Voice Separation Using Deep Complex Network

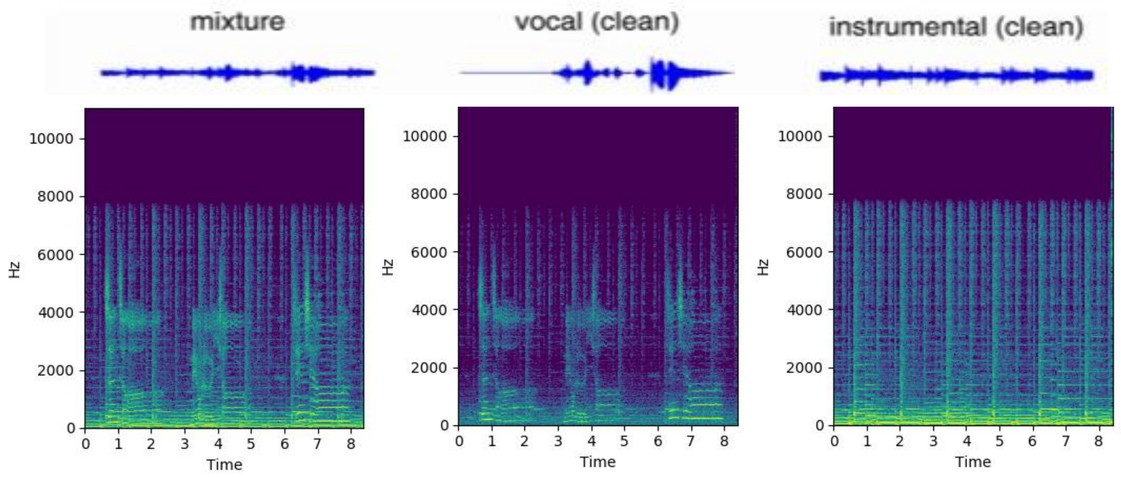
Wei-Che Chen, Chih-Ting Liu, Chih-Hsuan Lo, Ching-Yen Shih
National Taiwan University

Motivation

Deep Complex Network, which means every neuron outputs and weights are complex number, are proposed recently to boost network performance by more sophisticated calculation. In the related work, it didn't improve on MNSITs and Cifar-10, but got a better performance on Musicnet . We assume the reason is that it did Fourier transform for speech signal first, so the input is in the complex domain. Thus, we want to analogously apply deep complex network on the voice separation task of which speech signal mapping to the complex domain by the Fourier transform.

