Abstract

We propose cross-modal convolutional neural networks (X-CNNs), a novel biologically inspired type of CNN architectures, treating gradient descent-specialised CNNs as individual units of processing in a larger-scale network topology, while allowing for unconstrained information flow and/or weight sharing between analogous hidden layers of the network—thus generalising the already well-established concept of neural network ensembles. The constituent networks are individually designed to learn the output function on their own subset of the input data, after which cross-connections between them are introduced after each pooling operation to allow for information exchange between them. This injection of knowledge into a model is expected to yield greatest returns in sparse data environments, which are typically less suitable for training CNNs. For evaluation purposes, we have compared a standard four-layer CNN as well as a sophisticated FitNet4 architecture against their cross-modal counterparts the feature maps being passed:

Model construction

The network design process is initiated by appropriately partitioning the input data—this may be done either manually or through an unsupervised pre-training step. Afterwards, an X-CNN is constructed such that a separate CNN superlayer is dedicated to each partition of the input data, attempting to learn the target function from its partition only. The purpose of the partitioning is to help the constituent CNNs become powerful predictors while requiring a smaller dimensionality of the input data. This, in turn, allows for a reduction in parameter counts in these CNNs, requiring a smaller training set to train efficiently. Finally, the superlayers may be interconnected by any sort of (feedforward) cross-connection as is best seen fit. Here, after each pooling operation, we exchange the feature maps between the superlayers, after first passing them through an additional 1×1 convolutional layer. This construction is biologically inspired by cross-modal systems wherein several cross-connections between various sensory networks have been discovered [1, 2].

Evaluation

To quantify the gains of this approach, our evaluation focuses on an already well-understood problem of coloured image classification, on established CIFAR-10/100 [4] benchmarks for which an abundance of data is available, so it is easier to investigate the effects of restricting the size of the training set on various CNN models. We compare two models against their X-CNN variants: KerasNet a simple CNN with four convolutional ReLU layers, and the 17-layer FitNet4 by Romero et al. [5], representing a sophisticated CNN close to the state-of-the-art. Here we provide results for a variety of data availability scenarios (using only 5% of the training dataset for training) on CIFAR-100, with and without data augmentation. The metric we report is accuracy, as the classes are balanced in the test set.

Cross-connection analysis

We have investigated the mode of operation for cross-connections in two ways. Initially, we visualise the learned weights of the 1 × 1 convolutions in the first cross-connection layer, revealing a nontrivial linear combination of the feature maps being passed:

 conclusions

We conclude that cross-connections selectively filter features, learning to combine them (e.g., horizontal & vertical), and that their mode of operations mimics human vision (Y features having higher frequency than U/V).

Cross-modal Convolutional Neural Networks for Sparse Datasets

Petar Veličković, Duo Wang, Nicholas D. Lane and Pietro Liò

University College London, Nokia Bell Labs

{petar.velickovic, duo.wang, pietro.lio}@cl.cam.ac.uk, niclane@acm.org

References


