WaveNet - A Generative Model for Raw Audio

출처 : https://www.youtube.com/watch?v=GyQnex_DK2k

논문 : https://arxiv.org/abs/1609.03499

Generative Model of Raw Audio Waveform

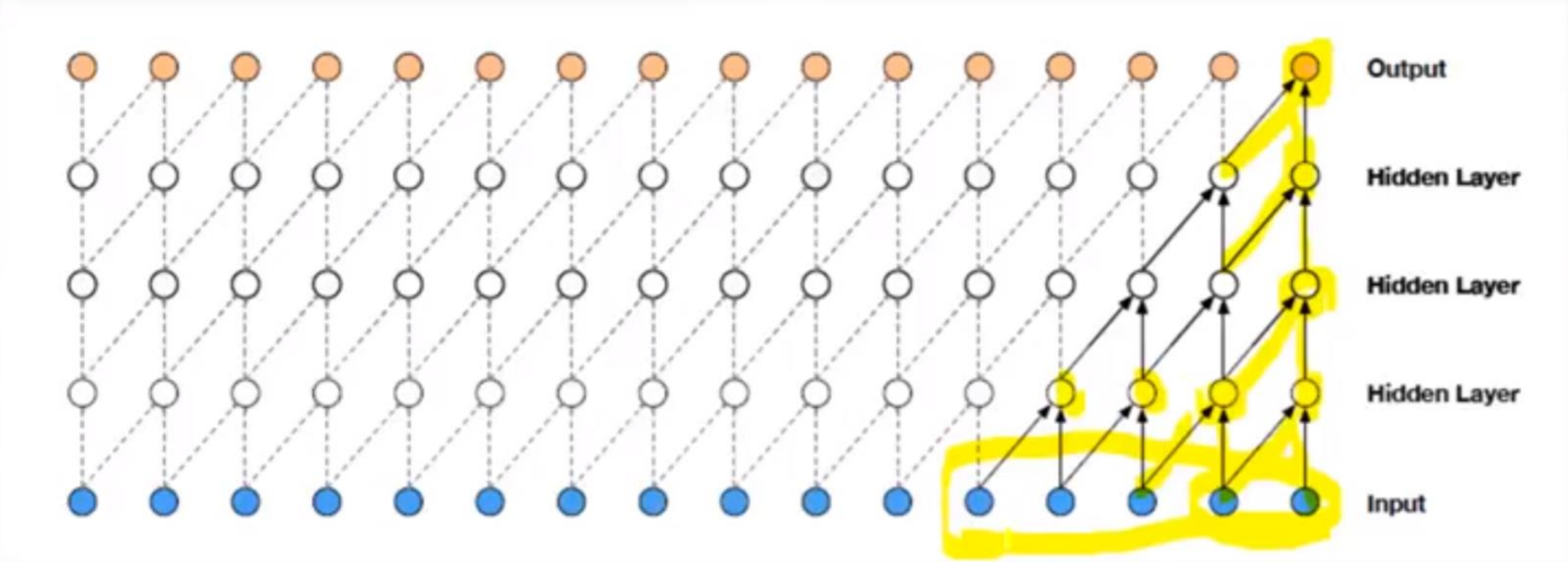
The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ is factorised as a product of conditional probabilities as follows:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \tag{1}$$

