Chapter 1

Learning in the Wavelet Domain

Some intro text

1.1 Gradients in critically sampled wavelet systems

If wavelet transforms are to have any place in a deep learning architecture, it is important that we are able to calculate the derivatives of wavelet activations with respect to inputs, and use that to define efficient methods to propagate gradients back from a loss function to any point.

Let us start with the 1-D discrete wavelet transform for this. The 1-D DWT is composed of the following components:

1. Decimation
2. Interpolation
3. Convolution
4. Padding (often used to handle the borders of images before convolution)

Once we define the derivative of the output w.r.t.the input for each of these blocks, we can use the chain rule to arbitrarily find the derivatives through any path in the system.

Proof. This is a well-known property but we can prove it here for the discrete 1-D case. Let

$$y[n] = (x * h)[n] = \sum_{m=-\inf}^{\inf} x[m]h[n-m] = \sum_{m=-\inf}^{\inf} x[n-m]h[m]$$

Then □

1.2 Introduction

Using wavelet based methods with deep learning is nascent but not novel. Wavelets have been applied to texture classification [1], [2], super-resolution [3] and for adding detail back into dense

pixel-wise segmentation tasks [4]. One exciting piece of work built on wavelets is the Scattering Transform [5], which has been used as a feature extractor for learning, firstly with simple classifiers [6], [7], and later as a front end to hybrid deep learning tasks[8], [9]. Despite their power and simplicity, scattering features are fixed and are visibly different to regular CNN features [10] - their nice invariance properties come at the cost of flexibility, as there is no ability to learn in between scattering layers.

For this reason, we have been investigating a slightly different approach, more similar to the Fourier based work in [11] in which Rippel et. al. investigate parameterization of filters in the Fourier domain. In the forward pass, they take the inverse DFT of their filter, and then apply normal pixel-wise convolution. We wish to extend this by not only parameterizing filters in the wavelet domain, but by performing the convolution there as well (i.e., also taking the activations into the wavelet domain). After processing is done, we can return to the pixel domain. Doing these forward and inverse transforms has two significant advantages: i) the layers can easily replace standard convolutional layers if they accept and return the same format; ii) we can learn both in the wavelet and pixel space.

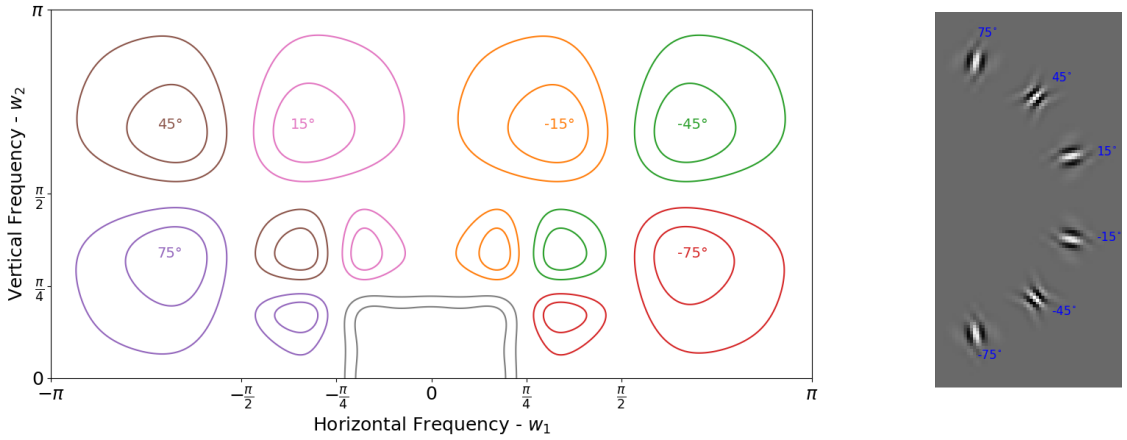
As neural network training involves presenting thousands of training samples, we want our layer to be fast. To achieve this we would ideally choose to use a critically sampled filter bank implementation. The fast 2-D Discrete Wavelet Transform (DWT) is a possible option, but it has two drawbacks: it has poor directional selectivity and any alteration of wavelet coefficients will cause the aliasing cancelling properties of the reconstructed signal to disappear. Instead we choose to use the Dual-Tree Complex Wavelet Transform ($DTCWT$) [12] as at the expense of limited redundancy (4:1), it enables us to have better directional selectivity, and allows us to modify the wavelet coefficients and still have minimal aliasing terms when we reconstruct [13].

section 1.3 of the paper describes the implementation details of our design, and section 1.4 describes the experiments and results we have done so far.

