At the Crossroads: Traffic Signs Recognition From HOG to CNN

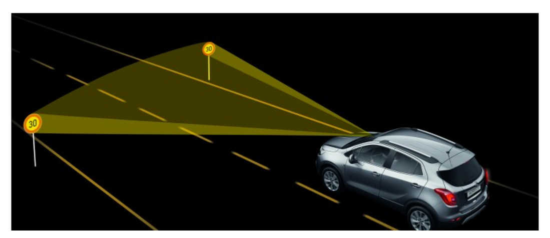
Yu Zhao *Zixiao Wang*

zhaoyu92@stanford.edu *zixiaow@stanford.edu*

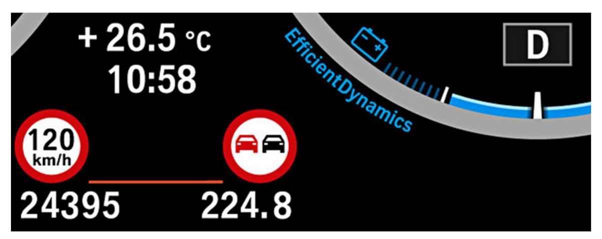
Stanford
CS221 Autumn 2019

Motivations and Problem Definition

- [NHTSA estimated that 36,750 people lost their lives in traffic accidents in 2018](#).
- Reducing human driving errors save life
- Traffic Sign Recognition could improve the safety and reliability of driverless cars and traditional vehicles
- Goals: build an accurate and robust "single-image, multi-class" traffic sign classifier
  - Input: an image containing a traffic sign
  - Output: correct class of that traffic sign
  - Using various machine learning or deep learning models and techniques



General Motors Traffic Sign Recognition Technology



BMW Traffic Sign Recognition for the all-new 1 Series

Dataset

Selective images of the 43 sign classes feature in *Stallkamp et al.* 2012 paper

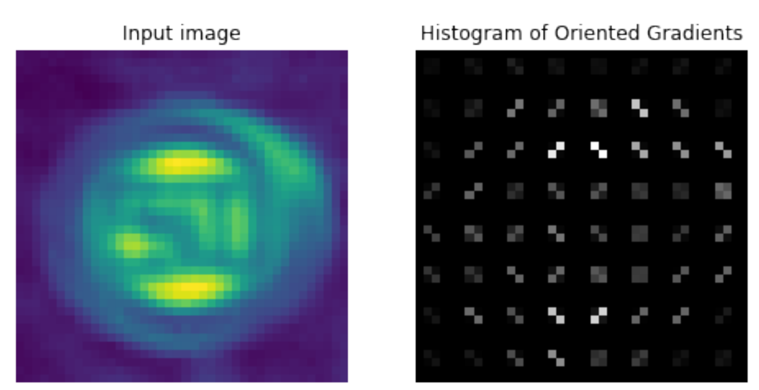
German Traffic Sign Recognition Benchmarks (GTSRB) dataset

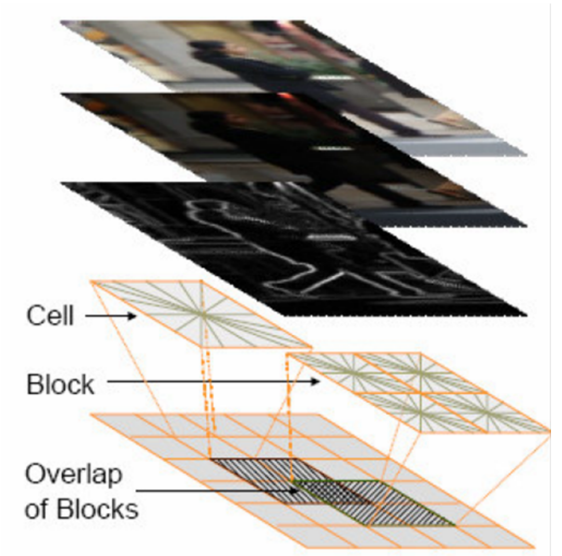
- 51,840 images of the 43 classes
  - 39,210 for training and validation (80:20 split)
  - 12,630 for testing
- 1,728 unique traffic sign occurrences, each occurrence with 30 images sampled from far away to close up
- Image sizes ranging from 15×15 to 222 × 193 pixels
- Distribution among 43 classes of traffic signs is very imbalanced, some has <=300 while some has >= 2000

