

Hierarchical CNN for Traffic Sign Recognition

Xuehong Mao, Samer Hijazi, Raúl Casas, Piyush Kaul, Rishi Kumar,
and Chris Rowen, *Fellow, IEEE*

Abstract— The Convolutional Neural Network (CNN) is a breakthrough technique in object classification and pattern recognition. It has enabled computers to achieve performance superior to humans in specialized image recognition tasks. Prior art CNNs learn object features by stacking multiple convolutional/non-linear layers in sequence on top of a classifier. In this work, we propose a Hierarchical CNN (HCNN) which is inspired by a coarse-to-fine human learning methodology. For a given dataset, we introduce a CNN-oriented clustering algorithm to separate classes into K subsets, which are referred to as families. Then, the HCNN algorithm trains $K+1$ classification CNNs: one CNN for family classification and K dedicated CNNs corresponding to each family for member classification. We evaluate this HCNN approach on the German Traffic Sign Recognition Benchmark (GTSRB), and achieve 99.67% correct detection rate (CDR), which is superior to the best reported results (99.46%) achieved by a single network.

I. INTRODUCTION

The Convolutional Neural Network (CNN) is a deep learning approach that stacks several convolutional, subsampling and non-linear activation layers in sequence. Recently, the CNN has become a breakthrough technique in the field of artificial intelligence for object classification and pattern recognition applications such as handwritten digit recognition [1], speech recognition [2], object classification [3], and face identification [4], [5]. Meanwhile, Advanced Driver Assistant Systems (ADAS) have received tremendous interest from both industry and academia. To improve driving safety, a vehicle needs to be able to see and understand its surroundings. Reliable detection and classification of objects such as pedestrians, vehicles, roads and traffic signs are highly demanded in ADAS. This work proposes a novel CNN algorithm and applies it to the problem of traffic sign recognition (TSR).

The conventional multiclass CNN classification approach [1]-[5] treats all classes the same way: the output layer

generates a signal for each class and the strongest signal determines the class of the input object. With this “N-way” classifier structure some classes are naturally more likely to be misclassified than others. For example, there are six subsets of traffic signs defined in Table 1 for the German Traffic Sign Recognition Benchmark (GTSRB) [6]. There is a strong likelihood of confusion between 30kmph and 80kmph speed limit signs since digits “3” and “8” are quite similar to each other. On the other hand, it is more difficult to miss-classify a priority road sign (the first of the *unique signs* on the left of row (f) in Table 1) because it looks quite different from all other signs in Table 1. The observation suggests a hierarchical approach to categorize objects into subsets, then to identify members in each set. This is a natural human “coarse-to-fine” learning approach that can be directly applied to machine learning. It has two advantages. First, the hierarchical architecture enables more optimized resource allocation: the size of the network can be tailored to the size of the classification problem. Second, a smaller network with less classes to be identified is easier to learn than a larger network. Label tree model has been recognized as an efficient approach to learn the hierarchical structure of large multi-class objects [7], [8]. A hierarchical CNN architecture has been proposed in [9], which obtains a confusion matrix from a randomly sampled held-out set and then applies a spectral clustering algorithm to partition the classes into coarse categories.












































In this paper, we propose a Hierarchical-CNN (HCNN) architecture for the GTSRB. Instead of using the human predefined subsets, we develop a CNN-oriented approach which clusters the GTSRB signs into new subsets, or families. The motivation of re-grouping signs by CNN is that the sign clustering is also based on CNNs, and the new clustering is demonstrated to improve the correct classification rate.

The rest of the paper is organized as follows: Section II is a brief introduction to the GTSRB, the dataset we used in our



Figure 1 Samples of the GTSRB dataset

Table 1 Pre-defined subsets in the GTSRB

(a) Speed Limit Signs	       
(b) Other Prohibitory Signs	   
(c) Derestriction Signs	   
(d) Mandatory Signs	       
(e) Danger Signs	              
(f) Unique Signs	   

work and previous work related to it. Details of the proposed algorithm are explained in Section III, while experimental results that validate the proposed approach are presented in Section IV. Section V provides concluding remarks.