1.3 Method

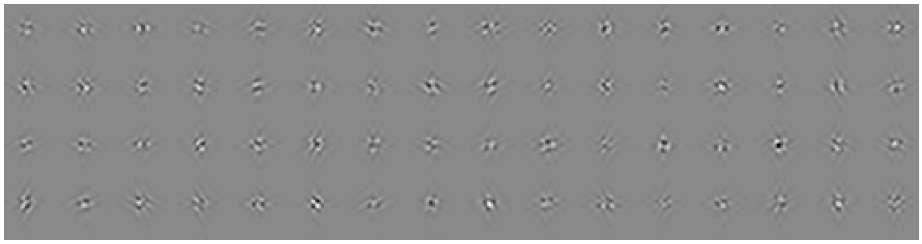
In a standard convolutional layer, an input with C channels, H rows and W columns is $X \in \mathbb{R}^{C\times H\times W}$, which is then convolved with F filters of spatial size K - $w \in \mathbb{R}^{F\times C\times K\times K}$, giving $Y \in \mathbb{R}^{F\times H\times W}$. In many systems like **he˙deep˙2015**, [14], the first layer is typically a selection of bandpass filters, selecting edges with different orientations and center frequencies. In the wavelet space this would be trivial - take a decomposition of each input channel and keep individual subbands (or equivalently, attenuate other bands), then take the inverse wavelet transform. Figure 1.1 shows the frequency space for the DT$\mathbb{C}$WT and makes it clearer as to how this could be done practically for a two scale transform. To attenuate all but say the 15° band at the first scale for the first input channel, we would need to have $13C$ gains for the 13 subbands and C input channels, $13C-1$ of which would be zero and the remaining coefficient one.

Instead of explicitly setting the gains, we can randomly initialize them and use backpropagation to learn what they should be. This gives us the power to learn more complex shapes rather than simple edges, as we can mix the regions of the frequency space per input channel in an arbitrary way.



(a)

(b)



(c)

Figure 1.1: (a) Contour plots at -1dB and -3dB showing the support in the Fourier domain of the 6 subbands of the DT$\mathbb{C}$WT at scales 1 and 2 and the scale 2 lowpass. These are the product $P(z)Q(z)$ from Equation 1.3.1.(b) The pixel domain impulse responses for the second scale wavelets. (c) Example impulses of our layer when g_1, and g_{lp} are 0 and $g_2 \in \mathbb{C}^{6\times1\times1}$, with each real and imaginary element drawn from $\mathcal{N}(0,1)$. I.e., only information in the 6 subbands with $\frac{\pi}{4} < |w_1|, |w_2| < \frac{\pi}{2}$ from (a) is passed through.

1.3.1 Forward propagation

Figure 1.2 shows the block diagram using Z-transforms for a single band of our system (it is based on Figure 4 in [13]). To keep things simple for the rest of section 1.3 the figure shown is for a 1-D system; it is relatively straightforward to extend this to 2-D[12]. The complex analysis filter (taking us into the wavelet domain) is $P(z) = \frac{1}{2}(A(z) + jB(z))$ and the complex synthesis filter (returning us to the pixel domain) is $Q(z) = \frac{1}{2}(C(z) - jD(z))$ where A, B, C, D are real. If $G(z) = G_r(z) + jG_i(z) = 1$ then the end-to-end transfer function is (from section 4 of [13]):

$$\frac{Y(z)}{X(z)} = \frac{2}{M}(P(z)Q(z) + P^*(z)Q^*(z)) \tag{1.3.1}$$

where P, Q have support only in the top half of the Fourier plane and P^*, Q^* are P and Q reflected in the horizontal frequency axis. Examples of $P(z)Q(z)$ for different subbands of a 2-D DT$\mathbb{C}$WT have spectra shown in 1.1a, $P^*(z)Q^*(z)$ make up the missing half of the frequency space.

Modifying this from the standard wavelet equations by adding the subband gains $G_r(z)$ and $G_i(z)$, the transfer function becomes:

$$\frac{Y(z)}{X(z)} = \frac{2}{M}\left[\, G_r(z^M)(P(z)Q(z) + P^*(z)Q^*(z)) + jG_i(z^M)(P(z)Q(z) - P^*(z)Q^*(z))\right] \tag{1.3.2}$$

1.3.2 Backpropagation

We start with the commonly known property that for a convolutional block, the gradient with respect to the input is the gradient with respect to the output convolved with the time reverse of the filter. More formally, if $Y(z) = H(z)X(z)$:

$$\Delta X(z) = H(z^{-1})\Delta Y(z) \tag{1.3.3}$$

where $H(z^{-1})$ is the Z-transform of the time/space reverse of $H(z)$, $\Delta Y(z) \triangleq \frac{\partial L}{\partial Y}(z)$ is the gradient of the loss with respect to the output, and $\Delta X(z) \triangleq \frac{\partial L}{\partial X}(z)$ is the gradient of the loss with respect to the input. If H were complex, the first term in Equation 1.3.3 would be $\bar{H}(1/\bar{z})$, but as each individual block in the DT$\mathbb{C}$WT is purely real, we can use the simpler form.