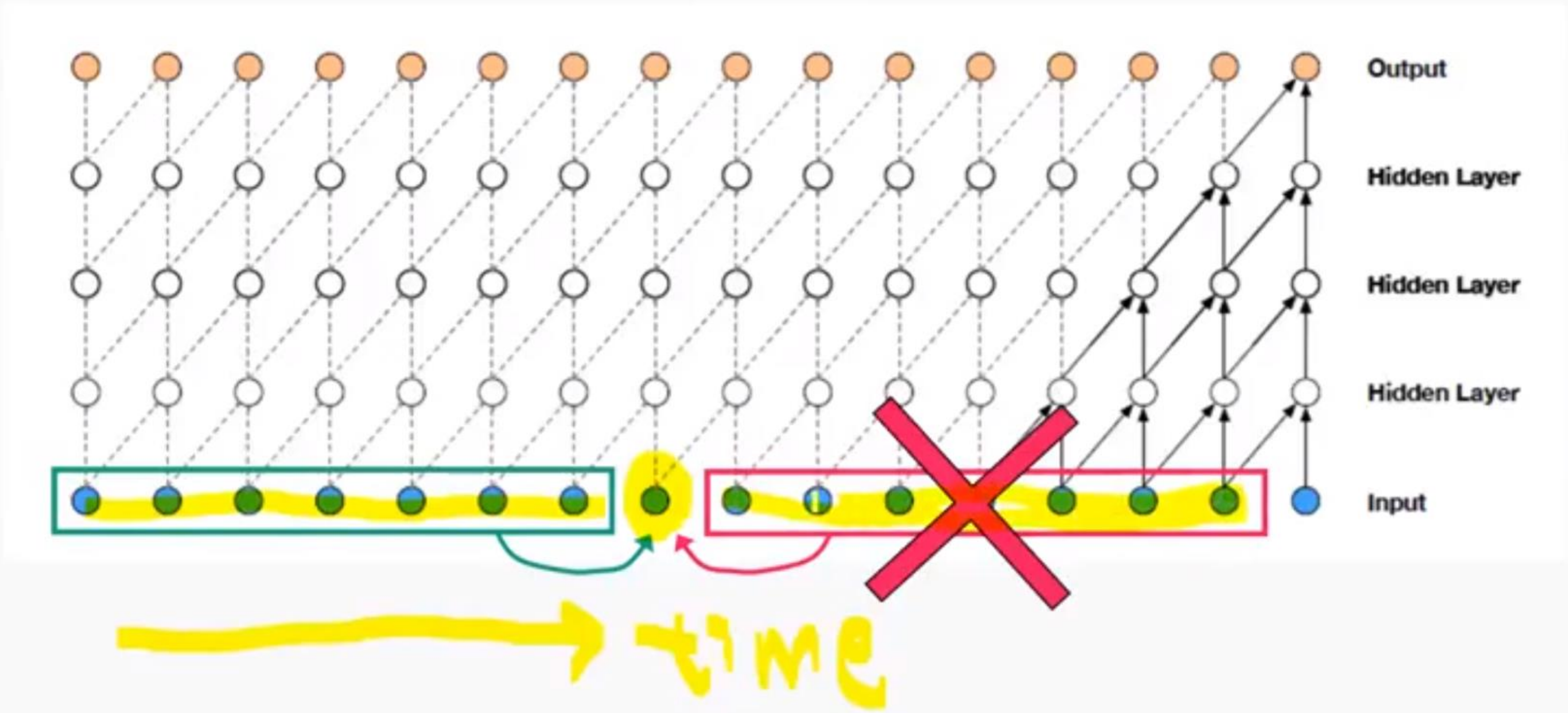
p(x1)
p(x1,x2)=p(x1)p(x2|x1)
p(x1,x2,x3)=p(x1,x2)p(x3|x1,x2)
=p(x1)p(x2|x1)p(x3|x1,x2)
...

