EnGraf-Net: Multiple Granularity Branch Network with Fine-Coarse Graft Grained for Classification Task

Riccardo La Grassa$^{1\,[0000-0002-4355-0366]}$, Ignazio Gallo$^{1\,[0000-0002-7076-8328]}$, and Nicola Landro$^{1\,[0000-0002-0565-7496]}$

University of Insubria, Department of Theoretical and Applied Sciences, Varese, Italy
{rlagrassa}@uninsubria.it

Abstract. Fine-Grained classification models can expressly focus on the relevant details useful to distinguish highly similar classes typically when the intra-class variance is high and the inter-class variance is low given a dataset. Most of these models use part annotations as bounding box, location part, text attributes to enhance the performance of classification and other models use sophisticated techniques to extract an attention map automatically. We assume that part-based approaches as the automatic cropping method suffers from a missing representation of local features, which are fundamental to distinguish similar objects. While Fine-Grained classification endeavours to recognize the leaf of a graph, humans recognize an object trying also to make a semantic association. In this paper, we use the semantic association structured as a hierarchy (taxonomy) as supervised signals and used them in an end-to-end deep neural network model termed as EnGraf-Net. Extensive experiments on three well-known datasets: Cifar-100, CUB-200-2011 and FGVC-Aircraft prove the superiority of EnGraf-Net over many Fine-Grained models and it is competitive with the most recent best models without using any cropping technique or manual annotations.

Keywords: Fine-Grained classification · Hierarchical Classification.

1 Introduction

In Neuroscience, pattern separation is a process defined as the capability to discriminate a set of similar patterns into less-similar sets of outputs patterns. In [28], the authors show the evidence of the capability of the pattern separation in the Dentate Gyrus (DG) neurons and the pattern completion (a complementary process of pattern separation) in the CA3 neurons. DG and CA3 area of the hippocampus have been long hypothesized to be responsible for these processes and [28, 7, 27] provides strong empirical support for this functional dissociation. In [29], entitled *CA3 Sees the Big Picture while Dentate Gyrus Splits Hairs*, the authors support the same idea and provide furthermore result to this conclusion. Again, in [25], theoretical models suggest the DG performs pattern separation of cortical inputs before sending its differentiated outputs to CA3. Indeed, DG is ideally located to do this, receiving signals via the major projection from the entorhinal cortex (EC), the perforant path (PP), and sending signals to CA3. These results provide vigorous support for long-standing hypotheses attributing each hippocampal sub-region with distinct roles in neural information processing and set the stage for exciting new research [25]. The deep learning models separate the main signal (e.g. images, sounds, text) in small signals using convolutional operation useful to improve the discrimination ability in the pattern recognition task. Recently, some works are considering forcing the pattern separation process using the semantic association (e.g. hierarchical structure) that comes from the hierarchy abstraction or by manual/automatic text annotation extracted for each image to achieve better performance of a deep learning model. Many of them apply sophisticated methods to extract specific crops on images in order to get more high discriminative features [38, 43, 13, 42, 36] instead to consider all manual annotations from a dataset to get them. In Computer Vision (e.g. Fine-Grained classification) implies a hierarchy organization structure composed by different levels of abstraction and it can be represented by a graph, in which all nodes closer to the root represents the abstract concept and as deep as we go far from the node root we find finer-grained abstraction. Also, humans use hierarchical information to recognize a specific object when it is unknown, therefore the categories hierarchy provides a rich semantic correlation among different categories across many levels of abstraction. In the learning process, this guidance can have a regulating effect on semantic space and can lead an algorithm to get better discriminative features for the fine-grained recognition task. In [4], the authors designed a model which considers different granularity levels and proves the usefulness to consider this information to enhance the capability of the main model. Again, in [17], the authors use hierarchical annotation taken by Word-Net to build an end-to-end model to focus on final classification jointly with the hierarchical classification task. They use a simple multi-layer perceptron considering 3 levels of abstraction demonstrating the capability of a model to solve both recognition tasks. The idea to feature fusion considering a multi-scale model was introduced in [20]. Recent works as [21, 12] brings to light new interesting architecture to feature fusion at different levels of a deep model. These approaches use the lateral connections of a deep model to carry out fusion operations and combine them widely. In our approach, we use the semantic association as a hierarchical supervised signal to improve the ability of pattern recognition. In Fine-Grained classification, the focus of the most recent deep models is to generate an attention map that contains high discriminative feature such that they can outperform the results in the classification. However, the spatial information (e.g. all regions that contain the environment of the object itself) can also contain useful features to help the pattern recognition ability by models. In [24], observations of five species of Warbler proves that species divide up the resources of a community in such a way that each species is limited by a different factor, such for example the tree partition. Authors show the tree partitioning where at a certain percentage it is possible to find a species in a specific location of the trees. The environment (spatial information) in which the objects can be found is very important and must be considered by modern deep learning models. In our proposal, we do not avoid spatial information using cropping technique, but we consider all information without using any specific region location. In this paper, leveraging by the action makes by DG in our brain we simulate the pattern separation ability of DG neurons using a supervised approach through the semantic association extracted from the hierarchical information of the datasets. We force the pattern separation in a deep model in order to get discriminative features useful to recognize the hierarchy of the objects and distinguish very similar objects. The scientific contribution of this work is concluded as follows:

