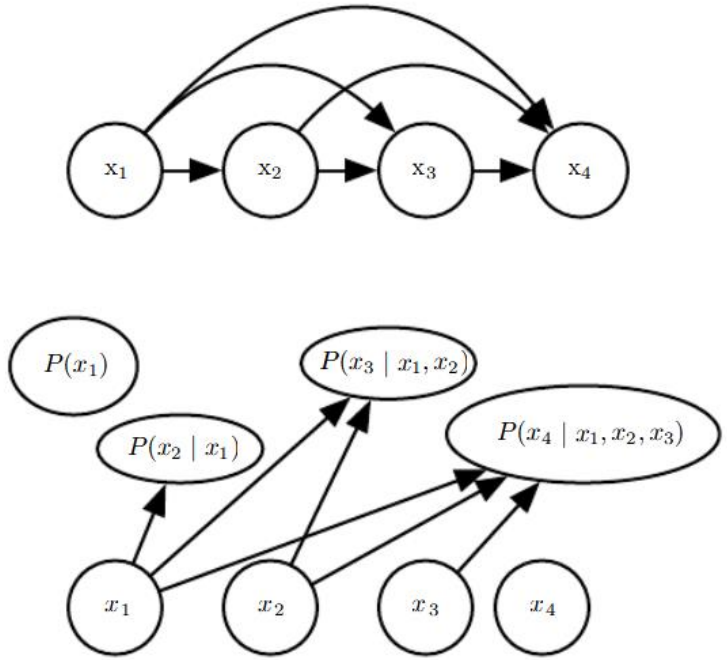
Chapter 06. 무엇이든 진짜처럼 생성하는 생성 모델(Generative Networks)

Autoregressive Model

자동회귀 모델 AR; Autoregressive Model



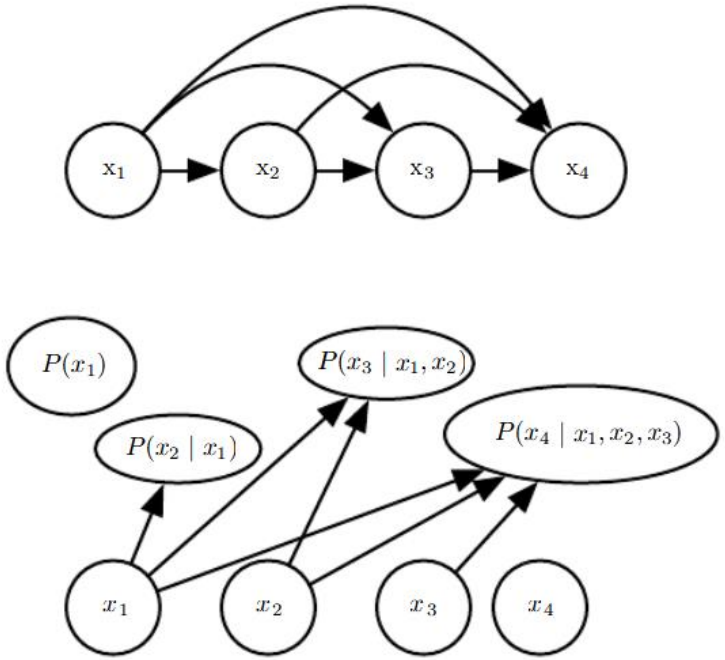
AR Model

Figure 1: A second of generated speech.

목소리를 모사하는 WaveNets

AR 모델을 이용한 Google DeepMind의 PixelCNN과 WaveNet을 알아보자.

AR Equations



AR Model

$$AR(p) \text{ 모델 : } x_t = c + \sum_{i=1}^{p} \varphi_i x_{t-i} + \varepsilon_t$$