Assume we already have access to the quantity $\Delta Y(z)$ (this is the input to the backwards pass). **??** illustrates the backpropagation procedure. An interesting result is that the backwards pass of an inverse wavelet transform is equivalent to doing a forward wavelet transform.[1] Similarly, the backwards pass of the forward transform is equivalent to doing the inverse transform. The weight update gradients are then calculated by finding $\Delta W(z) = \text{DT}\mathbb{C}\text{WT}\{\Delta Y(z)\}$ and then convolving with the time reverse of the saved wavelet coefficients from the forward pass - $V(z)$.

$$\Delta G_r(z) = \Delta W_r(z)V_r(z^{-1}) + \Delta W_i(z)V_i(z^{-1}) \tag{1.3.4}$$

$$\Delta G_i(z) = -\Delta W_r(z)V_i(z^{-1}) + \Delta W_i(z)V_r(z^{-1}) \tag{1.3.5}$$

[1] As shown in **??**, the analysis and synthesis filters have to be swapped and time reversed. For orthogonal wavelet transforms, the synthesis filters are already the time reverse of the analysis filters, so no change has to be done. The q-shift filters of the DT$\mathbb{C}$WT [15] have this property.

Unsurprisingly, the passthrough gradients have similar form to Equation 1.3.2:

$$\Delta X(z) = \frac{2\Delta Y(z)}{M}\left[G_r(z^{-M})\,(PQ + P^*Q^*) + jG_i(z^{-M})\,(PQ - P^*Q^*)\right] \tag{1.3.6}$$

where we have dropped the z terms on $P(z), Q(z), P^*(z), Q^*(z)$ for brevity.

Note that we only need to evaluate equations 1.3.4,1.3.5,1.3.6 over the support of $G(z)$ i.e., if it is a single number we only need to calculate $\Delta G(z)|_{z=0}$.

1.3.3 Memory Cost

Again considering a two scale transform — instead of learning $w \in \mathbb{R}^{F\times C\times K\times K}$ we learn complex gains at the two scales, and a real gain for the real lowpass:

$$\begin{aligned} g_1 &\in \mathbb{C}^{F\times C\times 6\times 1\times 1} \\ g_2 &\in \mathbb{C}^{F\times C\times 6\times 1\times 1} \\ g_{lp} &\in \mathbb{R}^{F\times C\times 1\times 1} \end{aligned}$$

We have set the spatial dimension to be 1×1 to show that this gain is identical to a 1×1 convolution over the complex wavelet coefficients. If we wish, we can learn larger spatial sizes to have more complex attenuation/magnification of the subbands. We also can use more/fewer than 2 wavelet scales. At first glance, we have increased our parameterization by a factor of 25 (13 subbands, of which all but the lowpass are complex), but each one of these gains affects a large spatial size. For the first scale, the effective size is about 5×5 pixels, for the second scale it is about 15×15.

1.3.4 Computational Cost

A standard convolutional layer needs K^2F multiplies per input pixel (of which there are $C\times H\times W$). In comparison, the wavelet gain method does a set number of operations per pixel for the forward and inverse transforms, and then applies gains on subsampled activations. For a 2 level DT$\mathbb{C}$WT the transform overhead is about 60 multiplies for both the forward and inverse transform. It is important to note that unlike the filtering operation, this does not scale with F. The learned gains in each subband do scale with the number of output channels, but can have smaller spatial size (as they have larger effective sizes) as well as having fewer pixels to operate on (because of the decimation). The end result is that as F and C grow, the overhead of the C forward and F inverse transforms is outweighed by cost of FC mixing processes, which should in turn be significantly less than the cost of FC $K\times K$ standard convolutions for equivalent spatial sizes.

1.3.5 Examples

1.1c show example impulse responses of our layer. These impulses were generated by randomly initializing both the real and imaginary parts of $g_2 \in \mathbb{C}^{6\times 1\times 1}$ from $\mathcal{N}(0,1)$ and g_1, g_{lp} are set to 0. I.e. each shape has 12 random variables. It is good to see that there is still a large degree of variability between shapes. Our experiments have shown that the distribution of the normalized cross-correlation between 512 of such randomly generated shapes matches the distribution for random vectors with roughly 11.5 degrees of freedom.