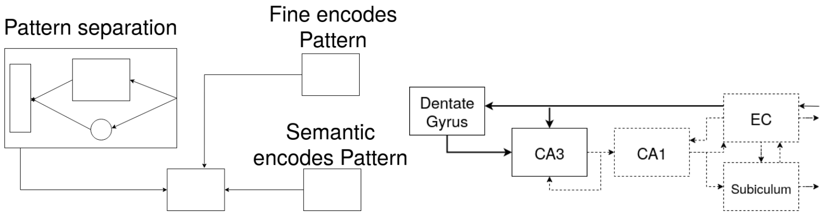
1. We introduce a Multiple Granularity Branch Network with Fine-Coarse graft grained for Fine-Grained classification task. Our model termed as EnGraf-Net, uses the hierarchical semantic associations from the datasets to force the pattern separation and improve the discrimination capability of a deep learning model.
2. We conduct experiments on Cifar-100, CUB200-2011 and FGVC-Aircraft datasets and demonstrating the effectiveness of our proposal over the baselines and proves to compete with the most recent algorithms compared. We investigating also in the contribution of each components using the Resnet family models conducting ablative studies. We released the code and all experimental reports at [16].

1.1 Related work

NTS-Net [38] introduces a self-supervised mechanism to locate informative regions without using the bounding box and part annotations. Many works as [37, 33, 15, 11], take advantage of fine-grained human annotations, like the location of some details of images. However, human annotations are expensive and far away from the deep learning concept where every single concept has to be automatic. NTS-Net [38] uses a mechanism to localize informative regions automatically (Navigator) and a Teacher module that evaluates the probability to belong to the ground-truth class using these regions extracted by the Navigator module. Finally, a Scrutinizer module uses these regions to make fine-grained classifications. The model takes the top-M informative regions with the highest score got and these last can represent a weakness of NTS-Net because of the fixed number of regions taken. In [14], authors developed a localization module integrated into an end-to-end setup that generates an attention map and then is used to predict the bounding box of the discriminative regions. The main model is composed of three main modules. The first two modules termed as *AttNet* and *AffNet* has the goal to perform the localization using a combined max-pooling method that merges the vertical/horizontal transformation. Finally, the last model represents the baseline useful to the make classification. In [13], similar to *Affnet*, a method was proposed to search relevant images regions introducing a module trained to build an attention map and a Global K-Max pooling function useful to find a single feature vector that describes the image. The final model requires multiple separate training runs instead to have an end-to-end model. In [43], authors proposed an *attentive pairwise interaction network* for Fine-grained classification based on the idea that humans often compare pairs of images jointly to recognize subtle differences between similar objects. Their method uses two paired images as input and cross-entropy (CE) loss function with a score ranking regularization. In this paper, we compare us with the most recent models who obtain excellent performance in Fine-Grained classification task and we conduct extensive ablation studies analyzing the performance of our proposal.