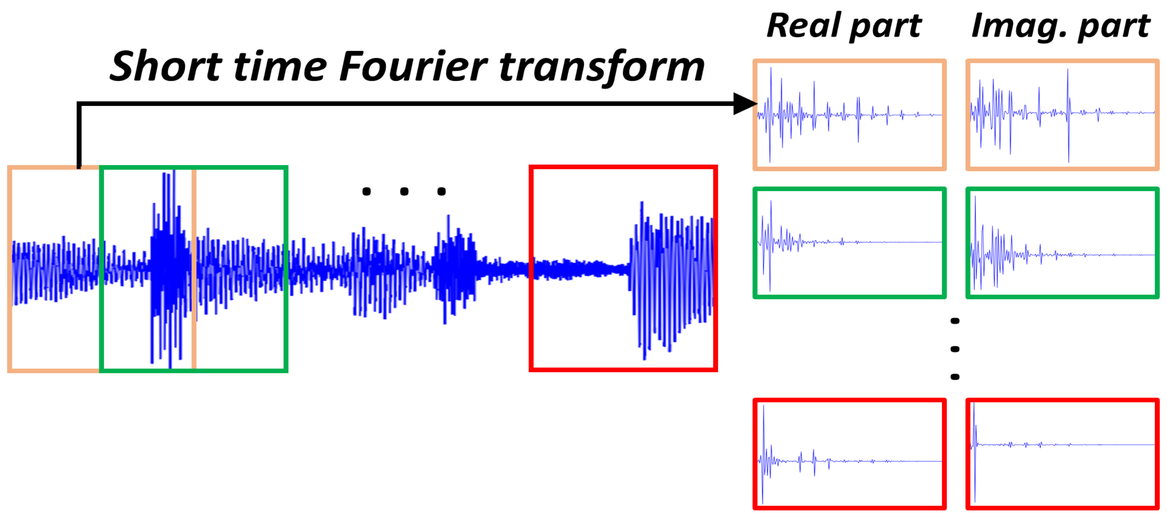
Voice separation

Separate singing voice from the background voice.



Method

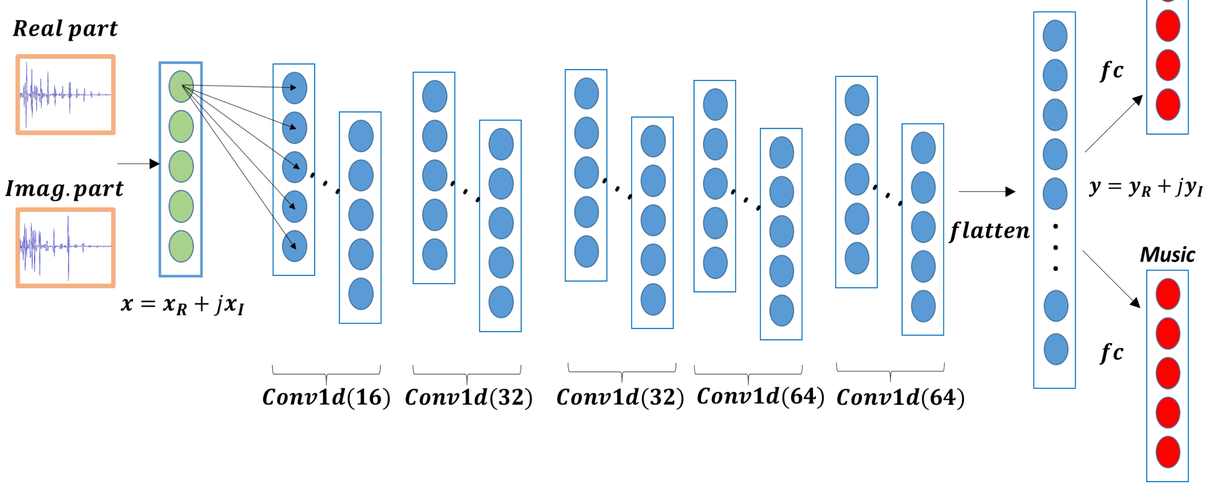
- Transform to spectrogram using short time Fourier transform (STFT) by a sliding window.



