Non-local Neural Networks

Cheng Guan

August 15, 2018

Abstract

Both convolutional and recurrent operations are building blocks that process one local neighborhood at a time. In this paper, the authors present non-local operations as a generic family of building blocks for capturing long-range dependencies. Inspired by the classical non-local means method [1] in computer vision, their non-local operation computes the response at a position as a weighted sum of the features at all positions. This building block can be plugged into many computer vision architectures. On the task of video classification, even without any bells and whistles, their non- local models can compete or outperform current competition winners on both Kinetics and Charades datasets.

1. Introduction

Capturing long-range dependencies is of central importance in deep neural networks. For sequential data (*e.g.*, in speech, language), recurrent operations [6, 5] are the dominant solution to long-range dependency modeling. For image data, long-distance dependencies are modeled by the large receptive fields formed by deep stacks of convolutional operations [3].

In this paper, the authors present non-local operations as an efficient, simple, and generic component for capturing long- range dependencies with deep neural networks. Their proposed non-local operation is a generalization of the classical non-local mean operation [1] in computer vision. Intuitively, a non-local operation computes the response at a position as a weighted sum of the features at all positions in the input feature maps as shown in Fig. 1. The set of positions can be in space, time, or spacetime, implying that their operations are applicable for image, sequence, and video problems.

There are several advantages of using non-local operations: (a) In contrast to the progressive behavior of recurrent and convolutional operations, non-local operations capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distance; (b) As we show in experiments, non-local operations are efficient and achieve their best results even with only a few layers (*e.g.*, 5); (c) Finally, their non-local operations maintain the variable input sizes and can be easily combined with other operations.

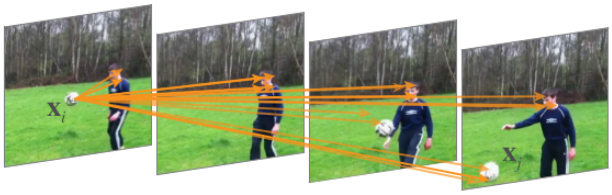

Figure 1. A spacetime non-local operation in their network trained for video classification in Kinetics. A position x_i's response is computed by the weighted average of the features of all positions x_j (only the highest weighted ones are shown here). In this example computed by their model, note how it relates the ball in the first frame to the ball in the last two frames.

2. Related Work

Non-local image processing. Non-local means [1] is a classical filtering algorithm that computes a weighted mean of all pixels in an image. It allows distant pixels to contribute to the filtered response at a location based on patch appearance similarity. This non-local filtering idea was later developed into BM3D (block-matching 3D) [2], which performs filtering on a group of similar, but non-local, patches. BM3D is a solid image denoising baseline even compared with deep neural networks.

Graphical models. Long-range dependencies can be modeled by graphical models such as conditional random fields (CRF). In the context of deep neural networks, a CRF can be exploited to post-process semantic segmentation predictions of a network. The iterative mean-field inference of CRF can be turned into a recurrent network and trained.

Feedforward modeling for sequences. Recently there emerged a trend of using feedforward (*e.g.*, non-recurrent) networks for modeling sequences in speech and language [4, 7] . In these methods, long-term dependencies are captured by the large receptive fields contributed by very deep

1-D convolutions.

Self-attention. Their work is related to the recent self-attention [8] method for machine translation. A self-attention module computes the response at a position in a sequence (*e.g.*, a sentence) by attending to all positions and taking their weighted average in an embedding space. As they will discuss in the next, self-attention can be viewed as a form of the non-local mean [1], and in this sense their work bridges self-attention for machine translation to the more general class of non-local filtering operations that are applicable to image and video problems in computer vision.

References

[1] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 1, 2

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE TIP*, 2007. 1

[3] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*. Springer, 1982. 1

[4] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017. 1

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 1

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 1986. 1

[7] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *arXiv preprint arXiv:1609.03499*, 2016. 1

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 2