2 Methodology

The hippocampus and related structures have the capability to minimizes the sets overlap between similar patterns (pattern separation) and to reconstruct complete stored rep-



(a) (b)

Fig. 1: a) Schematic diagram of our EnGraf-Net b) Schematic diagram of the regions of the hippocampus. The figure shows the feedforward pathway from the entorhinal cortex to the DG and the CA3 neurons. The EC, DG and CA3 blocks are very similar to ours blocks. We simulate the process of the pattern separation by DG and the others connection (EC, DG, CA3) with our proposed approach.

resentations from partial patterns that are part of the stored representation (pattern completion). In Fig. 1(b) we show the main pathway diagram of the hippocampus regions and the structure similarity than our proposed approach (see Fig. 1(a)). Observing the nature of this process, we try to simulate the pattern separation/completion as a module engrafted into a branch of a convolutional neural network and analyze the performance model in Fine-Grained classification task. We force a branch through a graft to obtain two supervised patterns that come from the truth of the fine labels and the coarse labels (pattern separation) and finally, we concatenate these patterns into one going towards the next steps of EnGraft-Net (pattern completion). Instead to use manual/automatic annotation comes from images as supervised truth, we extract the semantic association that comes from the hierarchy of the entire datasets used. A semantic association is a process that quantifies the strength of the semantic connection between textual units, taking into account different kinds of relationships and it is an indispensable section of various applications having a spot with a huge number of fields, for instance, Cognitive Psychology and Computer Science. When a semantic association is organized as hierarchical structure, it is called *Taxonomy*. We use the *Taxonomy* of datasets and use the semantic association (class, superclass) as supervised signals useful to compute the loss functions used and we use a combination of different patterns comes from different branches to enhance the discriminative power and increase the main performance of the model. More precisely, given y^K be the fine-grained label from a dataset, we build upon y^K label the superclasses label y^{K-1}. Each image x is annotated using different granularity y^{K-1}, y^K and C_{K-1}, C_K is the number of the class categories considered. Our goal is to correct classify images x across two different types of granularity using an end-to-end model and CE loss functions.

2.1 Network architecture

EnGraf-Net is based on Resnet family networks. We use a multi-branch approach (see Fig. 1) where the first two branches have the goal to find discriminative features using

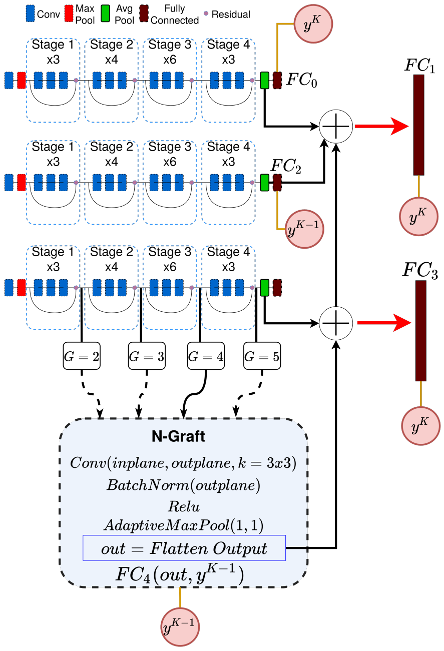


Fig. 2: An overview of our proposed EnGraf-Net model. It employs two branches to extract features at different grain and a third branch network where we engraft a subnetwork useful to apply the pattern separation process.

two types of supervised signals: all labeled classes of fine grained and all labeled superclass of coarse grained extracted by the semantic annotation of a the dataset. The third branch is responsible to make the pattern separation/completion through both supervised signals (fine/coarse grained labels). We can choose different type of grafting. The graft block is composed by a convolutional layer, batch normalization and *relu* activation function. Then, we use an adaptive max pooling with output 1×1 and finally after a flatten operation of the output we have a fully connected layer, where y^{K-1} represents the hierarchy class labels used. We concatenate all signals from different branches and use fully connect layers where the last loss function is applied. Depending on which model of Resnet we select the total numbers of parameters of EnGraf-Net is increased.

