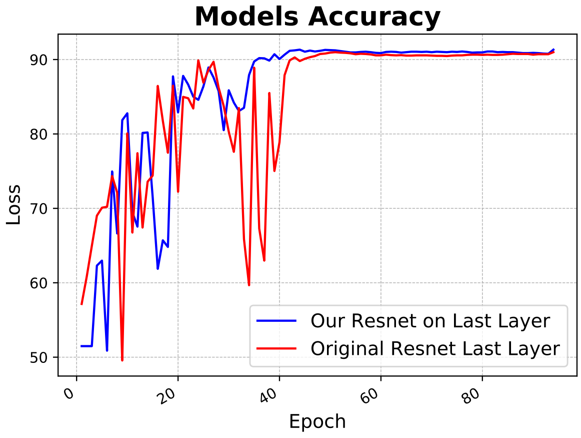


Fig. 10: HDL vs original Resnet18 with lr = 0.01

Table 1: Accuracies comparison using three different datasets on different learning rate. We use a batch size of 32 on VQA-MED and 64 on the other datasets.

		lr=0.005	
Datasets	11	21	31
VQA-Med	38.05%	74.04%	66.66%
Shapes	100%	100%	100%
Animals_Taxonomy8	71.98%	69.07%	92.82%
		lr=0.001	
VQA-Med	35.98%	70.20%	67.84%
Shapes	100%	100%	100%
Animals_Taxonomy8	72%	69.12%	92.89%
		lr=0.01	
VQA-Med	34.8%	71.97%	69.61%
Shapes	100%	100%	100%
Animals_Taxonomy8	69.2%	66.53%	91.32%

depth, indeed CIFAR100 contains only two levels of depth (Super Class, Classes) and other datasets with many depths find applicability only in text classification or in bioinformatics, where the inputs are not images.

6. Conclusion

In literature, multi-label classification is an important field in machine learning and it is strongly related to many real-world applications for example, in biomedical images annotation, document categorization and whatever problem which the instances inside the classes are not disjoint but they keep a hierarchical structure. In this work, we have conducted four empirical studies on different datasets to prove by experimental results the effectiveness and robustness of our proposed model, that can be applied as an extension to any Convolutional Neural Network.

References

Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: CLEF2019 Working Notes. CEUR Workshop Proceedings, pp. 09–12.

Cerri, R., Barros, R.C., de Carvalho, A.C., 2011. Hierarchical multi-label classification for protein function prediction: A local approach based on neural networks, in: 2011 11th International Conference on Intelligent Systems Design and Applications, IEEE. pp. 337–343.

Cesa-Bianchi, N., Gentile, C., Zaniboni, L., 2006. Incremental algorithms for hierarchical classification. Journal of Machine Learning Research 7, 31–54.

Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P., 2018. Deep hierarchical multi-label classification of chest x-ray images .

Costa, E.P., Lorena, A.C., Carvalho, A.C., Freitas, A.A., Holden, N., 2007. Comparing several approaches for hierarchical classification of proteins with decision trees, in: Brazilian Symposium on Bioinformatics, Springer. pp. 126–137.

Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S., 2011. Hierarchical annotation of medical images. Pattern Recognition 44, 2436–2449.

Hobbs, J.R., 1990. Granularity, in: Readings in qualitative reasoning about physical systems. Elsevier, pp. 542–545.

Table 2: Accuracies comparison of our model with a original ResNet on different layers. The batch size used for these experiments is 64.

lr=0.005				
Our Model	11	21	31	ResNet18
Animals_Taxonomy8	**71.98%**	-	-	71.19%
	-	**69.07%**	-	68.58%
	-	-	**92.82%**	90.86%
lr=0.01				
Our Model	11	21	31	ResNet18
Animals_Taxonomy8	**69.2%**	-	-	68.34%
	-	**66.53%**	-	65.36%
	-	-	**91.32%**	90.98%

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

McCalla, G., Greer, J., Barrie, B., Pospisil, P., 1992. Granularity hierarchies. Computers & Mathematics with Applications 23, 363–375.

Silla, C.N., Freitas, A.A., 2011. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery 22, 31–72.

Su, H., et al., 2015. Multilabel classification through structured output learning-methods and applications .

Valentini, G., 2009. True path rule hierarchical ensembles, in: International Workshop on Multiple Classifier Systems, Springer. pp. 232–241.

Wehrmann, J., Cerri, R., Barros, R., 2018. Hierarchical multi-label classification networks, in: International Conference on Machine Learning, pp. 5225–5234.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition, in: European conference on computer vision, Springer. pp. 499–515.

Xu, D., Shi, Y., Tsang, I.W., Ong, Y.S., Gong, C., Shen, X., 2019. A survey on multi-output learning. arXiv preprint arXiv:1901.00248 .

Yan, X., Li, L., Xie, C., Xiao, J., Gu, L., 2019. Zhejiang university at imageclef 2019 visual question answering in the medical domain. Working Notes of CLEF .