- Model architecture

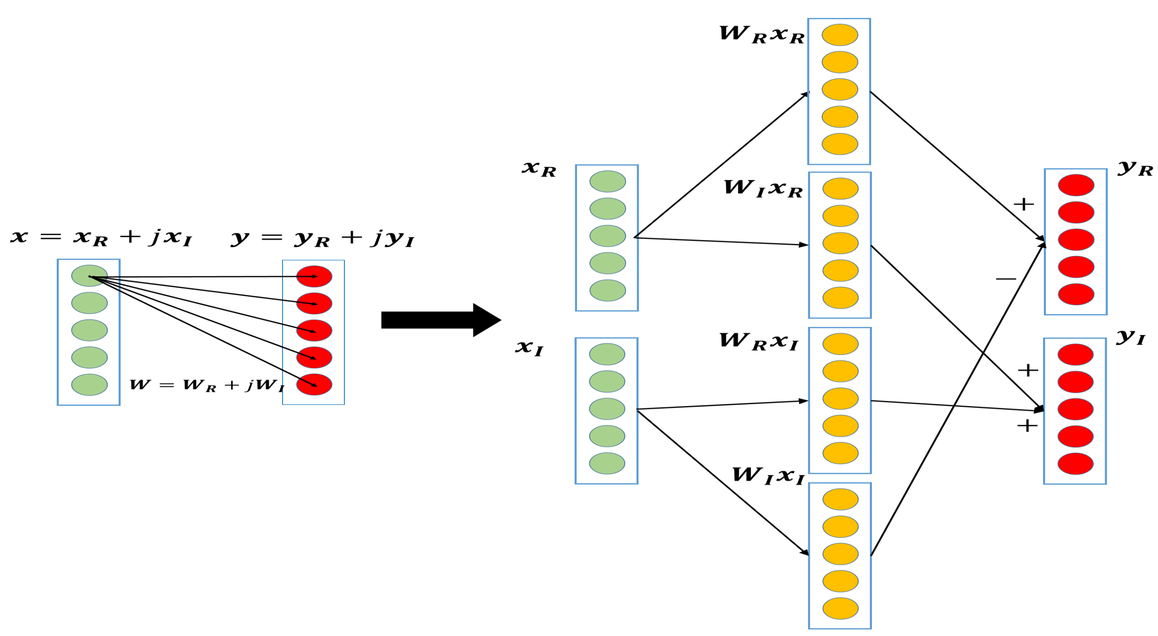
1. We use 5 complex convolutional layer followed by two complex fully connected layers to generate voice and music respectively.

2. $Loss = MSELoss_{voice} + MSELoss_{music}$



Because most libraries don't support complex operation, we have to build the networks manually.
Neural networks are formed by a series of multiply-accumulate operation, so we have to handle the multiplication of complex, such as

$$(\mathbf{W_R} + \mathbf{jW_I}) \times (\mathbf{x_R} + \mathbf{jx_I}) = (\mathbf{W_R x_R} - \mathbf{W_I x_I}) + \mathbf{j}(\mathbf{W_R x_I} + \mathbf{W_I x_R})$$



MIRIK-Dataset

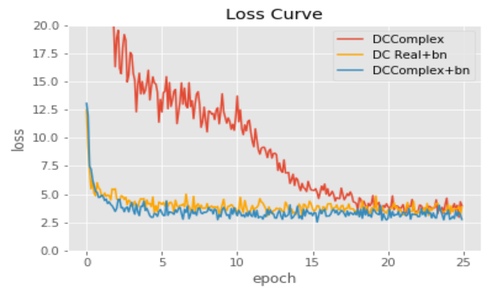
Multimedia Information Retrieval lab, 1000 song clips, dataset for **singing voice separation.**

- 1000 song clips which the **music accompaniment** and the **singing voice** are recorded at **left and right channels**, respectively.

Results

We experiment the following setting in order to figure out the efficacy for complex operation and vanilla model.

- **Deep Convolution Complex(DCComplex):** Setting described above
- **Deep Convolution Real + Batch Normalization(DCReal+BN):** We separate real and imaginary part as two real channel and feed them to the vanilla convolutional network with batch normalization.
- **DCComplex + BN**: DCComplex with batch normalization



Quantitative evaluation on separation metrics

Model	GNSDR	GSIR	GSAR
DCComplex	5.41	8.13	**9.87**
DCComplex+BN	**5.86**	**9.89**	8.90
*DCComplex+Cos	4.98	9.14	8.06
DCReal	5.63	8.79	9.55

*DCComplex + Cos: $Loss = MSELoss_{voice} + MSELoss_{music} + \lambda \cdot |CosineSimilarity(\tilde{y}_{Voice}, \tilde{y}_{Music})|$

Spectrogram visualization

Compared with DCReal model(Fig. 2), our proposed method(Fig. 3) generate clearer spectrogram of voice and less noise in high frequency