2.2 Loss functions

In the training process we use CE as loss function in the form:

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

where W_{yi} is the weight associated to class y of i-th instance, x_i are the deep feature of i-th instance and b is the bias term to class y of i-th instance.

Table 1: Experimental results

(a) CUB-200-2011

Method	Top-1
	Prior Work
Resnet-50	84.5
PN-DCN [1](BMVA 14)	85.4
DT-RAM [19](ICCV 17)	86.0
MC-Loss [3](Trans. Img Proc. 20)	87.3
MaxEnt [10](NeurIPS 18)	86.5
MA-CNN [39](ICCV 17)	86.5
KERL [6](IJCAI 18)	87.0
AP-CNN 1 st. [9](Trans. Img Proc. 21)	87.2
NTS-Net [38](ICCV 18)	87.5
DBTNet-50 [40](NeurIPS 19)	87.5
Cross-X [22](ICCV 19)	87.7
TASN [42](CVPR 19)	87.9
HSE [5](ACM-MM 18)	88.1
DBTNet-101 [40](NeurIPS 19)	88.1
CDL [36](ACM-MM 19)	88.4
AP-CNN 2 st. [9](Trans. Img Proc. 21)	88.4
Elope [13](WACV 20)	88.5
API-Net [43](AAAI 20)	88.6
	Our Results
EnGraf-Net50 (G=4, H=1)	87.94
EnGraf-Net101 (G=4, H=1)	88.00
EnGraf-Net152 (G=4, H=1)	88.31

(b) FGVC-Aircraft

Method	Top-1
	Prior Work
Kernel-Act [2](ICCV 17)	88.3
MaxEnt [10](NeurIPS 18)	89.8
MA-CNN [39](ICCV 17)	89.9
PA-CNN [41](Trans. Img Proc. 19)	91.0
DBTNet-50 [40](NeurIPS 19)	91.2
NTS-Net [38](ICCV 18)	91.4
iSQRT-COV [18](CVPR 18)	91.4
DBTNet-101 [40](NeurIPS 19)	91.6
DFL-CNN [35](CVPR 18)	92.0
SEF [23](IEEE Sign. Proc. Lett. 20)	92.1
AP-CNN 1 st [9](Trans. Img Proc. 21)	92.2
Cross-X [22](ICCV 19)	92.7
S3Ns [8](ICCV 19)	92.8
MC-Loss [3](Trans. Img Proc. 20)	92.9
EfficientNet-B7 [31](ICML 19)	92.9
API-Net [43](AAAI 20)	93.4
Elope [13](WACV 20)	93.5
AP-CNN 2 st [9](Trans. Img Proc. 21)	94.1
	Our Results
EnGraf-Net50 (G=4, H=1)	92.14
EnGraf-Net101 (G=4, H=1)	93.34

(c) Hierarchy classification

	CUB	AIR
Method	acc coarse-fine	acc coarse-fine
EnGraf-Net50	92.32-87.94	95.44-92.14
EnGraf-Net101	92.70-88.00	96.10-93.34

(d) Cifar-100

Method	top-1
Resnet-18	72.43
Two-Branch	72.95
Graft	73.85
EnGraf-net18 (G=2, H=1)	75.52
EnGraf-net18 (G=3, H=1)	75.13
EnGraf-net18 (G=4, H=1)	**75.85**
EnGraf-net18 (G=5, H=1)	75.41