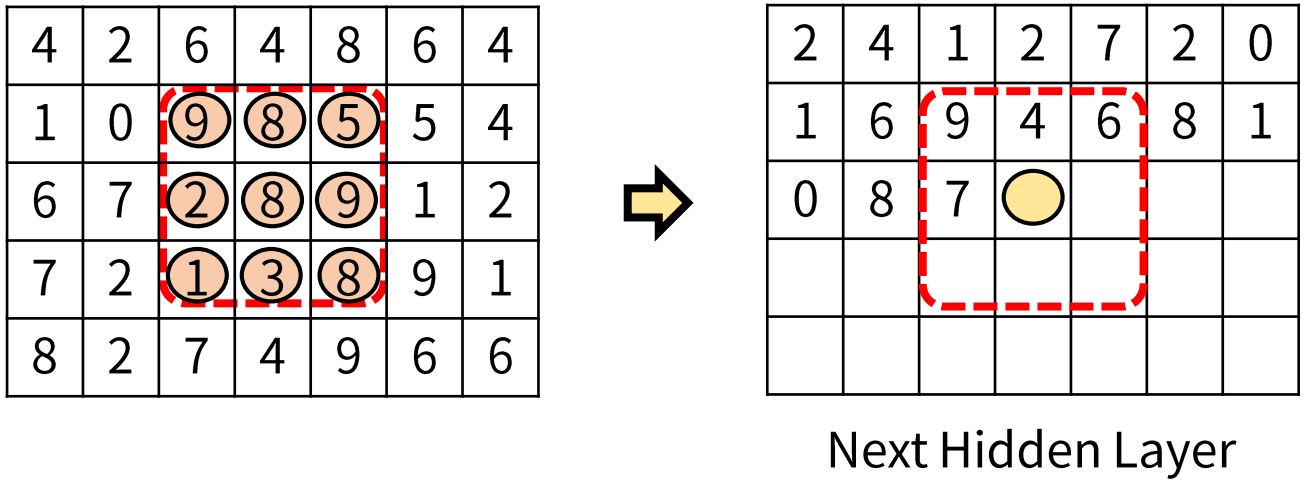
- x_t : Random Process
- c : Constant(Bias)
- φ_i : 모델 파라미터
- ε_t : White noise

결국 p개의 이전 샘플을 이용해 Linear Regression하여 현재 샘플을 예측하는 것과 동일하다.

수식도 그리 복잡하지 않다. 이 모델을 기반으로 Generative Model을 만들어 보자.

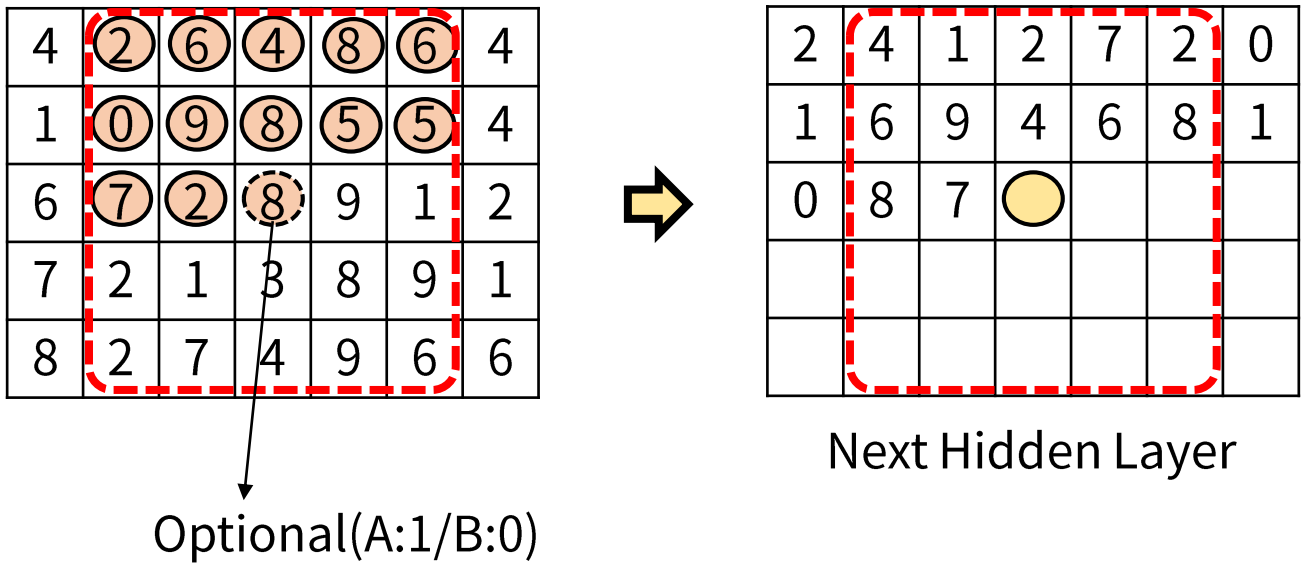
Convolutional Layer

$$h_{i,j}^{n} = c + \sum_{k=-N}^{N} \sum_{l=-M}^{M} \varphi_{k,l}^{n} h_{i-k,j-l}^{n-1}$$