→ 이것을 Stack of Convolution Layer로 표현

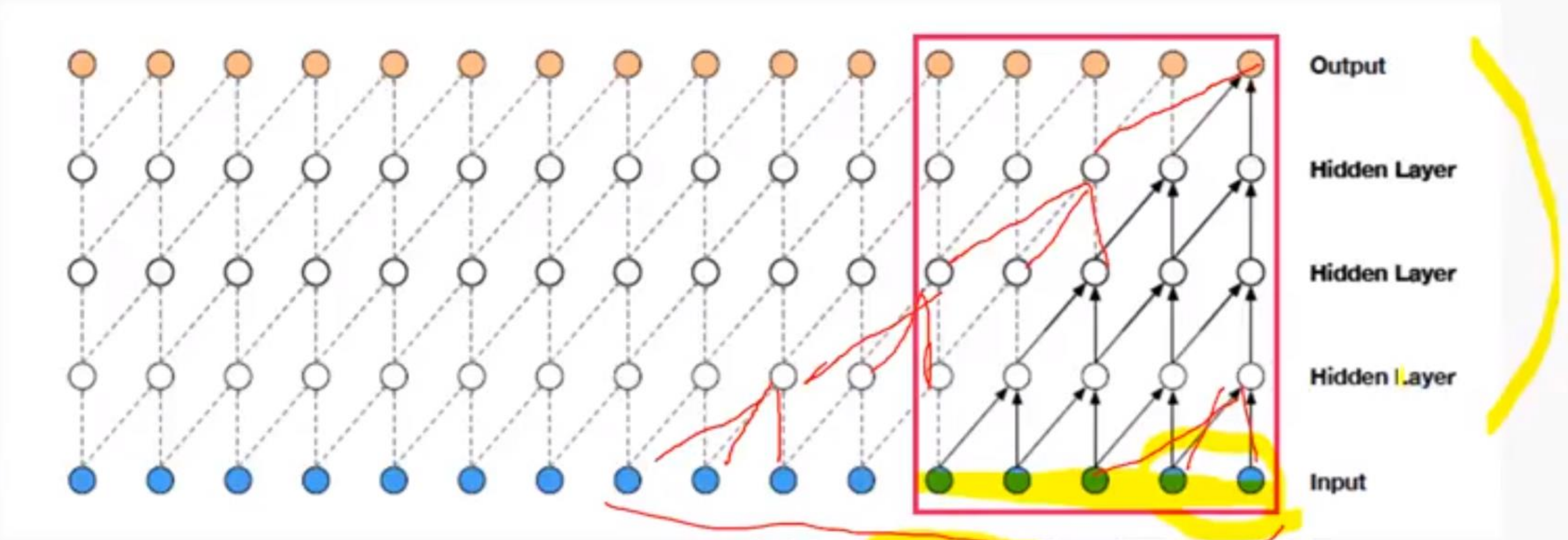
Stack of Causal Convolutional Layers



Stack of "Causal" Convolutional Layers

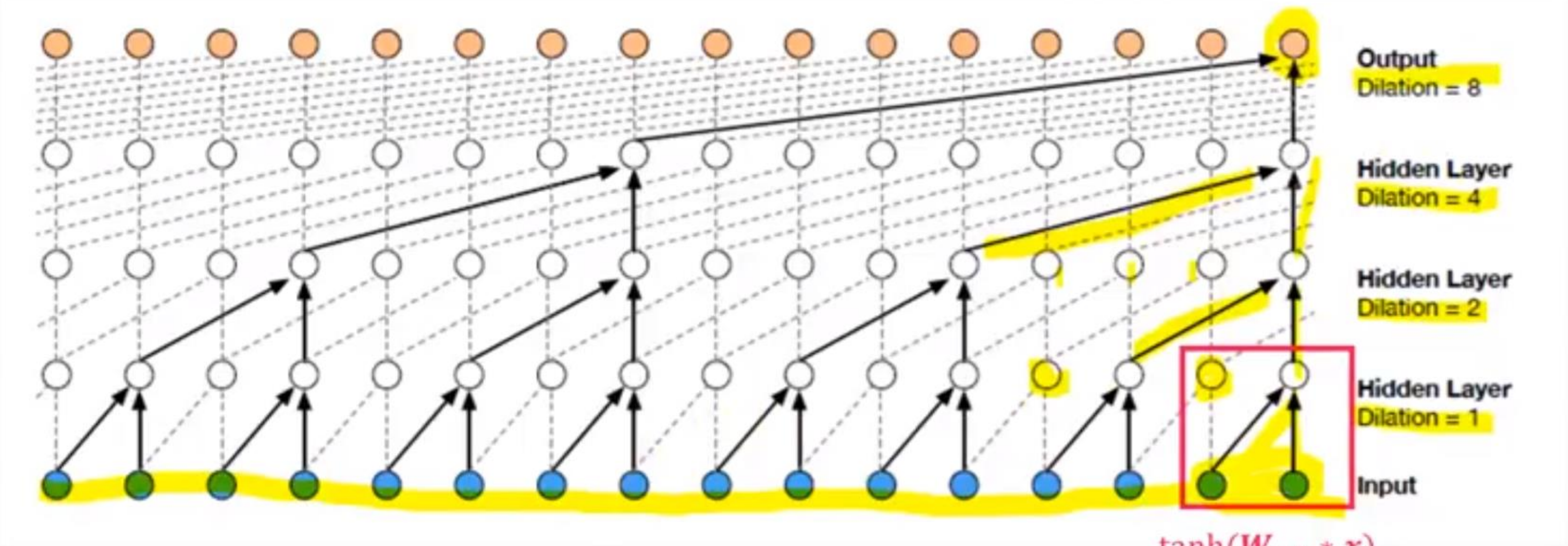


Stack of Causal Convolutional Layers



Receptive field= #layers + filter length -1 (?)
가 너무 좁다. 여기서는 5.
이것을 키우기 위해서는 너무 많은 Layer가 필요.