(e) Cifar-100

Method	top-1	Ours	top-1
Resnet-18	72.43	EnGraf-net18	75.85
Resnet-50	75.42	EnGraf-net50	77.27
Resnet-101	75.49	EnGraf-net101	77.13

Considering the network proposed (see Fig. 2) we compute multiple CE loss in a different part of our proposal $(FC_0, FC_1, FC_2, FC_3, FC_4)$ where each of them jointly with supervised signals is used in the learning process with Stochastic Gradient Descendent method to achieve the global minima (or a good approximation of it). To summarize our total loss function, we use the following formulation:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{xent}(FC_0, y^K) + \mathcal{L}_{xent}(FC_1, y^K) + \\ & \mathcal{L}_{xent}(FC_2, y^{K-1}) + \mathcal{L}_{xent}(FC_3, y^K) + \mathcal{L}_{xent}(FC_4, y^{K-1})\end{aligned} \tag{2}$$

The cardinality of the classes y considered in *EnGraf-Net* is different in FC_2, FC_4 than FC_0, FC_1, FC_3 due to the supervised signals selected (it depends on the datasets and by how many hierarchy annotations we consider).

3 Experiments

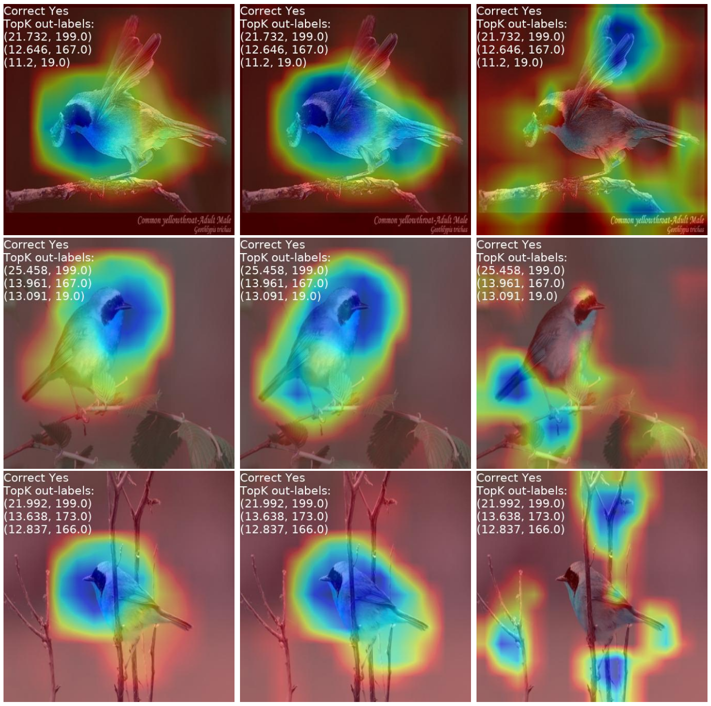
We conduct experiments on three well-known datasets: Cifar-100, CUB-200-2011 and FGVC-Aircraft and we investigate our performance model using the Resnet family comparing our proposal with the relative baselines and with some most recent architectures proposed in the literature (Table (1a) and Table (1b)). We conduct an ablation study on Resnet-18 using different type of *graft* and with some variations of it (Table (1d) and Table (1e)). We use Cifar100 dataset as a toy dataset to analyze the behaviour of our proposal. It contains 50,000 images 32×32 of training and 10,000 test images, labelled over 100 fine-grained classes. We use 20 coarse-grained classes as y^{K-1} semantic association in our hierarchical extraction. All other experiments have been performed on challenging Fine-Grained image classification benchmark datasets. **CUB-200-2011** [32] contains 11788 images of 200 species of birds split in 5994 and 5794 images for train and test respectively. In addition, we use 122 class labelled as *genera* of the species as supervised signals. **FGVC-Aircraft** [26] contains $10,000$ images of airplanes annotated with the model, specifically splitted in 6667 and 3333 for train and test set. This dataset is organised in four-level hierarchy. In addition to 100 classes (fine-labels) we use 70 classes (family) as superclass labelled. In all our experiments we use different pre-preprocessing data (see our code [16]). We report the upper-bound computational time of 19:43h in CUB-200-2011 over 150 epochs using a learning rate optimizer (SGD in all our experiments) of 0.001 and batch-size 20 using an EnGraf-Net152.