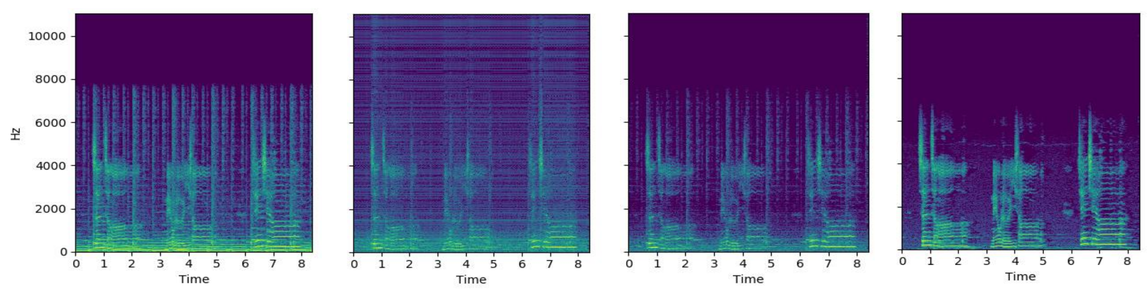
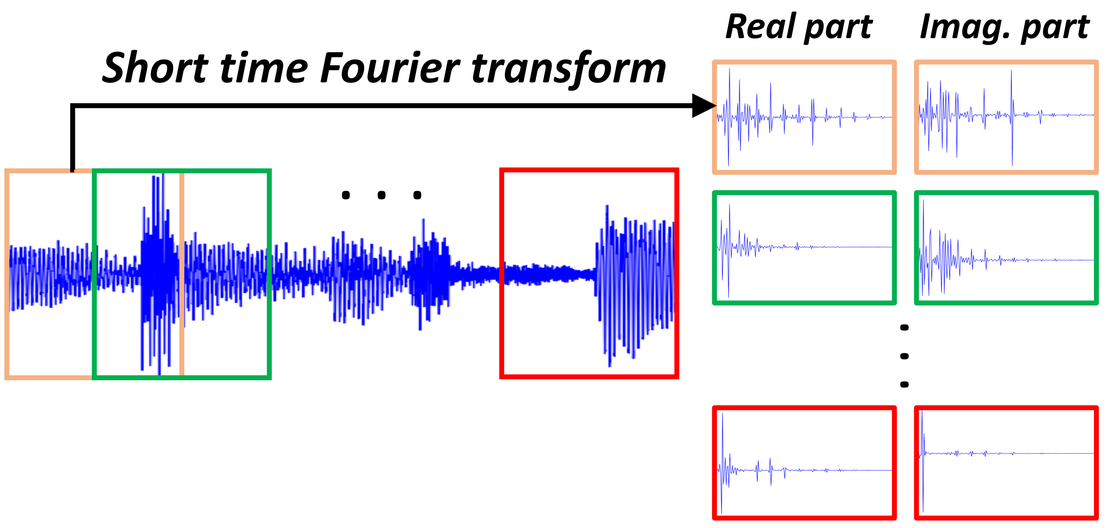
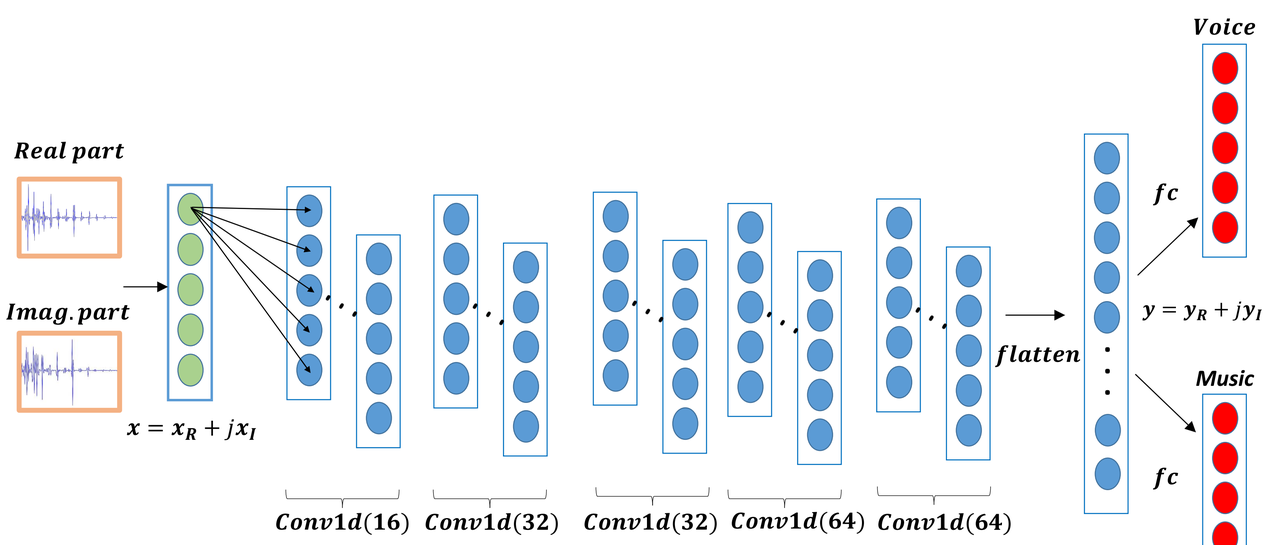