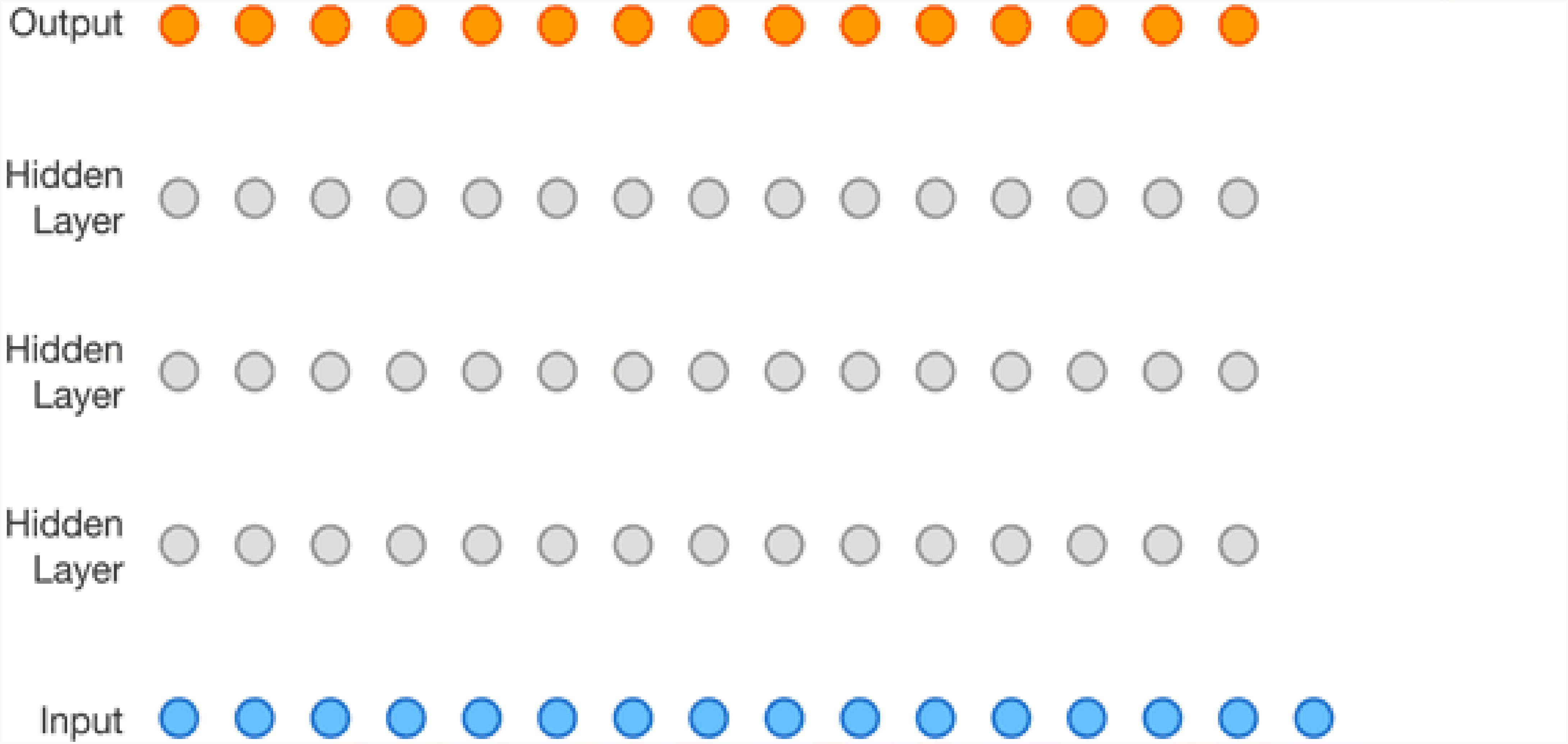
Stack of Dilated Causal Convolutional Layers