Challenges

- Uneven distribution of traffic sign classes
- Traffic signs from far are low in resolution (15 x 15)
- Traffic signs up close are blurry due to motion
- Complex environment lighting
- Different orientations for the same class
- Best human accuracy as 99.22% (Stallkamp et. al)

Approaches - HOG Features





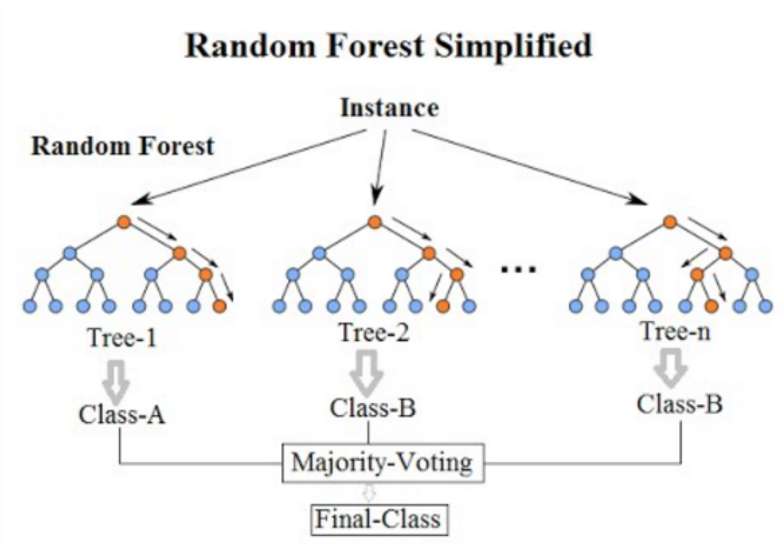
Top: An example of grayscale version of the image input and its HOG output
Right: Structure of HOG by Dalal et al. 2005

Histograms of Oriented Gradient (HOG) descriptors

- Proposed by Dalal and Triggs (2005) for pedestrian detection.
- Weighted and normalized histogram to represent gradient of color images.
- All input images scaled to 40 x 40.
- GTSRB provides training and testing images converted to HOG:
  - cell size 5×5 pixels
  - block size of 2×2 cells
  - an orientation resolution of 8
  - Total feature length 1568

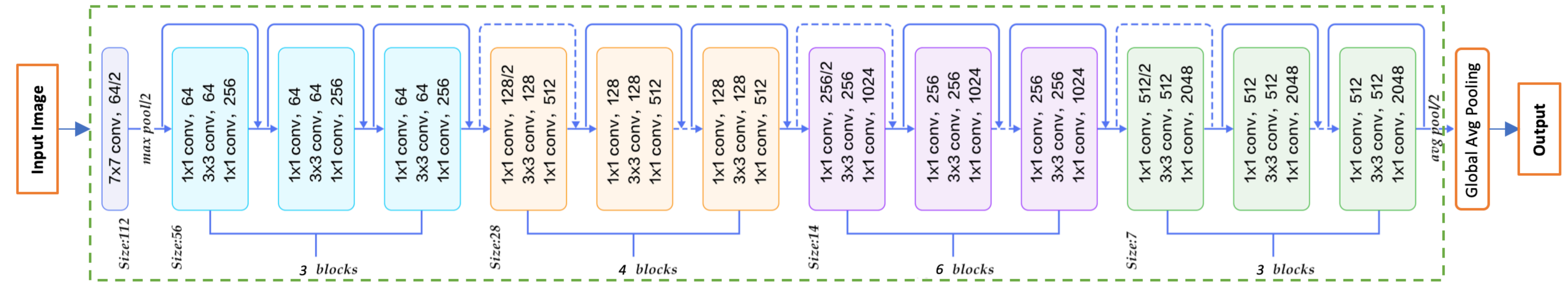
Classification Model - Random Forest Model

