UNIVERSITY OF WATERLOO

TECHNOLOGY SUMMARY

A Novel, Lightweight Architecture for Deep Convolutional Neural Networks

Model	No. of Param.(MB)	Error Rates
Conv kernel	-	17.82%
Stochastic pooling	-	15.13%
ResNet	$1.7M$	13.63%
Maxout	$> 5M$	11.68%
NIN (9 conv layers)	$0.97M$	10.41%
Ours (3 conv layers)	$0.092M$	17.23%
Ours (6 conv layers)	$0.565M$	11.75%

Training error rates using CIFAR-10 images without data augmentation (Canadian Institute For Advanced Research- CIFAR)

Model	No. of Param.(MB)	Error Rates
ResNet	$1.7M$	44.74%
Stochastic pooling	-	42.51%
Maxout	$> 5M$	38.57%
Prob Maxout	$> 5M$	38.14%
NIN (9 conv layers)	$1M$	35.68%
Ours (3 conv layers)	$0.16M$	44.79%
Ours (6 conv layers)	$0.51M$	38.63%
Ours (6 conv layers)	$0.74M$	36.68%

Training error rates using CIFAR-100 images without data augmentation.

Compared to other techniques, UW solution requires one order of magnitude lesser training parameters while achieving similar error rates. This enables fast training of Machine Learning (ML) systems.

Reference

10135

Patent status

US Patent Pending

Stage of development

Prototype built and tested for video classification

Ongoing research

Contact

Scott Inwood
Director of Commercialization
Waterloo Commercialization Office
519-888-4567, ext. 33728
sinwood@uwaterloo.ca
uwaterloo.ca/research

Background

Convolutional Neural Networks (CNNs) are widely used in Machine Learning (ML) systems. While deep CNNs (utilizing many “neuron layers”) are very effective in extracting features from raw pixel values and have achieved amazing performance for classification and segmentation tasks in computer vision, they are suffering from problems affecting their accuracy and speed.

Deep CNNs are subject to a number of problems including loss of learned features between layers and the well-known “Vanishing Gradient Problem” which adds to the difficulty of training the ML system if not making it impossible.

Description of the invention

Deploying a novel architecture, researchers at the University of Waterloo have successfully designed a unique CNN that is capable of retaining rich information fused in its hidden layers thus, greatly alleviating the gradient-vanishing problem while avoiding the overfitting risk. Due to its unique construct, the new CNN’s activation function is capable of approximating very complex functions. The new CNN architecture along with its uniquely designed activation function has enabled the network to perform even better than the state-of-the-art while requiring much lesser parameters and “shallower structure” (lesser layers).

Advantages

Due to its unique architecture and activation function, the new deep CNN requires an order of magnitude lesser training parameters and can be implemented in a “shallower” structure. This means that the CNN is fast and requires much lesser computational power while delivering the same performance as the state-of-the-art.

Potential applications

Because the new CNN requires much lesser parameters and is light-weighted, it can be used in:

- Real time video/image classification
- Autonomous driving