뜬금없이 CNN이 나왔다. 그런데 왠지 AR Model과 상당히 닮아 있다.

Masked Convolution



일부 값을 Masking하여 사용하면 AR 모델처럼 '이전' 값들만 참조할 수 있다.

Multi-Scale Case

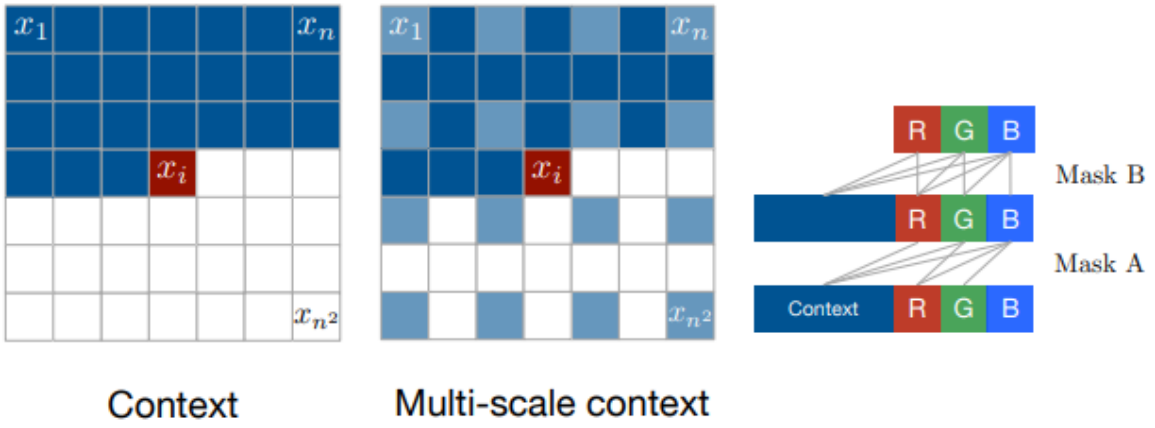


Figure 2. **Left**: To generate pixel x_i one conditions on all the previously generated pixels left and above of x_i. **Center**: To generate a pixel in the multi-scale case we can also condition on the subsampled image pixels (in light blue). **Right**: Diagram of the connectivity inside a masked convolution. In the first layer, each of the RGB channels is connected to previous channels and to the context, but is not connected to itself. In subsequent layers, the channels are also connected to themselves.