- Ensemble of decision trees
- Subsect of features when forming questions
- Random set of the training data points

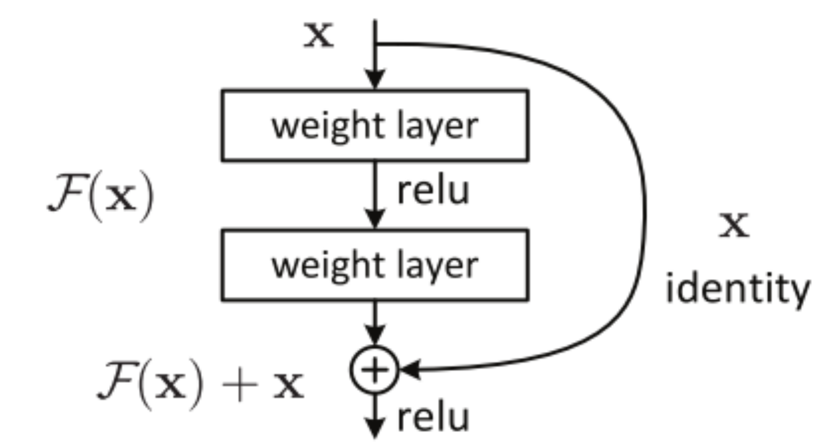


[An example of Random Forest Model](#)

Approaches – Deep Learning CNN



- **Model**: ResNet-50 (keras.applications.resnet50.ResNet50)
- **Input size**: Experimented with 32x32x3, 64x64x3, and 96x96x3 (best)
- **Loss Function**: Categorical Cross-Entropy
- **Optimizer**: Experimented Adam (lr = 0.001) or SGD (lr = 0.01, decay = 1e-6)
- **Initialization**: Model was initialized with weights pre-trained on image-net.
- **Data Augmentation**: Rotation (up to 20°), Zoom (up to 20% closer), brightness (between 20% darker and 20% brighter). Experimented with 1) balancing all categories, or 2) only augment to 1000 images per category
- **Image preprocessing**: 1) Central crop to ensure input images are square, 2) Resize all images to 96x96x3, 3) Adjust illumination by doing histogram normalization in V channel of HSV (hue, saturation, value) representation