II. GTSRB AND RELATED WORK

The GTSRB is a challenging image classification problem because real world scene variations such as illumination, weather conditions and partial occlusions impact the visual appearance of the signs. The GTSRB is provided by the Institut für Neuroinformatik group and was published for a competition held in IJCNN 2011. Among various solutions proposed for the TSR problem, the CNN approach seems to be the most promising method to compete against the best human performance [6]. The GTSRB dataset was collected by capturing videos from a moving car at different velocities and weather conditions. Still images of each traffic sign instance were extracted and annotated. As a result, the GTSRB dataset consists of images from 43 classes. Sample images are shown in Figure 1. The sizes of the traffic signs vary between 15×15 and 222×192 pixels. The full GTSRB dataset was randomly split into three mutually exclusive subsets for training, validation and test. The partitions of the three sets are 50%, 25% and 25%. The training set is used by the designed classification algorithm to train the parameters. The test set is shuffled such that no temporal information is available. The validation set is a subset of the database which is reserved for optimization of learning parameters such as the learning rate.

The results of the GTSRB competition are summarized in [6]. Among the algorithms that were evaluated with the GTSRB dataset, the multi-column deep neural network (Multicolumn-DNN) [10] proposed by Ciresan et al. and the Multiscale-CNN (MSCNN) approach proposed by Sermanet and Lecun [11] show better performance than other machine learning algorithms. In [12], higher recognition rate than the best result reported in [6] has been reported. However it ensembles 20 CNNs for training and testing, hence the complexity is prohibitive. Due to the lower complexity of MSCNN comparing to the approach in [10] and [12], we

adopt the MSCNN as a prototype for our basic CNN architecture. Unlike a traditional CNN which only takes the last stage output for classification, the MSCNN combines the outputs from previous stages with the last stage. Hence, the MSCNN enables the classifier to see both local and global features.

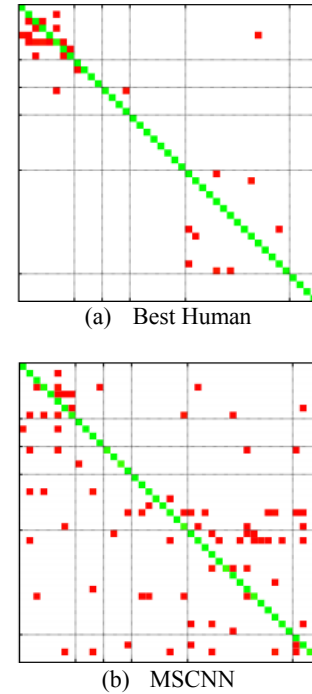


Figure 2 Confusion matrix of best human performance and Sermanet MSCNN algorithm [11]

Figure 2 shows an interesting observation from the GTSRB evaluation results for the best human performance and the MSCNN [6]. This figure was reported in [6] to visualize the confusion matrix that represents distribution of error over different classes. The rows and columns denote the ground truth and detected classes respectively. The classes are

ordered by the subsets in Table 1 (a) to (f) from left-to-right, top-to-bottom, and grid lines separate the subsets. The location of red markers (i, j) on the confusion matrix indicate that class i is misclassified as class j . As shown in Figure 2, the best human performance seldom misclassifies between subsets, whereas the MSCNN classifier suffers from the confusion among subsets. For example, the stop sign has been confused with six different signs across subsets (a), (b), (d) and (e).

As discussed above, the N-way flat CNN classifiers are not able to utilize the hierarchical structure in the classes. In this paper, we focus on this problem and develop a two-level HCNN for traffic sign recognition to improve the recognition rate. The proposed HCNN designs a hierarchical network which is able to learn the hierarchical structure in classes. Furthermore, we also suggest a family clustering algorithm that benefits the hierarchical classification CNNs.