Residual Connections

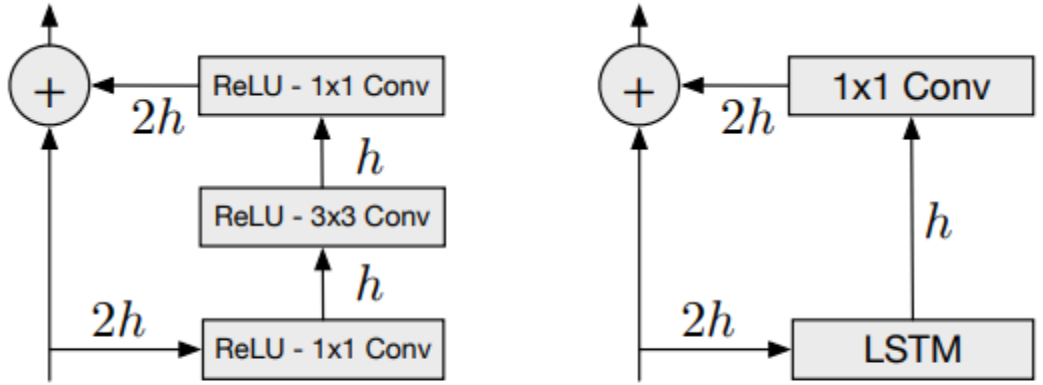
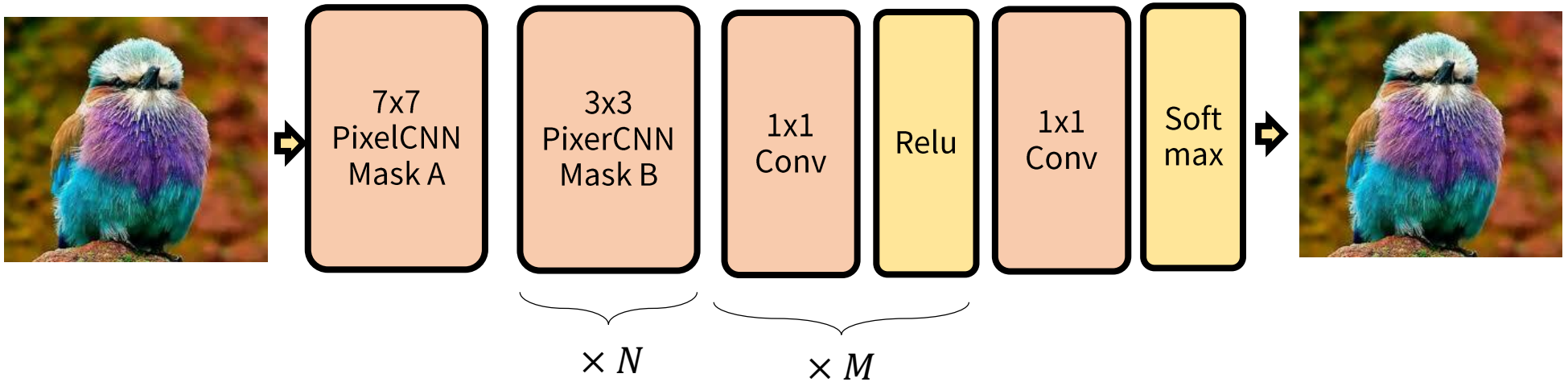


Figure 5. Residual blocks for a PixelCNN (left) and PixelRNNs.

ResNet과 유사하게 Pre-Activation을 이용하여 Layer를 구성하고 있다.

PixelCNN



Unsupervised Learning 형태이므로, Autoencoder의 형태와 상당히 유사하다.