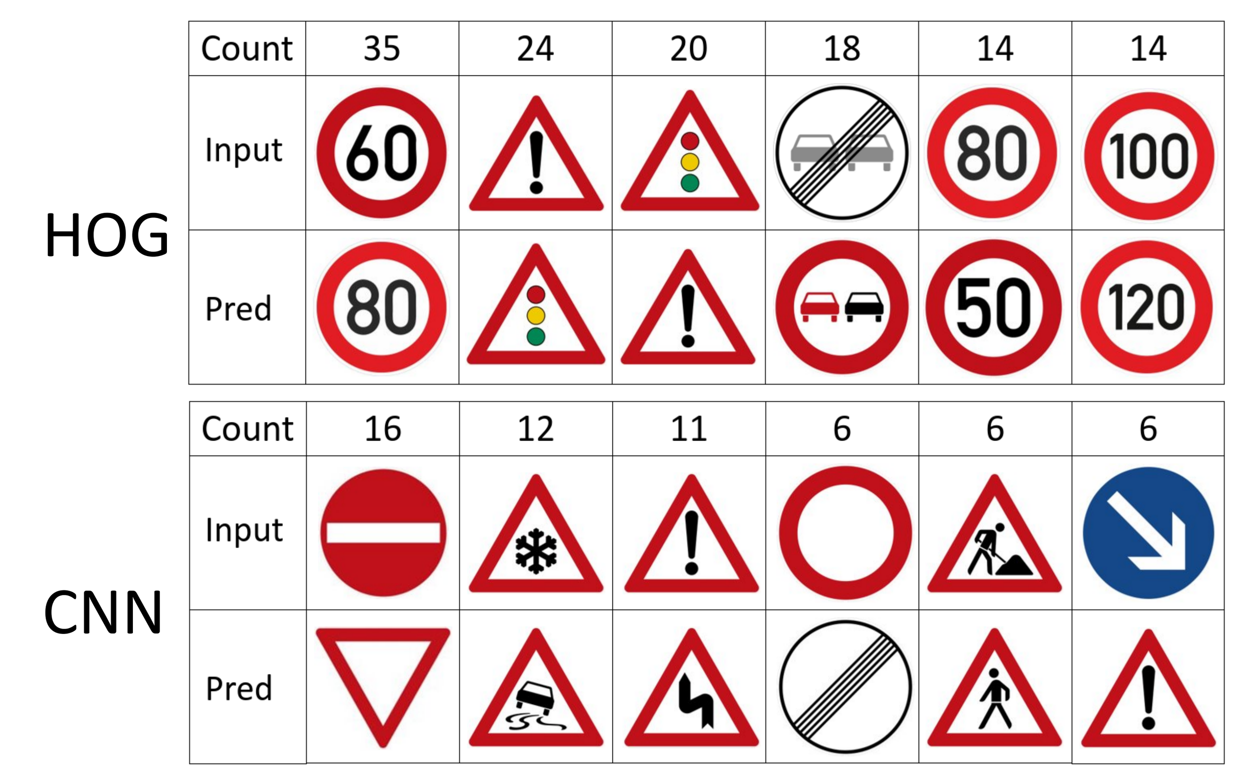
Advantage of ResNet-50: The identity mappings of the input are directly added to the outputs of the corresponding convolution layers, thereby "short-cutting" them if they are not helping the model, and therefore ensures that the accuracy of the deeper networks should be as good as its shallower counterpart.

Results

Input	Data Aug	Model	Accuracy
Raw	Full	Base CNN	0.8879
GTSRB HOG 2	None	Random Forest	0.9673*
GTSRB HOG 2	None	SVM	0.9579
Raw	None	ResNet-50	0.9813
Raw	None	ResNet-50 + SGD	0.9838
Raw	None	ResNet-50 + transfer	0.4451
Raw	Full	ResNet-50	0.9878**
Raw	Limited	ResNet-50	0.9853
Raw	Full	VGG-16	0.9610
Raw	None	DenseNet-121	0.9935***

* Best HOG and Random Forest outperformed model by Zaklouta et al.
** Best ResNet50 featured in error analysis.
*** Best DenseNet with preliminary results that outperformed the best human accuracy of 0.9922 reported by Stallkamp et al.

Analysis



References

- Pierre Sermanet and Yann LeCun. "Traffic sign recognition with multi-scale Convolutional Networks." In: IJCNN. 2011, pp. 2809–2813.
- Johannes Stallkamp et al. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: Neural networks 32 (2012), pp. 323–332.
- Johannes Stallkamp et al. "The German Traffic Sign Recognition Benchmark: A multi-class classification competition." In: IJCNN. Vol. 6. 2011, p. 7.
- Fatin Zaklouta, Bogdan Stanciulescu, and Omar Hamdoun. "Traffic sign classification using kd trees and random forests". In: The 2011 International Joint Conference on Neural Networks. IEEE. 2011, pp. 2151–2155.
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 886–893
- Zaklouta, F., Stanciulescu, B., and Hamdoun, O. (2011). Traffic sign classification using k-d trees and random forests. In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 2151–2155. IEEE Press.
- Kaiming He et al. "Deep residual learning for image recognition". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770–778.
- Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: arXiv preprint arXiv:1409.1556 (2014).
- G. Huang, Z. Liu, K. Q. Weinberger, and L. Maaten. Densely connected convolutional networks. In CVPR, 2017.