III. METHOD

The proposed hierarchical CNN shown in Figure 2 consists of six CNNs: one family classifier CNN (FC-CNN) and five member classifiers CNNs (MC-CNN). Please note that we will justify why only 5 member classifiers are necessary later. The objective of the FC-CNN is to cluster the $C = 43$ traffic signs of GTSRB into families. The similarity between any members of the same family should be higher than the similarity between members from different families. FC-CNN annotates an image with a family ID $\in \{1, \dots, K\}$. According to the family index k , the MC-CNN k is invoked, which produces in a member ID $mid \in \{1, \dots, M^{(k)}\}$. Therefore, the

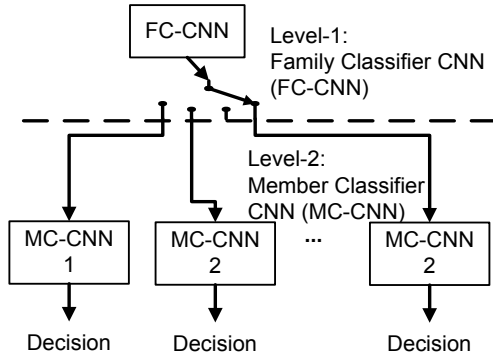


Figure 3 Flowchart of Hierarchical CNN

class ID of a sign is determined by fid and mid .

Although the hierarchical annotation is natural and used in some applications, such as [3], other applications and databases do not have explicit family classes. Furthermore, in applications such as GTSRB, the predefined subsets are not necessarily the best hierarchical categories for the next level CNN member classifiers. Therefore, we propose a CNN-oriented solution to cluster classes into families instead of using the predefined subsets. The reasons of using CNN-oriented family clustering algorithm to partition the classes are 1) the features of the CNN-oriented families are preferable to the FC-CNN such that the classification errors at the top level is minimized, and 2) the members within each family

are selected by CNN such that the MC-CNN at the next level is unlikely to be distracted by the uncorrelated features. The advantage of the proposed CNN-oriented family is demonstrated by the experiments in Section 4.

A. Similarity Matrix

In order to cluster members into families, we need a measure of the similarity between classes. In this paper, we propose the following algorithm to obtain the similarity metric λ_{ij} ($1 \leq i, j \leq C$) directly from the training set. Note that there are other approaches to obtain a similarity metric. For instance, in [12], a similarity metric is obtained by evaluating the network on a held-out subset from the training set. The major advantage of our method is that it can be obtained directly from the training set without a network that is pre-trained and evaluated.

Algorithm 1. Similarity Metric Generation

Input: training dataset

Output: λ_{ij} , $1 \leq i, j \leq C$

Begin

1. Randomly pick L samples for class $i \in \{1, \dots, C\}$ from training set and denote these samples by $P_i^{(1)}, \dots, P_i^{(L)}$, $i = 1, \dots, C$. Apply the same pre-processing to the sample images, such as global and local normalization, image resizing. Let denote results by $\tilde{P}_i^{(l)}$, where $i = 1, \dots, C$, $l = 1, \dots, L$.
2. Transfer the images to frequency domain by 2-D FFT, i.e., $Q_i^{(l)} = \text{FFT2D}(\tilde{P}_i^{(l)})$.
3. The similarity metric between class i and j is
$$\lambda_{ij} = \max_{l=1, \dots, L} \sum_{m,n} \text{real}\{Q_i^{(k)} \circ \text{conj}(Q_j^{(l)})\}_{m,n} \quad (1)$$

where \circ denotes the Hadamard matrix product.