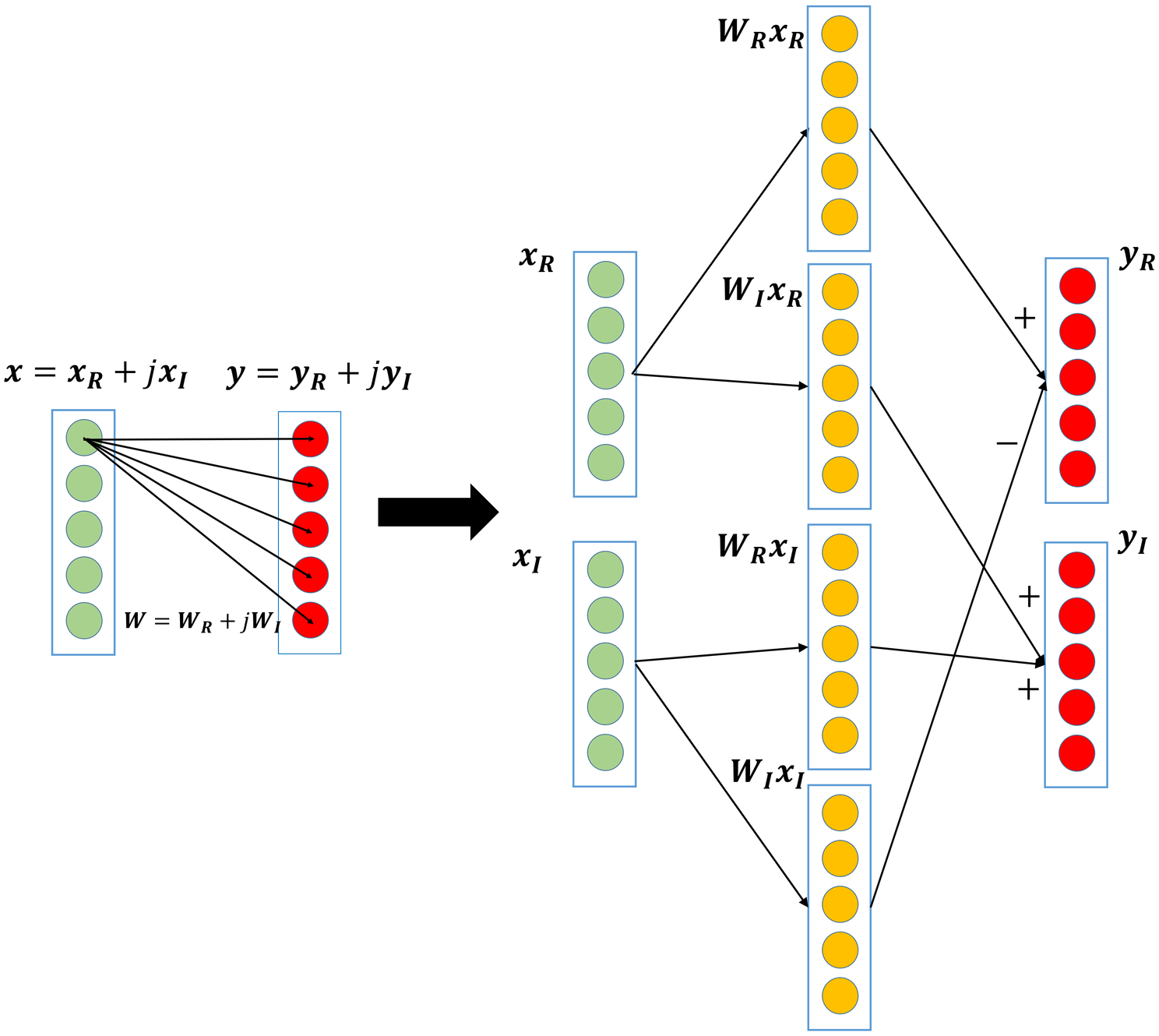
Fig. 1. Mixture | Fig. 2. $Voice_{DCReal}$ | Fig. 3. $Voice_{DCComplex+BN}$ | Fig. 4. $Voice_{Ground\ truth}$