$\tanh(\boldsymbol{W}_{x,k} * \boldsymbol{x})$

Receptive field= (1+2+4+8)+1=16

Stack of Dilated Causal Convolutional Layers



Stack of Dilated Causal Convolutional Layers

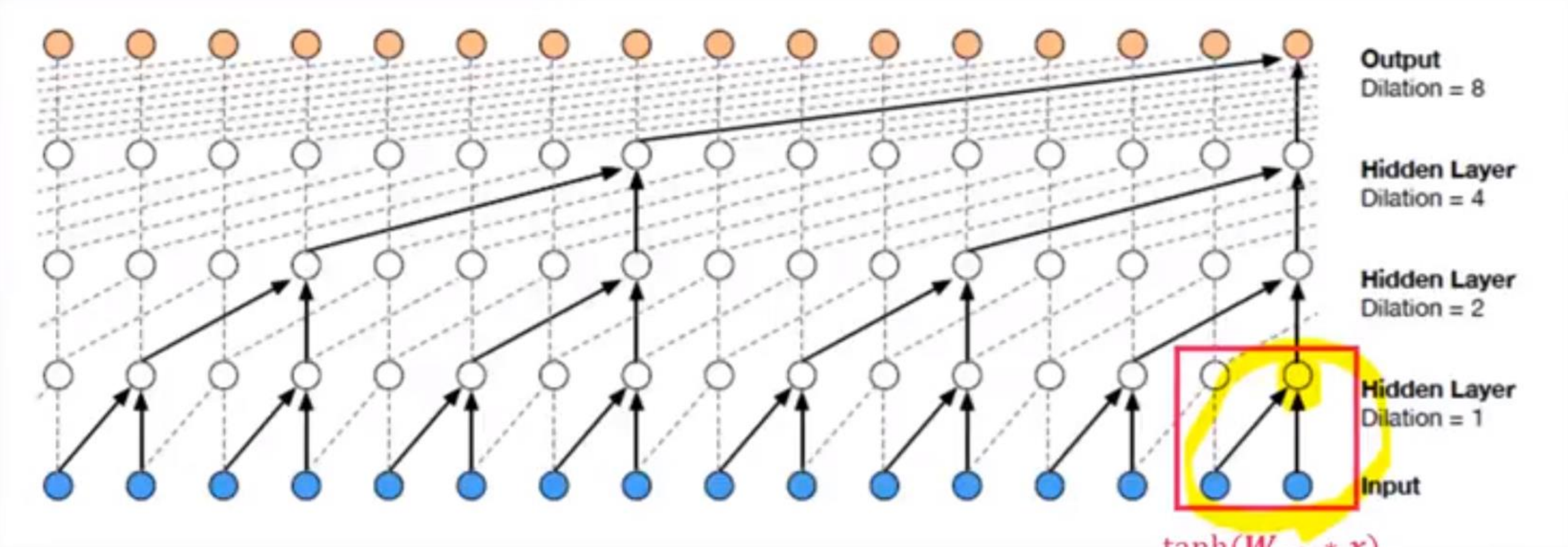
$$1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512.$$

→ 30 Layers

Softmax Distributions

- 이 논문에서 conditional probability 를 modeling하는데 있어서, softmax distributions 을 사용함.
- Audio 신호는 16bit로 quantization 하는 경우가 많음
  이걸 softmax로 표현하려면 sample마다 65536 개의 output이 필요. (너무 많다.)
- mu-law companding 기법을 사용.
  사람의 귀는 소리 크기가 작을 때는 작은 변화에도 민감
  소리 크기가 클 때는 비교적 큰 변화에도 둔감함.
  → quantization을 nonlinear하게 해줌. 이렇게 하면 8bit(256 outputs)로도 꽤 좋은 성능으로 encoding/decoding이 가능