End

B. CNN-oriented Family Clustering (CFC)

A two-fold CNN-oriented Family Clustering (CFC) algorithm is used to partition the GTSRB traffic signs into K families according to the similarity between signs. Figure 4 shows the

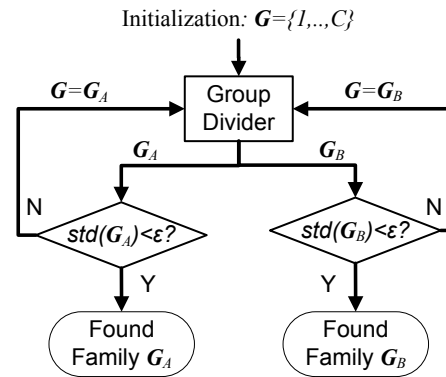


Figure 4 Flow chart of CNN-oriented Family Clustering (CFC) Algorithm

flowchart of the CFC algorithm, and details of the Group Divider is explained in Algorithm 2. The CFC is initialized with the group that contains all classes, i.e., $\mathbf{G} = \{1, \dots, C\}$. Then, the group divider separates \mathbf{G} into two groups: \mathbf{G}_A and \mathbf{G}_B . To determine whether to stop or keep dividing a group, the standard deviation of the similarity metrics of the group is calculated and compared to a predetermined threshold. A group is determined to be a family when the similarity standard deviation is less than the threshold. Otherwise, the group is further divided by the Group Divider. The family clustering algorithm runs recursively until all groups satisfy the termination conditions.

Algorithm 2. Group Divider Algorithm

Input: Group \mathbf{G} , similarity metric Λ

Output: Group \mathbf{G}_A and \mathbf{G}_B

Begin

1. Find seed sets \mathbf{S}_A and \mathbf{S}_B
 - 1.1 Initialize seed set \mathbf{S}_A and \mathbf{S}_B such that they are least correlated, i.e., $(a, b) = \underset{i, j \in \mathbf{G}, i \neq j}{\operatorname{argmin}} \lambda_{ij}$.
 - 1.2 Remove members in \mathbf{S}_A and \mathbf{S}_B from \mathbf{G} , i.e., $\mathbf{G} := \mathbf{G} \setminus (\mathbf{S}_A \cup \mathbf{S}_B)$.
 - 1.3 For $i \in \mathbf{G}$, if $\operatorname{std}(\mathbf{S}_x \cup \{i\}) < \epsilon$ then $\mathbf{S}_x := \mathbf{S}_x \cup \{i\}$,
where $\mathbf{x} = \begin{cases} \mathbf{A} & , \lambda_{iA} > \lambda_{iB} \\ \mathbf{B} & , \text{else} \end{cases}$.
 - 1.4 Update set: $\mathbf{G} := \mathbf{G} \setminus \{i\}$.
 - 1.5 Repeat Step 1.3 and 1.4 until no new members can be brought into either \mathbf{S}_A or \mathbf{S}_B .
2. Prepare sub-training set by selecting images belonging to $\mathbf{S}_A \cup \mathbf{S}_B$, and train a CNN to discriminate a sign between \mathbf{A} and \mathbf{B} .
3. Run CNN obtained in Step 2 on the original GTSRB training set to label all images by $J \in \{\mathbf{A}, \mathbf{B}\}$.
4. Post CNN processing: according to the labels J , obtain the probability distribution of class i : \mathbf{p}_i^A and \mathbf{p}_i^B . If $\mathbf{p}_i^A > \mathbf{p}_0$, include class i in \mathbf{G}_A ; otherwise, include class i in \mathbf{G}_B .

End

The operation $\operatorname{std}(\cdot)$ in step 1.3 stands for the standard deviation calculation. Threshold ϵ in Algorithm 2 controls how close the members in the seed sets are. In addition, threshold \mathbf{p}_0 manages the distance between families. The algorithm above groups the 43 GTSRB traffic signs into the five families in Table 2 that agree with the pre-defined substitutes shown in Table 1 with the following exceptions:

- The CNN-oriented Family clustering generates five families while there are six subsets shown in Table 1.
- Family #1 is determined after the first run of the group divider described in Algorithm 2. The standard deviation of similarity remains below the threshold even though a unique sign (the sign with a yellow rhombus in (f) of Table 1) has been included in this family.