fc



** Every weight and neuron are complex number*



Implementation of complex network

$$x = x_R + jx_I$$

Conv1d(16)

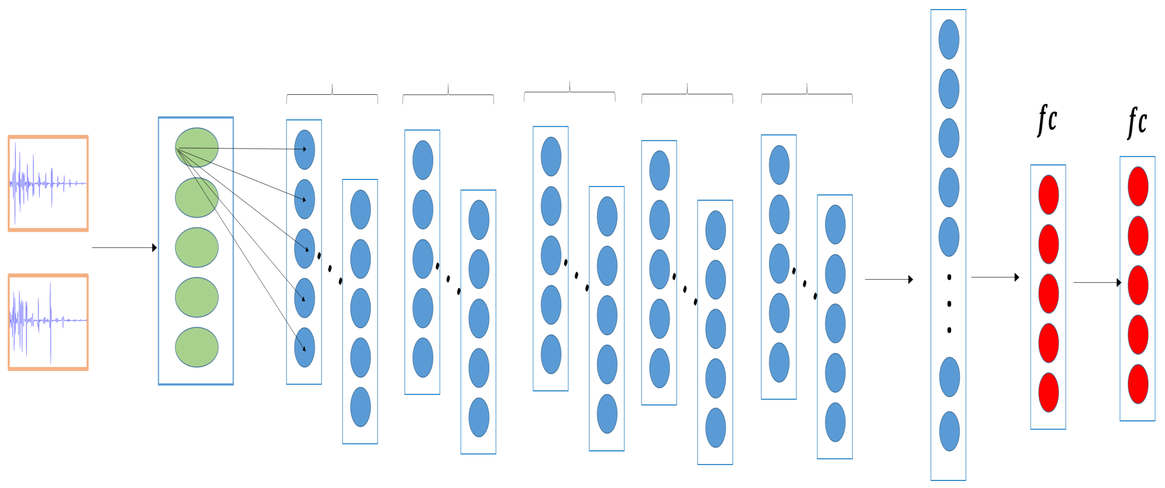
Conv1d(32)

Conv1d(32)

Real part(x_R)

Conv1d(64)

Imag. Part(x_I)



** Every weight and neur*