X-CNN: Cross-modal Convolutional Neural Networks for Sparse Datasets

Petar Veličković[1,3], Duo Wang[1], Nicholas D. Lane[2,3] and Pietro Liò[1]
[1]University of Cambridge, [2]University College London, [3]Nokia Bell Labs
{petar.velickovic, duo.wang, pietro.lio}@cl.cam.ac.uk, niclane@acm.org

UNIVERSITY OF CAMBRIDGE

UCL

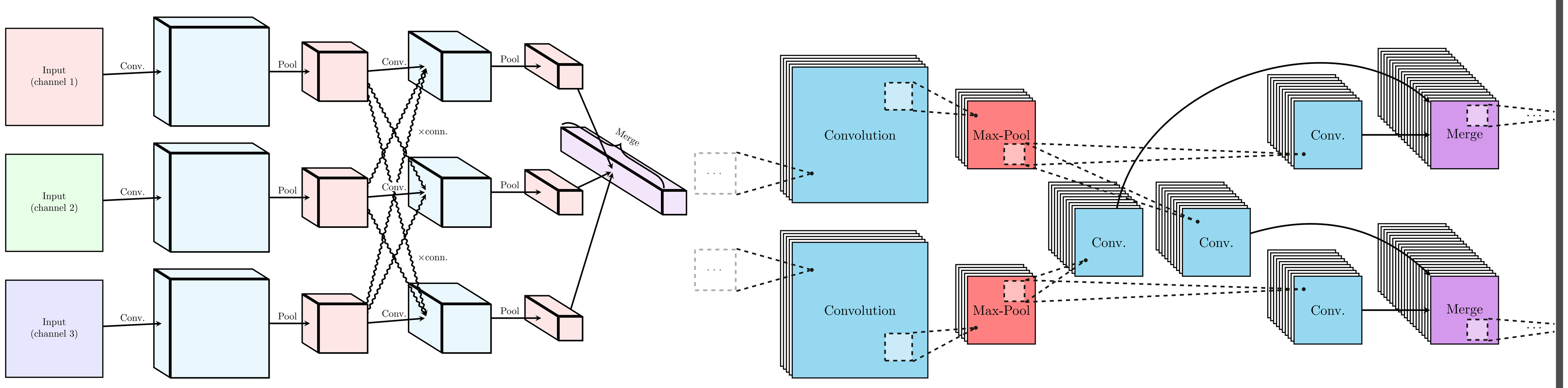
NOKIA Bell Labs

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 variants on the CIFAR-10 and CIFAR-100 datasets with differing percentages of the training data being removed, and find that at lower levels of data availability, the X-CNNs significantly outperform their baselines (typically providing a 2–6% benefit, depending on the dataset size and whether data augmentation is used), while still maintaining an edge on *all* of the full dataset tests.

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* within the visual and auditory systems of the human brain—wherein several cross-connections between various sensory networks have been discovered [1, 2].



(a) (b)

(a) Diagram of a simple cross-modal CNN for image classification, generated from a baseline CNN of the form $[Conv \rightarrow Pool] \times 2$.

(b) Illustration of a single cross-connection segment within an X-CNN with two superlayers.

Evaluation

To quantify the gains of this approach, our evaluation focusses 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 p% 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 dataset.

Comparative evaluation results on CIFAR-100 without data augmentation

Model \ p	1%	5%	10%	15%	20%	30%	40%	50%	∼	100%
KerasNet	7.55%	15.10%	20.24%	24.76%	28.18%	32.43%	36.29%	38.61%	…	48.26%
X-KerasNet	**8.05%**	**16.45%**	**23.04%**	**26.91%**	**30.08%**	**35.39%**	**39.13%**	**41.88%**	…	**49.98%**
FitNet4	6.48%	16.84%	22.12%	28.30%	35.52%	39.28%	43.59%	49.69%	…	59.78%
X-FitNet4	**6.64%**	**18.73%**	**27.57%**	**33.59%**	**38.38%**	**45.53%**	**49.68%**	**52.21%**	…	**62.20%**

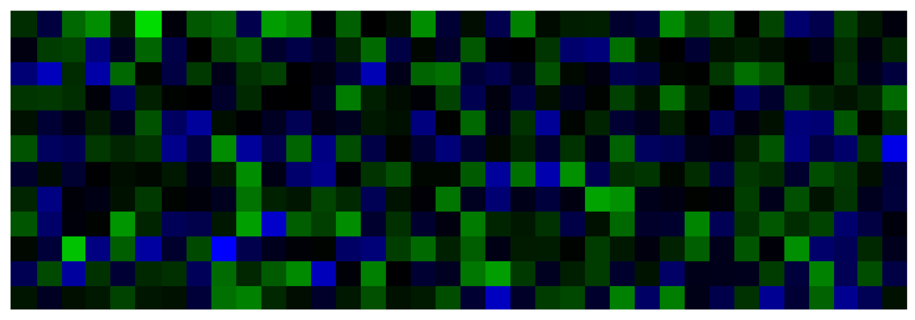
Comparative evaluation results on CIFAR-100 with data augmentation