Table 2 CNN-oriented Families for the GTSRB

Family 1	
Family 2	
Family 3	
Family 4	
Family 5	

- Members of unique signs shown in row (f) of Table 1 have been scattered into Families #1, #2 and #4 according to their similarity metric.
- Surprisingly, one of the blue signs is grouped together with the Derestriction signs (gray signs) in Family #1. However, the CNN perceives the diagonal pattern more than the color.

IV. EXPERIMENT RESULTS

We evaluate the proposed HCNN algorithm on the GTSRB database. The input images are resized to 32×32 pixels in R, G and B channels. To expand the existing GTSRB training set, we add jittered copies as described in [11]. Instead of random position, scale and rotation perturbation, we add saturation scaling before the global and local contrast normalization. For each instance in the training set, we create 5 extra jittered copies, thus the jittered training dataset is 6 times larger.

Table 3 Network Architecture of HCNN

Layer	Kernel in FC-CNN	Size-out in FC-CNN	Kernel in MC-CNN	Size-out in MC-CNN
Conv1	5×5×3 (40)	28×28×40	5×5×3 (100)	28×28×100
Pool1	2×2, max	14×14×40	2×2, max	14×14×100
ReLU1	NA	14×14×40	NA	14×14×100
Conv2	5×5×40 (70)	10×10×70	5×5×100 (200)	10×10×200
Pool2	2×2, max	5×5×70	2×2, max	5×5×200
ReLU2	NA	5×5×70	NA	5×5×200
MS2 Conv	5×5×70 (40)	1×1×40	5×5×200 (100)	1×1×100
MS2, Atan	NA	1×1×40	NA	1×1×100
MS2, ReLU	NA	1×1×40	NA	1×1×100
MS1 Pool	2×2, max	7×7×40	2×2, max	7×7×100
MS1 Conv	7×7×40 (40)	1×1×40	7×7×100 (100)	1×1×100
MS1 ReLU	NA	1×1×40	NA	1×1×100
Full Connection	1×1×80 (4)	1×1×4	1×1×200 (M ^(k))	1×1×M ^(k)

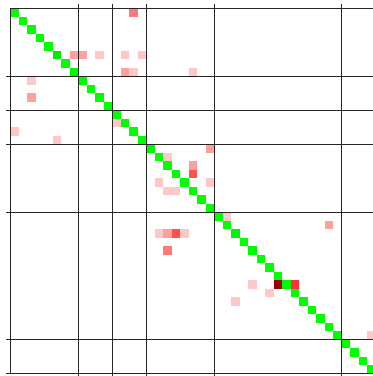
The network architectures for both FC-CNN and MC-CNN in Figure 3 are summarized in Table 3. The two types of CNN share the same network structure except for the number of neurons (the number shown in the parenthesis).

Table 4 CDR comparison to GTSRB Benchmark

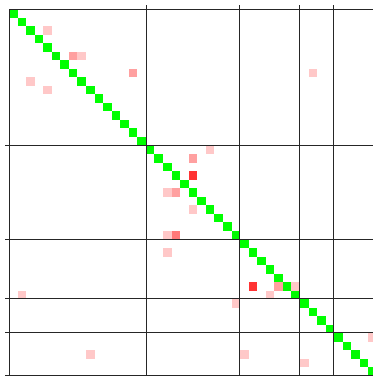
CDR%	Team	Method
99.67	Cadence	HCNN
99.46	IDSIA	Committee of CNNs
99.22	INI-RTCV	Human (best individual)
99.17	Sermanet (LeCun)	Updated multi-scale CNN
98.84	INI-RTCV	Human (average)
98.31	Sermanet (LeCun)	Multi-scale CNN
96.14	CAOR	Random Forest
95.68	INI-RTCV	LDA (HOG 2)

Table 4 compares the CDR of the proposed HCNN traffic sign recognition algorithm to the benchmark of the GTSRB. The proposed HCNN achieves the highest CDR among all the evaluated methods with the GTSRB database.