Gated Activation Units

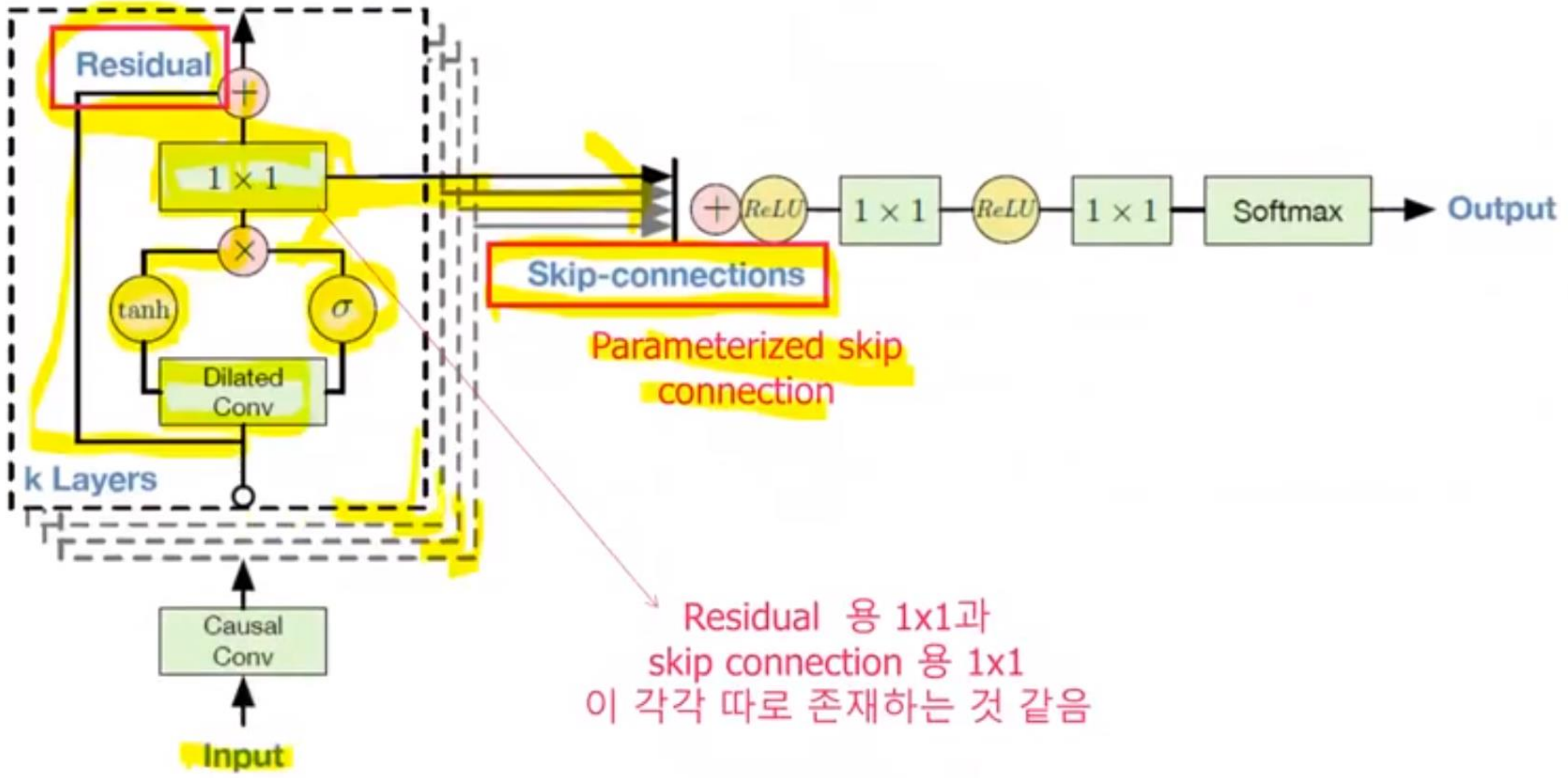


$\tanh(\boldsymbol{W}_{x,k} * \boldsymbol{x})$
에 gate units을 추가

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \boxed{\sigma\left(W_{g,k} * \mathbf{x}\right)}$$

