FREQUENCY DILATION LEARNING FOR TEMPORAL CONVOLUTIONAL NEURAL NETWORKS

PREPRINT- WORK IN PROGRESS

August 12, 2019

ABSTRACT

Convolutional neural networks (CNNs) make up the bedrock in modern machine learning. With ever increasing data sets and increasing efforts to port these foundational systems to mobile and robotics applications there is an ever strong desire to reduce the computational and memory footprint of CNNs. In this paper we introduce frequency dilation, a novel technique to increase efficiency at minimal cost in terms of network accuracy.

1 Introduction

Integral transforms lie at the core of computational efficiency trough sparsity [2][6]. Previous works have reduced the parameter count of fully connected layers trough the fast Hadamard transforms [9]. Fully connected layers are re-parametrized trough a fixed basis. Using a fixed bases [1] proposes a unitary RNN, later [7] finds that in the RNN case using a fixed basis is detrimental to network performance. As we believe that the fixed basis in [9] is equally restricted, and therefore detrimental to performance. Instead we proposed a learn-able representation based on the dual tree wavelet transform [4], which we apply to both input data and network weights.

2 Related work

Learnable filters [3]

Dimensionality reduction in inputs (FourierRNN)

Dimensionality reduction in weights [9]

3 Theory

Dilation in modern convolutional neural networks is equivalent to low-pass filtering in the frequency domain. This connection becomes apparent if one looks at the dilated CNN output as a downsampled version of the original CNN. Down-sampling is a form of low-pass filtering, because high frequency information present above frequencies not covered by the new Nyquist rate vanishes. When using dilated CNNs one broadens the receptive field by ignoring high frequency information. But depending on the data this may not be the best approach. In this work we explore an alternative dilation method. Instead of hard coding a low pass filter we explore broadening the receptive field trough compressive wavelet basis learning. The data is represented in terms of a learned wavelet basis and only a small number of coefficients is kept. In order to minmize the loss the optimizer will be forced to move the wavelet-filters into frequency bands where most of the relevant information is present. Sparsity and thereby the effectiveness of a representation is basis dependent [6][5], when a suitable basis is chosen few coefficients are are required in order to represent a signal, only a limited number of basis functions components are able to do the job.

3.1 Single tree wavelet basis-optimization

Wavelets and filter banks are closely related. This section explores how methods from the field of digital filter design can be used to optimize wavlet basis representations.

3.1.1 Filter coefficient optimization

Perfect reconstruction (PR) [4][6, page 107] requires, no distortion:

$$H_0(z)F_0(z) + H_1(z)F_1(z) = 2 \tag{1}$$

and alias cancellation:

$$H_0(-z)F_0(z) + H_1(-z)F_1(z) = 0 \tag{2}$$

For alias cancellation $F_0(z) = H_1(-z)$, $F_1(z) = -H_0(-z)$ is typically chosen [5]. A product filter approach is chosen to deal with the reconstruction condition.

$$P_0(z) = F_0(z)H_0(z); P_1(z) = F_1(z)H_1(z) \tag{3}$$

Alias cancellation leads to $P_1(z) = -P_0(-z)$ and turns the no distortion condition into:

$$P(z) + P(-z) = 2 \tag{4}$$

Even powers have to be zero, odd powers are design variables[6, page 107]. In theory it should be possible to optimize the odd power coefficients by gradient descent. One way designers choose the coeffecients is to picke the coefficients for the zeroth phase and generate subsequent phase filters by alternating the signs or flip-inversion [6, page 109]. After hard coding the alternation or inversion conditions, this process should lend itself to optimization by gradient descent.

3.1.2 Orthogonal polyphase matrix optimization

The filter design problem can be expressed in terms of choosing the analysis and reconstruction polyphase matrices. A polyphase system requires $\mathbf{F}_p\mathbf{H}_p = \mathbf{I}$, to work [6, page 116]. In the orthogonal case we have $\mathbf{H}_p = \mathbf{F}_p^{-1}$. Given an initial orthogonal wavelet basis it should be possible to use Stiefel manifold optimization as outlined in [7][8], to solve the wavelet optimization problem by using:

$$\mathbf{H}_{p,k+1} = (\mathbf{I} + \frac{\lambda}{2}\mathbf{A}_k)^{-1}(\mathbf{I} - \frac{\lambda}{2}\mathbf{A}_k)\mathbf{H}_{p,k} \tag{5}$$

Where $\mathbf{A}_k = \mathbf{H}_{p,k}\overline{\bigtriangledown_w F}^T - \overline{\mathbf{H}_{p,k}}^T \bigtriangledown_w F$, with the cost function F and $\mathbf{H}_{p,k}$ the orthogonal polyphase matrix at time k. Ideally the pros and cons of orthogonal wavelets will become apparent in comparison to the biorthogonal case described before.

3.2 Dual vs. single Tree approach

(MW: TODO: Write)

3.3 The dual tree wavelet basis optimization

(MW: I think dual tree wavelet basis optimization is based on using single tree methods plus the phase constraint, which connects the two trees. Dual tree wavelets are always complex valued, so I am hoping to connect this to [8]. I am going to finish this section later:
Dual tree wavelet filters should satisfy [4]:

- **approximate half sample delay property (?).**
- **Perfect reconstruction (orthogonal or bi-orthogonal)**
- **finite support (FIR filters)**
- **vanishing moments good stopband**
- **linear phase filters (desired, but not required)**

Where and how does the phase shift condition fit into this?

The wavelet literature [mallat] tells us $\int_{-\infty}^{\infty} \Psi(t)dt = 0$ and $\int_{-\infty}^{\infty} \|\Psi(t)\|dt = 1$.)

3.4 Multi-resolution analysis

Top-down vs bottom up in image processing, relations?

4 Experiments

4.1 Time series compression, is the wavelet basis code ok?

Compare a highly compressed example using Fourier, Harr, and learned wavelets on mackey glass and a bumpy rectangularly time series for example a staircase mackey glass or lorenz.

(MW: TODO: Choose a good real data source. To test this too.)

4.2 CNN compression, or can we use wavelets to replace the Welsh-Hadamard transform?

Revisit [9] deploy an optimized basis and see if we can do better.

4.3 Frequency dilation, or can wavelets help us broaden CNN receptive fields?

I am not yet sure how to best explore this idea.

References

[1] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary Evolution Recurrent Neural Networks. 48, 2016.

[2] Stephane Mallat and Gabriel Peyré. *A Wavelet Tour of Signal Processing: The Sparse Way*.

[3] Daniel Recoskie and Richard Mann. Learning Sparse Wavelet Representations. pages 1–7, 2018.

[4] Ivan W Selesnick, Richard G Baraniuk, and Nick G Kingsbury. The Dual-Tree Complex Wavelet Transform. (November 2005):123–151, 2005.

[5] Gilbert Strang. Wavelets. 82(3):250–255, 1994.

[6] Gilbert Strang and Truong Nguyen. Wavelets and Filter Banks, 1997.

[7] Scott Wisdom, Thomas Powers, John R Hershey, Jonathan Le Roux, and Les Atlas. Full-Capacity Unitary Recurrent Neural Networks arXiv : 1611 . 00035v1 [stat . ML] 31 Oct 2016. (Nips):1–9, 2016.

[8] Moritz Wolter and Angela Yao. Complex gated recurrent neural networks. In *Advances in Neural Information Processing Systems 31*, 2018.

[9] Ziyu Yang, Zichao and Moczulski, Marcin and Denil, Misha and de Freitas, Nando and Smola, Alex and Song, Le and Wang, Le Song, and Ziyu Wang. Deep Fried Convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476—-1483, 2015.