LP-3DCNN: Unveiling Local Phase in 3D Convolutional Neural Networks Supplementary Document

Sudhakar Kumawat and Shanmuganathan Raman
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
{sudhakar.kumawat, shanmuga}@iitgn.ac.in

Abstract

This document contains the supplementary materials to the main paper.

1. Detailed Mathematical Formulation of *Layer 2* of the ReLPV Block

In this section, we elaborate on each step of the *Layer 2* of ReLPV block which is at the core of our ReLPV block.

Let $f(\mathbf{x})$ be the single channel feature map of size $1 \times d \times h \times w$ that is output by *Layer 1* of the ReLPV block. Here, h, w, and d denotes the height, width, and depth of the feature map, respectively. For simplicity, we will drop the channel dimension and rewrite the size of $f(\mathbf{x})$ as $d \times h \times w$. Here, $\mathbf{x} \in \mathbb{Z}^3$ are the 3D coordinates of the elements in $f(\mathbf{x})$.

Every $\mathbf{x}$ in $f(\mathbf{x})$ has a $n \times n \times n$ 3D neighborhood denoted by $\mathcal{N}_{\mathbf{x}}$ which is defined in Equation 1. We provide detailed experimental analysis in the manuscript on the effect of varying n on the performance of the ReLPV block in 3D CNNs meant for video classification task.

$$\mathcal{N}_{\mathbf{x}} = \{\mathbf{y} \in \mathbb{Z}^3 ; \parallel (\mathbf{x}-\mathbf{y}) \parallel_{\infty} \leq r ; n = 2r+1; r \in \mathbb{Z}_+\} \tag{1}$$

For all positions $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{d \cdot h \cdot w}\}$ of the feature map $f(\mathbf{x})$, we use local 3D neighborhoods, $f(\mathbf{x}-\mathbf{y}), \forall \mathbf{y} \in \mathcal{N}_{\mathbf{x}}$ to derive the local frequency domain representation using Short Term Fourier Transform (STFT) as defined in Equation 2.

$$F(\mathbf{v}, \mathbf{x}) = \sum_{\mathbf{y}_i \in \mathcal{N}_{\mathbf{x}}} f(\mathbf{x}-\mathbf{y}_i) \exp^{-j2\pi \mathbf{v}^T \mathbf{y}_i} \tag{2}$$

Here $i = 1, \ldots, n^3$, $\mathbf{v} \in \mathbb{R}^3$ is a 3D frequency variable, and $j = \sqrt{-1}$. Using vector notation [3], we can rewrite Equation 2 as shown in Equation 3.

$$F(\mathbf{v}, \mathbf{x}) = \mathbf{w}_{\mathbf{v}}^T \mathbf{f}_{\mathbf{x}} \tag{3}$$

Here, $\mathbf{w}_{\mathbf{v}}$ is a complex valued basis function (at frequency variable $\mathbf{v}$) of a linear transformation, and is defined as shown in Equation 4.

$$\mathbf{w}_{\mathbf{v}}^T = [\exp^{-j2\pi \mathbf{v}^T \mathbf{y}_1}, \exp^{-j2\pi \mathbf{v}^T \mathbf{y}_2}, \ldots, \exp^{-j2\pi \mathbf{v}^T \mathbf{y}_{n^3}}], \tag{4}$$

and $\mathbf{f}_{\mathbf{x}}$ is a vector containing all the elements from the neighborhood $\mathcal{N}_{\mathbf{x}}$, and is defined as shown in Equation 5.

$$\mathbf{f}_{\mathbf{x}} = [f(\mathbf{x}-\mathbf{y}_1), f(\mathbf{x}-\mathbf{y}_2), \ldots, f(\mathbf{x}-\mathbf{y}_{n^3})]^T \tag{5}$$

In our work, we consider 13 lowest non-zero frequency variables $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{13}$. Low frequency variables are used because they usually contain most of the information, and therefore they have better signal-to-noise ratio than the high frequency components [1] (see Section 2). The values of these frequency variables are already discussed in the main paper. Thus, from Equation 3, the local frequency domain representation for the above frequency variables is defined as shown in Equation 6.