Model \ p	1%	5%	10%	15%	20%	30%	40%	50%	∼	100%
KerasNet	9.09%	24.68%	32.63%	38.64%	42.62%	47.64%	49.91%	52.46%	…	55.45%
X-KerasNet	**10.16%**	**27.15%**	**35.58%**	**42.05%**	**43.77%**	**48.80%**	**50.48%**	**54.25%**	…	**57.18%**
FitNet4	7.25%	17.94%	23.55%	29.24%	38.76%	48.07%	50.06%	56.01%	…	65.59%
X-FitNet4	**7.35%**	**20.39%**	**28.69%**	**37.86%**	**43.75%**	**50.48%**	**55.40%**	**57.92%**	…	**67.19%**

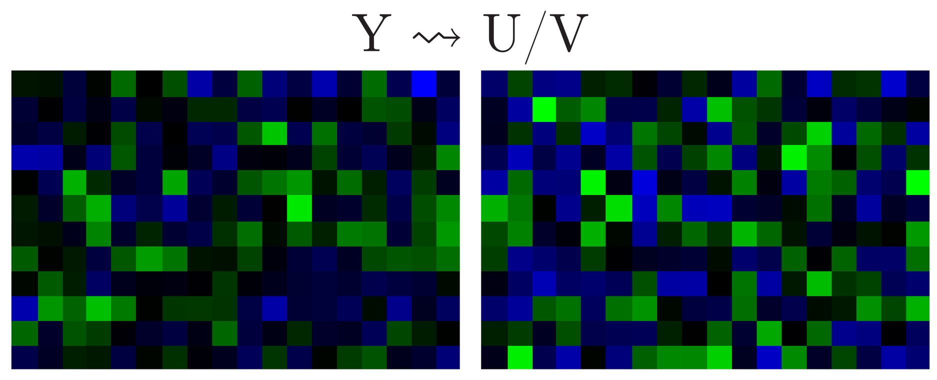
Cross-connection analysis

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:

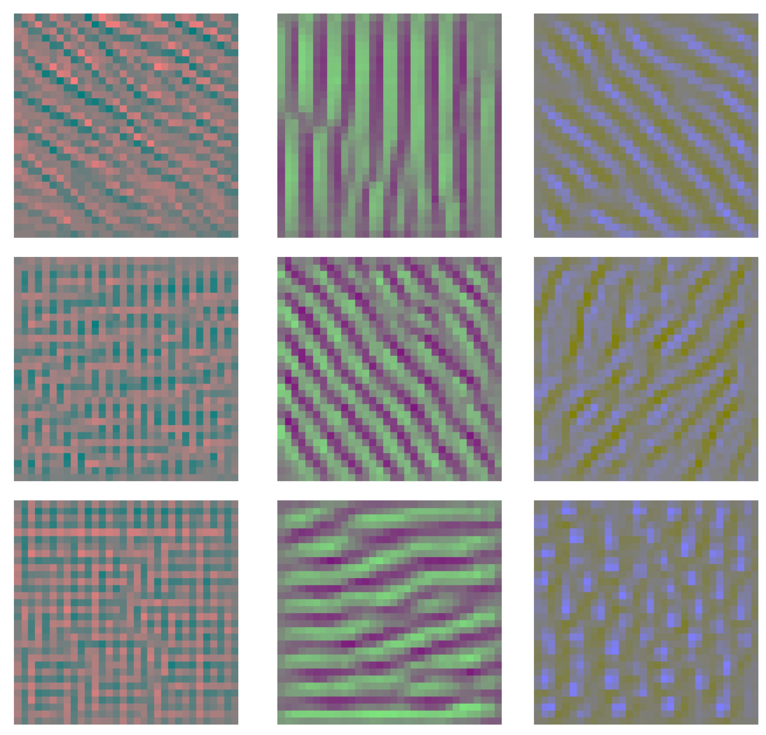


$Y \rightsquigarrow U/V$



$U \rightsquigarrow Y$ $V \rightsquigarrow Y$

We also applied layer-wise feature-map activation techniques proposed by Simonyan *et al.* [3], performing gradient ascent on a white-noise input image to maximise activations of the first cross-connection feature maps:



$Y \rightsquigarrow U/V$ $U \rightsquigarrow Y$ $V \rightsquigarrow Y$

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).

Conclusion

We have introduced X-CNNs, a novel type of CNN architecture derivable directly from a baseline CNN to decouple processing of separate views into an image into multiple CNN streams, with cross-connections allowing for information flow between them. This construction reduces the number of trainable parameters per-stream, significantly improving performance in a variety of data sparsity scenarios.

References

[1] Anton L. Beer, Tina Plank, and Mark W. Greenlee. Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Experimental Brain Research*, 213(2):299–308, 2011.

[2] Weiping Yang, Jingjing Yang, Yulin Gao, Xiaoyu Tang, Yanna Ren, Satoshi Takahashi, and Jinglong Wu. Effects of sound frequency on audiovisual integration: An event-related potential study. *PLoS ONE*, 10(9):1–15, 09 2015.

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[4] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[5] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[6] Petar Veličković, Duo Wang, Nicholas D Lane, and Pietro Liò. **X-CNN: Cross-modal Convolutional Neural Networks for Sparse Datasets**. *arXiv preprint arXiv:1610.00163*, 2016.