3.1 Results

In Table (1a) and Table (1b) we report the comparison results between the proposed model and other existing models on the two widely used fine-grained classification benchmark. We measure the top-1 accuracy in each experiment demonstrating the improvements across both datasets used. We obtained 88.31% and 93.34% respectively on CUB-200-2011 and FGVC-Aircraft datasets overcoming the best models used of fine-grained task and being very competitive with the most recent algorithms designed for this specific task (e.g API-Net). We investigated on Cifar-100 dataset analyzing the performance of our proposal using different graft (Table (1d)). In this last we achieve the best performance using a $graft = 4$ than the other type of graft applied in our experiments and overcome the performance than the baseline Resnet-18, two-branch or using only a branch with a simple graft. Starting to this assumption, we applied a $graft = 4$ for all experiment reported in Tables 1(a,b,c,e). In Table (1c) we report the accuracy from coarse and fine classes demonstrating the capability of our model to solve both tasks in a unique model end-to-end.

3.2 Visualization Analysis

In literature many techniques to visualize the class activation map has been proposed [30, 34]. Gradient-weighted Class Activation mapping (Grad-CAM) is widely used because it can be applied in the pretrained models highlighting the discriminative regions of the images. These approaches are useful to analyze the behaviour of the main model and



(a) fine-branch (b) coarse-branch (c) graft-branch

Fig. 3: Visualization of the attentions regions captured by EnGraf-Net50 in 3 types of layers (columns) and 3 different images (rows) of CUB. Using semantic association of the taxonomy, our method has the capability to detect subtle differences and spatial discriminative information without using part annotations. The third column show the effectiveness to focus the attention to other regions usually not considered in fine-grained models.

make it more transparent and understandable. In Fig. 3, we use Grad-CAM approach to visualize the attention map build by our model applying it on three different types of layers of our model. We emphasize that the three features obtained from these layers are combined using a concatenation function and feed into a fully connected layer where we make the final classification. In Fig. 3, we show the activation map build by the branch guided by the supervised signal that represents the class label (1st column), the branch guided by the supervised signal that represents the y^{K-1} class (2nd column) and at last, the branch responsible to make the graft using both supervised signals (3rd column). The discriminative regions usually considered in a fine-grained model belong in the object (1st column), however we force the model to find other discriminative regions from different area of images (3rd column) as the environment information or other useful details. It is extremely curious to observe the different highlighted regions from the graft branch (3rd column) than the others. The exploration of new discriminative regions (spatial information) using our approach is detected and combined with the regions from the others branch to increase the performance of the baselines without using any annotations (e.g bounding box, location parts).

4 Conclusions

In this paper, we simulate the pattern separation/completion process follow the behaviour of the hippocampus brain circuit. We explore a way to fine-grained classification using only semantic association without the requirement to use bounding-box/part annotations or sophisticated cropping techniques. We conduct experiments along the Resnet models demonstrating that our proposed model can easily be integrated into recent convolutional neural networks. Experiments in CUB-200-2011, FGVC-Aircraft and Cifar-100 have demonstrated the effectiveness of proposed model across many models designed for fine-grained task overcoming the performance of them and to be competitive with the most recent models.