gate units

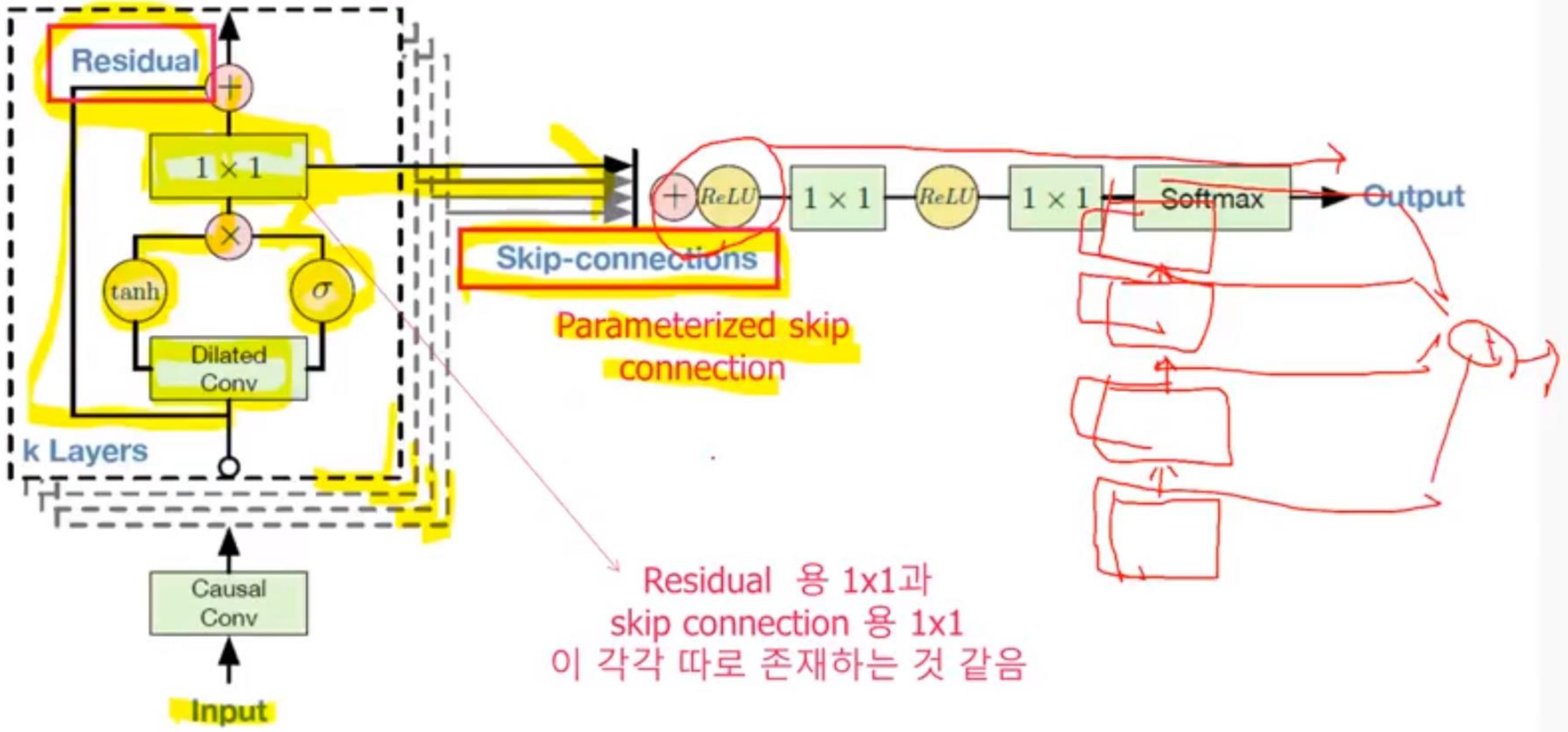
Residual and Skip Connections



Parameterized skip connection

Residual 용 1x1과
skip connection 용 1x1
이 각각 따로 존재하는 것 같음

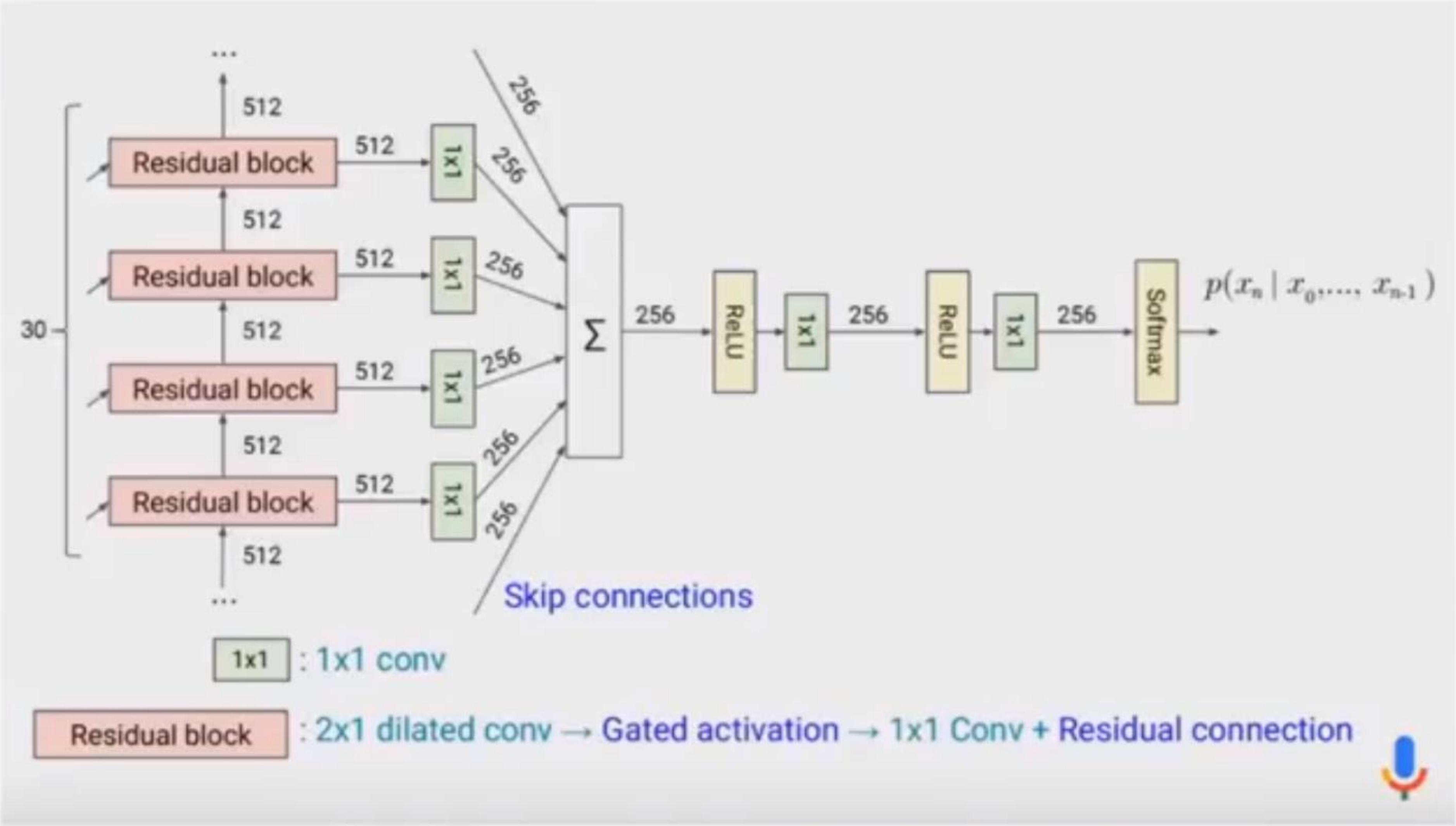
Residual and Skip Connections



Parameterized skip connection

Residual 용 1x1과
skip connection 용 1x1
이 각각 따로 존재하는 것 같음

WaveNet – Architecture



Heiga Zen, https://www.youtube.com/watch?v=nsrSrYtKkT8

Conditional WaveNets

We can guide WaveNet's
generation to produce audio
with the required characteristics

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}).$$

Ex1. In a multi-speaker setting : Speaker identity
Ex2. TTS : text

Conditional WaveNets : Global Conditioning

Global Conditioning is characterized by a single latent representation **h**
that influences the
output distribution across all timesteps

[Ex] Speaker Identity

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^{T}\mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^{T}\mathbf{h}\right).$$

Learnable linear projection

Condition 을 addition 으로..

Conditional WaveNets : Local Conditioning

For Local Conditioning, we have a second timeseries $\mathbf{h}_t$, possibly with a lower sampling frequency

[Ex] Linguistic Feature in a TTS model

1. Upsampling by transposed convolutional network

$$\mathbf{y} = f(\mathbf{h})$$

2. 1x1 convolution in activation unit

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right),$$

1x1 convolution