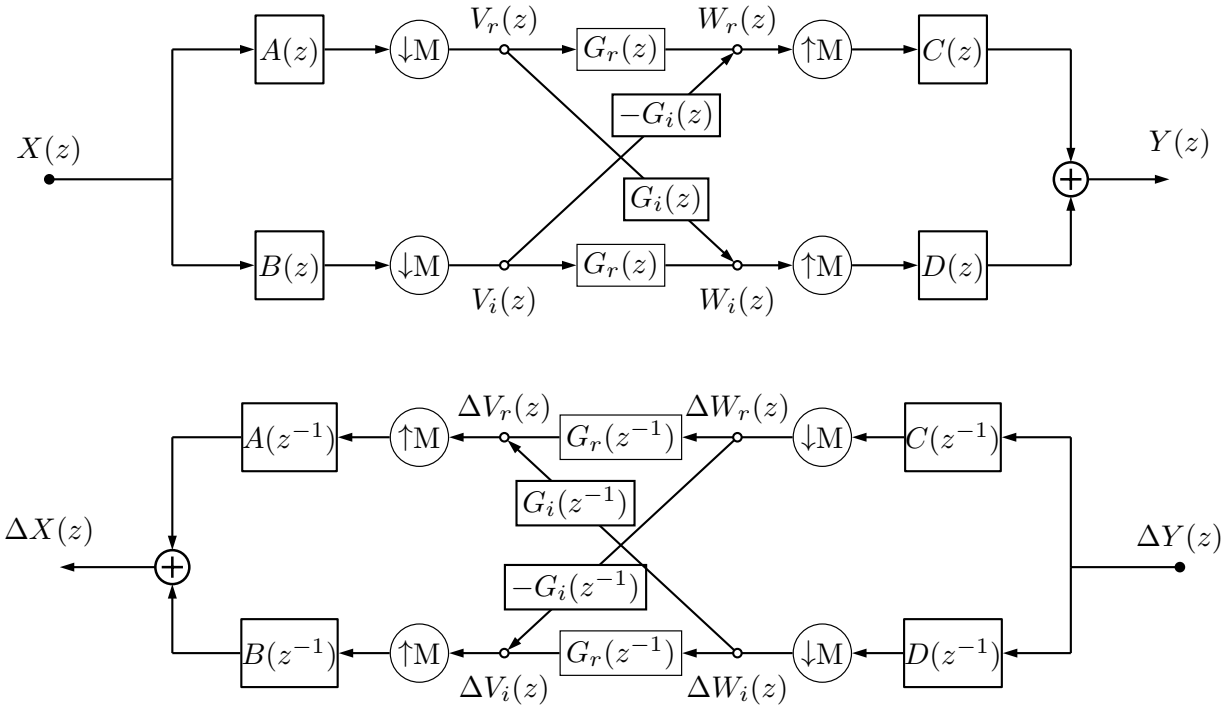


Figure 1.2: **Forward and backward block diagrams for** DT$\mathbb{C}$WT **gain layer.** Based on Figure 4 in [13]. Ignoring the G gains, the top and bottom paths (through A, C and B, D respectively) make up the the real and imaginary parts for *one subband* of the dual tree system. Combined, $A + jB$ and $C - jD$ make the complex filters necessary to have support on one side of the Fourier domain (see Figure 1.1). Adding in the complex gain $G_r + jG_i$, we can now attenuate/shape the impulse response in each of the subbands. To allow for learning, we need backpropagation. The bottom diagram indicates how to pass gradients $\Delta Y(z)$ through the layer. Note that upsampling has become downsampling, and convolution has become convolution with the time reverse of the filter (represented by z^{-1} terms).

1.4 Experiments and Preliminary Results

To examine the effectiveness of our convolutional layer, we do a simple experiment on CIFAR-10 and CIFAR-100. For simplicity, we compare the performance using a simple yet relatively effective convolutional architecture - LeNet [16]. LeNet has 2 convolutional layers of spatial size 5×5 followed by 2 fully connected layers and a softmax final layer. We swap both these convolutional layers out for two of our proposed wavelet gain layers (keeping the ReLU between them). As CIFAR has very small spatial size, we only take a single scale DT$\mathbb{C}$WT. Therefore each gain layer has 6 complex gains for the 6 subbands, and a 3×3 real gain for the lowpass (a total of $21C$ parameters vs $25C$ for the original system). We train both networks for 200 epochs with Adam [17] optimizer with a constant learning rate of 10^{-3} and a weight decay of 10^{-5}. The code is available at [18]. Table 1.1 shows the mean of the validation set accuracies for 5 runs. The different columns represent undersampled training set sizes (with 50000 being the full training set). When undersampling, we

Table 1.1: Comparison of LeNet with standard convolution to our proposed method which learns in the wavelet space (WaveLenet) on CIFAR-10 and CIFAR-100. Values reported are the average top-1 accuracy (%) rates for different train set sizes over 5 runs.