References

1. Branson, S., et al.: Bird species categorization using pose normalized deep convolutional nets (2014)
2. Cai, S., et al.: Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: ICCV (2017)
3. Chang, D., et al.: The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Trans Image Process (2020)
4. Chang, D., et al.: Your" labrador" is my" dog": Fine-grained, or not. arXiv preprint arXiv:2011.09040 (2020)
5. Chen, T., et al.: Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: ACM-MM (2018)
6. Chen, T., et al.: Knowledge-embedded representation learning for fine-grained image recognition. In: IJCAI (2018)
7. Deshmukh, S.S., Knierim, J.J.: Representation of non-spatial and spatial information in the lateral entorhinal cortex. Frontiers in behavioral neuroscience (2011)
8. Ding, Y., et al.: Selective sparse sampling for fine-grained image recognition. In: ICCV (2019)
9. Ding, Y., et al.: Ap-cnn: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Trans Image Process (2021)
10. Dubey, A., et al.: Maximum-entropy fine grained classification. In: NIPS (2018)
11. Fu, J., et al.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR (2017)
12. Ghiasi, G., et al.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: CVPR (2019)
13. Hanselmann, H., et al.: Elope: Fine-grained visual classification with efficient localization, pooling and embedding. In: WACV (2020)
14. Hanselmann, H., et al.: Fine-grained visual classification with efficient end-to-end localization. arXiv (2020)
15. Jaderberg, M., et al.: Spatial transformer networks. arXiv preprint arXiv:1506.02025 (2015)

16. La Grassa, R., Gallo, I., Landro, N.: Engraf-net: Multiple granularity branch networkwith fine-coarse graft grained for classification task. https://gitlab.com/artelabsuper/engraf-net
17. La Grassa, R., et al.: Learn class hierarchy using convolutional neural networks. Applied Intelligence (2021)
18. Li, P., et al.: Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: CVPR (2018)
19. Li, Z., et al.: Dynamic computational time for visual attention. In: ICCV (2017)
20. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: CVPR (2017)
21. Liu, S., et al.: Path aggregation network for instance segmentation. In: CVPR (2018)
22. Luo, W., et al.: Cross-x learning for fine-grained visual categorization. In: ICCV (2019)
23. Luo, W., et al.: Learning semantically enhanced feature for fine-grained image classification. IEEE Signal Processing Letters (2020)
24. MacArthur, R.H.: Population ecology of some warblers of northeastern coniferous forests. Ecology (1958)
25. Madar, A.D., et al.: Pattern separation of spiketrains in hippocampal neurons. Sci.rep (2019)
26. Maji, S., et al.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
27. Neunuebel, J.P., et al.: Conflicts between local and global spatial frameworks dissociate neural representations of the lateral and medial entorhinal cortex. Neuroscience (2013)
28. Neunuebel, J.P., et al.: Ca3 retrieves coherent representations from degraded input: direct evidence for ca3 pattern completion and dentate gyrus pattern separation. Neuron (2014)
29. Newman, E.L., et al.: Ca3 sees the big picture while dentate gyrus splits hairs. Neuron (2014)
30. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
31. Tan, M., et al.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
32. Wah, C., et al.: The Caltech-UCSD Birds-200-2011 Dataset (2011)
33. Wang, D., et al.: Multiple granularity descriptors for fine-grained categorization. In: ICCV (2015)
34. Wang, H., et al.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: CVPR (2020)
35. Wang, Y., et al.: Learning a discriminative filter bank within a cnn for fine-grained recognition. In: CVPR (2018)
36. Wang, Z., et al.: Weakly supervised fine-grained image classification via correlation-guided discriminative learning. In: ACM-MM (2019)
37. Xie, L., et al.: Hierarchical part matching for fine-grained visual categorization. In: ICCV (2013)
38. Yang, Z., et al.: Learning to navigate for fine-grained classification. In: ECCV (2018)
39. Zheng, H., et al.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: ICCV (2017)
40. Zheng, H., et al.: Learning deep bilinear transformation for fine-grained image representation. In: NIPS (2019)
41. Zheng, H., et al.: Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Trans Image Process (2019)
42. Zheng, H., et al.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: CVPR (2019)
43. Zhuang, P., et al.: Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)