PixelCNN - Results

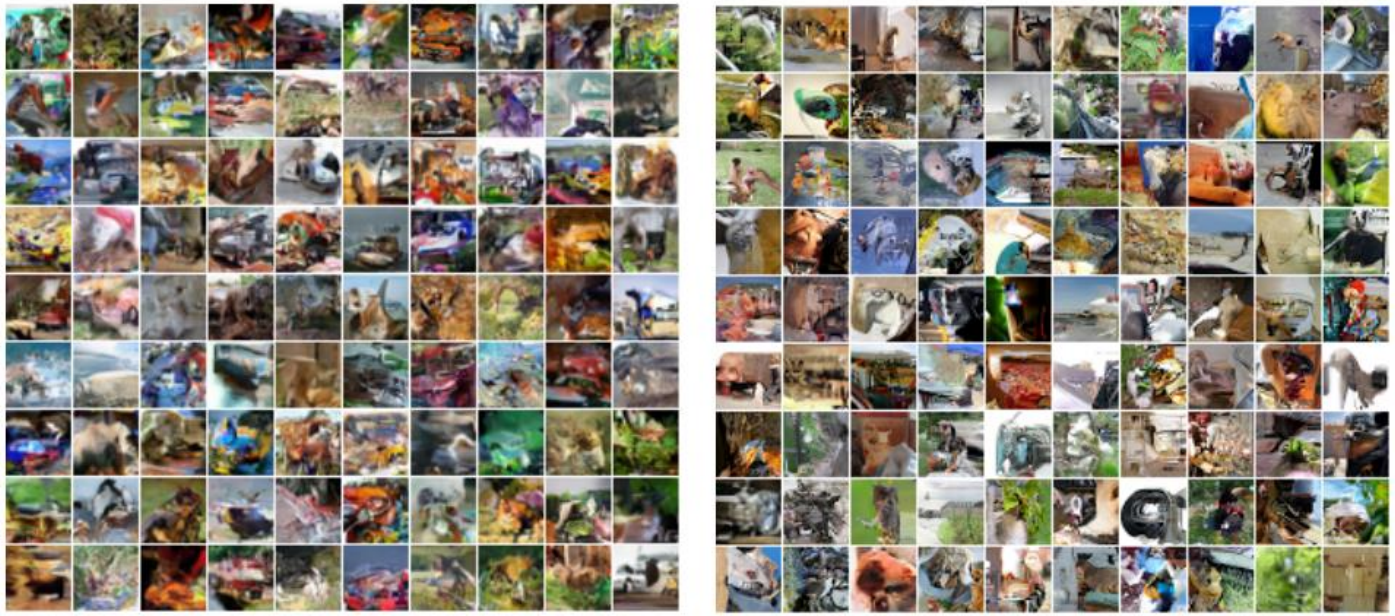
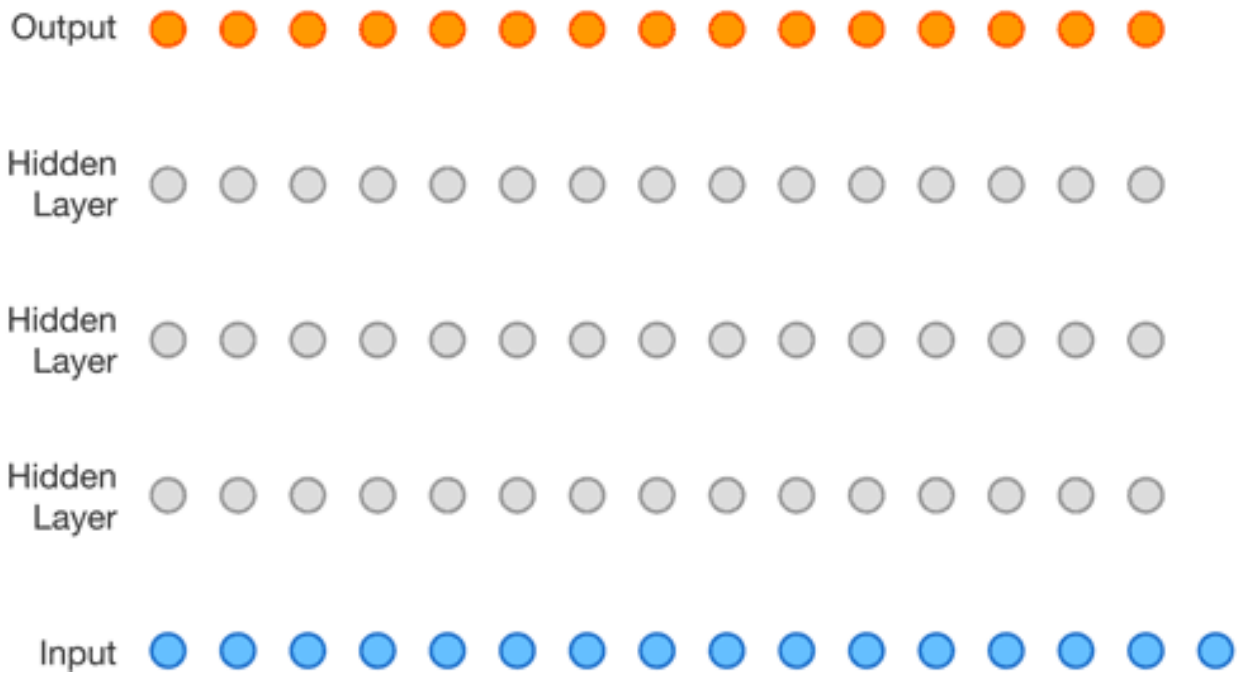


Figure 7. Samples from models trained on CIFAR-10 (left) and ImageNet 32x32 (right) images. In general we can see that the models capture local spatial dependencies relatively well. The ImageNet model seems to be better at capturing more global structures than the CIFAR-10 model. The ImageNet model was larger and trained on much more data, which explains the qualitative difference in samples.

작게 보면 그럭저럭 Natural Image 같지만 뭔가 부족하다...