	Train set size	1000	2000	5000	10000	20000	50000
CIFAR-10	LeNet	48.5	52.4	59.5	65.0	69.5	73.3
	WaveLeNet	47.3	52.1	58.7	63.8	68.0	72.4
CIFAR-100	LeNet	11.1	15.8	23.1	29.5	34.4	41.1
	WaveLeNet	11.1	15.4	23.2	28.4	33.9	39.6

keep the samples per class constant. We see our system perform only very slightly worse than the standard convolutional layer.

1.5 Conclusion and Future Work

In this work we have presented the novel idea of learning filters by taking activations into the wavelet domain, learning mixing coefficients and then returning to the pixel space. This work is done as a preliminary step; we ultimately hope that learning in both the wavelet and pixel space will have many advantages, but as yet it has not been explored. We have considered the possible challenges this proposes and described how a multirate system can learn through backpropagation.

Our experiments so far have been promising. We have shown that our layer can learn in an end-to-end system, achieving very near similar accuracies on CIFAR-10 and CIFAR-100 to the same system with convolutional layers instead. This is a good start and shows the plausibility of such an idea, but we need to search for how to improve these layers if they are to be useful. It will be interesting to see how well we can learn on datasets with larger images - our proposed method naturally learns large kernels, so should scale well with the image size.

In our experiments so far, we only briefly go into the wavelet domain before coming back to the pixel domain to do ReLU nonlinearities, however we plan to explore using nonlinearities in the wavelet domain, such as soft-shrinkage to denoise/sparsify the coefficients [19]. We feel there are strong links between ReLU non-linearities and denoising/sparsity ideas, and that there may well be useful performance gains from mixing real pixel-domain non-linearities with complex wavelet-domain shrinkage functions. Thus we present these ideas here as a starting point for a novel and exciting avenue of deep network research.

References

[1] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet Convolutional Neural Networks for Texture Classification", *arXiv:1707.07394 [cs]*, Jul. 2017. arXiv: 1707.07394 [cs].

[2] L. Sifre and S. Mallat, "Combined scattering for rotation invariant texture analysis", in *European Symposium on Artificial Neural Networks (ESANN) 2012*, 2012.

[3] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga, "Deep Wavelet Prediction for Image Super-Resolution", in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1100–1109.

[4] L. Ma, J. Stückler, T. Wu, and D. Cremers, "Detailed Dense Inference with Convolutional Neural Networks via Discrete Wavelet Transform", *arXiv:1808.01834 [cs]*, Aug. 2018. arXiv: 1808.01834 [cs].

[5] S. Mallat, "Group Invariant Scattering", en, *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[6] J. Bruna and S. Mallat, "Invariant Scattering Convolution Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[7] A. Singh and N. Kingsbury, "Scatternet hybrid deep learning (SHDL) network for object classification", in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017, pp. 1–6.

[8] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the Scattering Transform: Deep Hybrid Networks", in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5619–5628. arXiv: 1703.08961.

[9] A. Singh, "ScatterNet Hybrid Frameworks for Deep Learning", PhD thesis, University of Cambridge, May 2018.

[10] F. Cotter and N. Kingsbury, "Visualizing and Improving Scattering Networks", in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017, pp. 1–6. arXiv: 1709.01355.

[11] O. Rippel, J. Snoek, and R. P. Adams, "Spectral Representations for Convolutional Neural Networks", in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015, pp. 2440–2448.

[12] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, "The dual-tree complex wavelet transform", *Signal Processing Magazine, IEEE*, vol. 22, no. 6, pp. 123–151, 2005.

[13] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, May 2001.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *NIPS*, Curran Associates, Inc., 2012, pp. 1097–1105.

[15] N. Kingsbury, "Design of Q-shift complex wavelets for image processing using frequency domain energy minimization", in *2003 International Conference on Image Processing, 2003. ICIP 2003. Proceedings*, vol. 1, Sep. 2003, I-1013-16 vol.1.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv:1412.6980 [cs]*, Dec. 2014. arXiv: 1412.6980 [cs].

[18] F. Cotter, *DTCWT Gainlayer*, Nov. 2018.

[19] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage", en, *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994.