To further demonstrate the advantage of the proposed HCNN algorithm, Figure 5 compares the confusion matrix of the HCNN with different subsets/families. The confusion matrices are shown in the same manner as in **Error!**



(a) HCNN with Predefined Subsets



(b) HCNN with CNN-oriented Families

Figure 5 Confusion Matrix Comparison

Reference source not found.2, while we use color to indicate the occurrence of miss-classification the occurrence of miss-classification. The darker the red marker is, the more errors appear. The comparison of miss-classification is provided in Table 5. From Figure 5(a) to Figure 2(b), it is obvious that the proposed HCNN algorithm reduces the number of miss-classifications across the pre-defined defined subsets significantly. In addition, from Figure 5(a) and (b) and Table 5 it can be observed that using the CNN-oriented families further reduces the occurrence of miss-classification across families/subsets. The reason for CNN-oriented families outperforming the human defined subsets is that the partitioning of the input space for CNN-oriented families is accomplished with the same network components (e.g., ReLU, max-pooling and convolutional filtering) and with similar network architecture of HCNN. For example, the features extracted by the FC-CNN are likely to be similar to the features extracted unsupervised CNN family clustering algorithm. As a result, we observed that the HCNN with CNN-oriented families achieves 99.67% CDR, while the HCNN using the pre-defined subsets only achieves 99.44% CDR. Note that the ensemble recognition rate 99.65% reported in [12] is at the cost of 80 times more complexity. The performance of our algorithm can be further improved if we use the same ensemble approach proposed in [12].

Table 5 Miss-Classification Comparison

	Pre-defined Subsets	CNN-oriented Families
Miss-classification across subsets/Families	0.23%	0.10%
Miss-classification within Subsets/Families	0.33%	0.24%

V. CONCLUSION

In this paper, we proposed a hierarchical CNN for traffic sign recognition, and demonstrated its superior performance to the best reported results with the GTSRB. The benefit of the proposed HCNN is two-fold. First, the hierarchical CNN is able to solve a difficult problem by partitioning it into multiple easier sub-problems and distributing the effort of solving these sub-problems according to their difficulty. In this paper, we have used a secondary member classifier that is of equal size and architecture for all families. Alternatively, the size and architecture of each member classifier could be tailored to the size of each corresponding family. The second benefit of our approach comes from the use of an unsupervised CNN to learn the hierarchy of classes because the HCNN perceives the CNN-oriented families better than the human defined families. As part of our future work, we will apply the HCNN to other applications and investigate recursive multi-level classification.

REFERENCES

- [1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L.D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541-551, 1989.

- [2] Abdel-Hamid O., Mohamed A., Hui Jiang, Penn G., "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, 2012.
- [3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in Neural Information Processing Systems, 2012.
- [4] Schroff F., Kalenichenko D., Philbin J., "Facenet: A unified embedding for face recognition and clustering," arXiv preprint arXiv:1503.03832, 2015.
- [5] Taigman Y., Ming Yang, Ranzato M., Wolf L., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2014.
- [6] Stallkamp J., Schlipsing M., Sakneb J. Igel C., "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," Neural networks, vol. 32, pp. 323-332, 2012.
- [7] S. Bengio, J. Weston, and D. Grangier, "Label Embedding Trees for Large Multi-Class Tasks", NIPS 2010, pp. 163-171
- [8] J. Deng, S. Satheesh, A.C.Berg, and F. Li, "Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognition", NIPS 2011, pp. 567-575.
- [9] Yan Z., Jagadeesh V., DeCoste D., Di W., Piramuthu R. , "HD-CNN: Hierarchical Deep Convolutional Neural Network for Image Classification," arXiv preprint arXiv:1410.0736, 2014.
- [10] Ciresan D., Meier U., Masci J., Schmidhuber J., "A committee of neural networks for traffic sign classification," in Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011.
- [11] Sermanet P., Lecun Y., "Traffic sign recognition with multi-scale Convolutional Networks," in Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011.
- [12] J. Jin, K. Fu, C. Zhang, "Traffic Sign Recognition With Hinge Loss Trained Convolutional Neural Networks", IEEE Trans on Intelligent Transportation systems, Vol. 15, No. 5, Oct. 2014