$$\mathbf{F}_{\mathbf{x}} = [F(\mathbf{v}_1, \mathbf{x}), F(\mathbf{v}_2, \mathbf{x}), \ldots, F(\mathbf{v}_{13}, \mathbf{x})]^T \tag{6}$$

At each position $\mathbf{x}$, after separating the real and imaginary parts of each component, we get a vector as shown in Equation. 7.

$$\begin{aligned}\mathbf{F}_{\mathbf{x}} = [&\Re\{F(\mathbf{v}_1, \mathbf{x})\}, \Im\{F(\mathbf{v}_1, \mathbf{x})\}, \Re\{F(\mathbf{v}_2, \mathbf{x})\}, \\ &\Im\{F(\mathbf{v}_2, \mathbf{x})\}, \ldots, \Re\{F(\mathbf{v}_{13}, \mathbf{x})\}, \Im\{F(\mathbf{v}_{13}, \mathbf{x})\}]^T\end{aligned} \tag{7}$$

Here, $\Re\{\cdot\}$ and $\Im\{\cdot\}$ return the real and imaginary parts of a complex number, respectively. The corresponding $26 \times n^3$ transformation matrix can be written as shown in Equation 8.

$$\mathbf{W} = [\Re\{\mathbf{w}_{\mathbf{v}_1}\}, \Im\{\mathbf{w}_{\mathbf{v}_1}\}, \ldots, \Re\{\mathbf{w}_{\mathbf{v}_{13}}\}, \Im\{\mathbf{w}_{\mathbf{v}_{13}}\}]^T \tag{8}$$

Hence, from Equation 3 and 8, the vector form of STFT for all the 13 frequency points $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{13}$ can be written as shown in Equation 9.

$$\mathbf{F_x} = \mathbf{W}\mathbf{f_x} \tag{9}$$

Since, $\mathbf{F_x}$ is computed for all positions $\mathbf{x}$ of the input $f(\mathbf{x})$, it results in an output feature map with size $26 \times d \times h \times w$. This feature map is then passed as input to the *Layer 3* of the ReLPV block.

2. Decorrelation Property of STFT and Reason for Selecting Low Frequency Variables

As mentioned in the manuscript, some important properties of Short Term Fourier Transform (STFT) is its ability to decorrelate the input signal and to compact the energy (information) contained in a signal. These properties are inherent to STFT since it belongs to the family of orthogonal transforms such K-L transform, Walsh-Hadamard transform (WHT), and Discrete Cosine Transform (DCT) [4]. All the above orthogonal transforms have the following properties in common.

- Orthogonal transforms have the tendency of decorrelating the input signals [4]. For example, consider a signal containing temperature as a function of time. Now, given the value of a current sample of the signal, the value of its next sample can be predicted with reasonable confidence to be close to the current one, i.e., two consecutive time samples are highly correlated. On the other hand, after an orthogonal transform, such as Fourier transform, knowing the magnitude of a certain frequency component, one has little idea in terms of the magnitude (or the energy) of the next frequency component, i.e., the two components are much less correlated than the time samples before the transform. The same property holds true for signals in multiple dimensions such as images and videos [2]. In images and videos, decorrelation is achieved due to STFT's insensitivity to the correlation coefficient of images and videos [2].

- Orthogonal transforms tend to compact the energy (information) contained in the signal into a small number of signal components [4]. For example, after Fourier transform, most of the energy (information) will be concentrated in a relatively small number of low frequency components. Most of the high frequency components carry little energy. Moreover, low frequency components have better signal-to-noise ratio than the high frequency components. It is for this reason that we chose low frequency variables while computing STFT.

References

[1] Janne Heikkila and Ville Ojansivu. Methods for local phase quantization in blur-insensitive image analysis. In *LNLA*, pages 104–111, 2009.

[2] B Hinman, Jared Bernstein, and D Staelin. Short-space fourier transform image processing. In *ICASSP*, volume 9, pages 166–169, 1984.

[3] Anil K Jain. *Fundamentals of digital image processing*. Englewood Cliffs, NJ: Prentice Hall,, 1989.

[4] Ruye Wang. *Introduction to orthogonal transforms: with applications in data processing and analysis*. Cambridge University Press, 2012.