이 PixelCNN 아이디어를 다른 곳에 또 쓸 수 있을까?

1-D WaveNet



AR의 본거지인 1-D Signal에 다시 PixelCNN을 적용해 보자.

Implementation Details (1/2)

Because raw audio is typically stored as a sequence of 16-bit integer values (one per timestep), a softmax layer would need to output 65,536 probabilities per timestep to model all possible values. To make this more tractable, we first apply a μ-law companding transformation (ITU-T, 1988) to the data, and then quantize it to 256 possible values:

$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu|x_t|)}{\ln(1 + \mu)}$$

where $-1 < x_t < 1$ and $\mu = 255$. This non-linear quantization produces a significantly better reconstruction than a simple linear quantization scheme. Especially for speech, we found that the reconstructed signal after quantization sounded very similar to the original.

2.3 Gated Activation Units

We use the same gated activation unit as used in the gated PixelCNN (van den Oord et al., 2016b):

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \odot \sigma(W_{g,k} * \mathbf{x}), \quad (2)$$

where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W is a learnable convolution filter. In our initial experiments, we observed that this non-linearity worked significantly better than the rectified linear activation function (Nair & Hinton, 2010) for modeling audio signals.

Implementation Details (2/2)

2.4 RESIDUAL AND SKIP CONNECTIONS

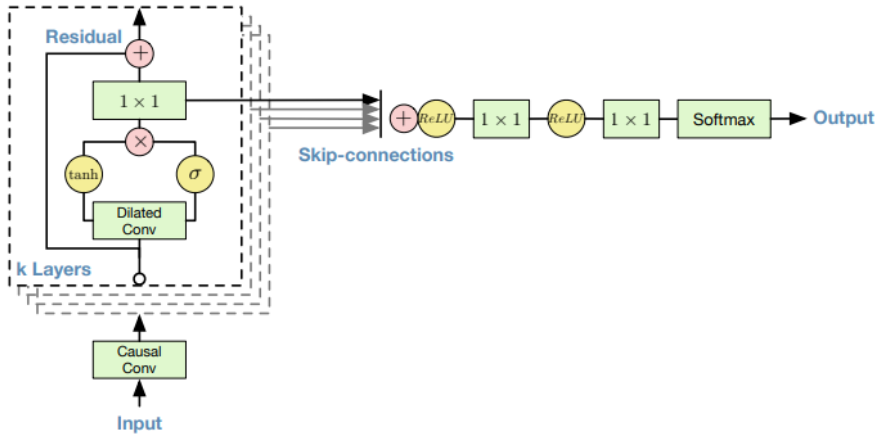
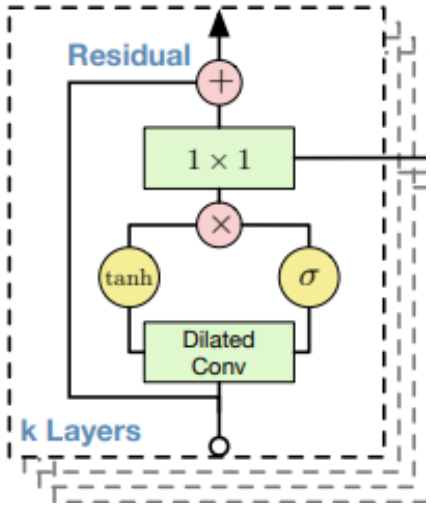
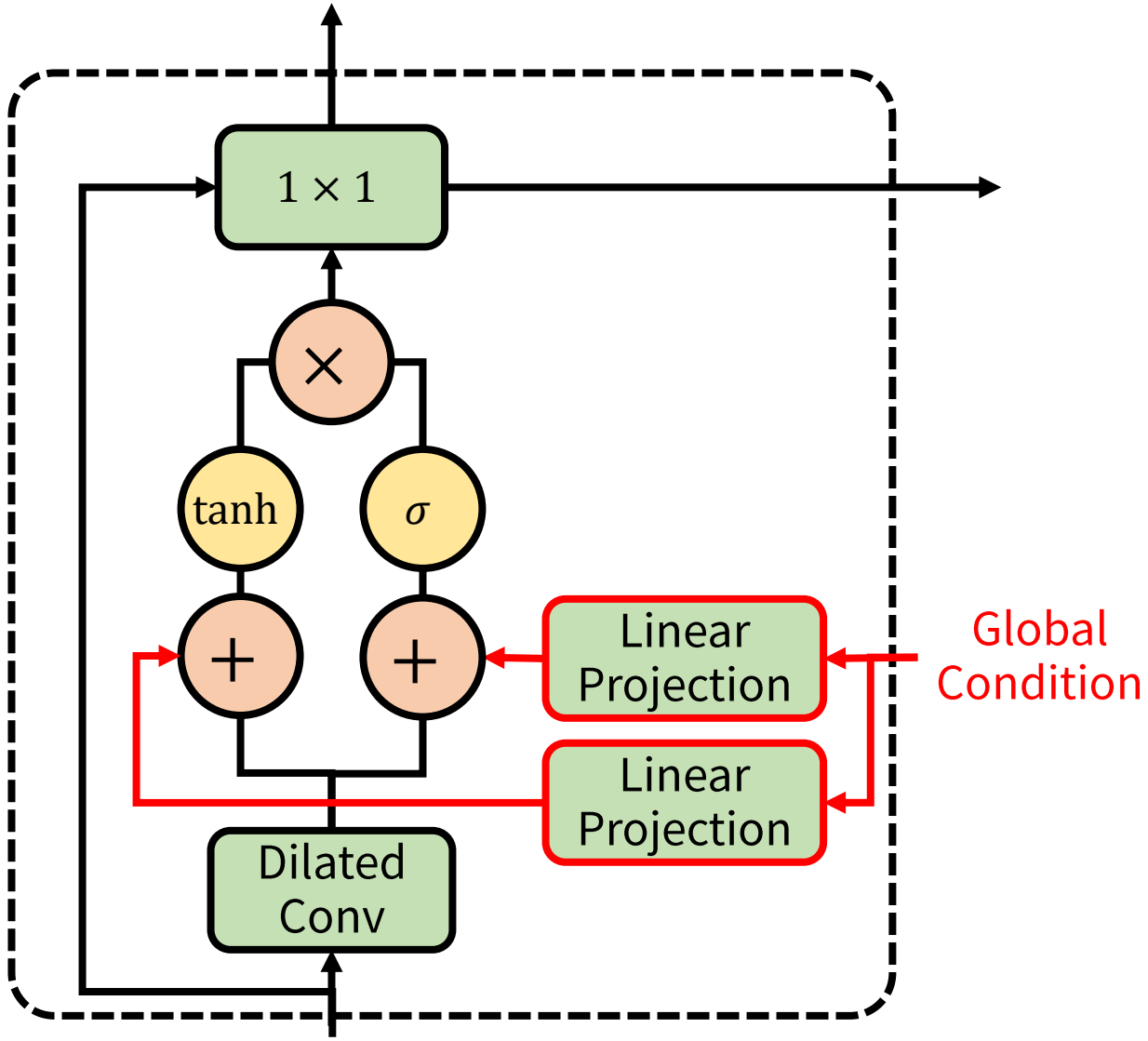


Figure 4: Overview of the residual block and the entire architecture.

Both residual (He et al., 2015) and parameterised skip connections are used throughout the network, to speed up convergence and enable training of much deeper models. In Fig. 4 we show a residual block of our model, which is stacked many times in the network.

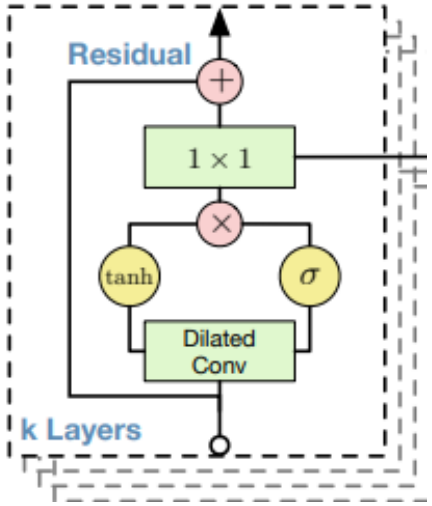
Global Condition

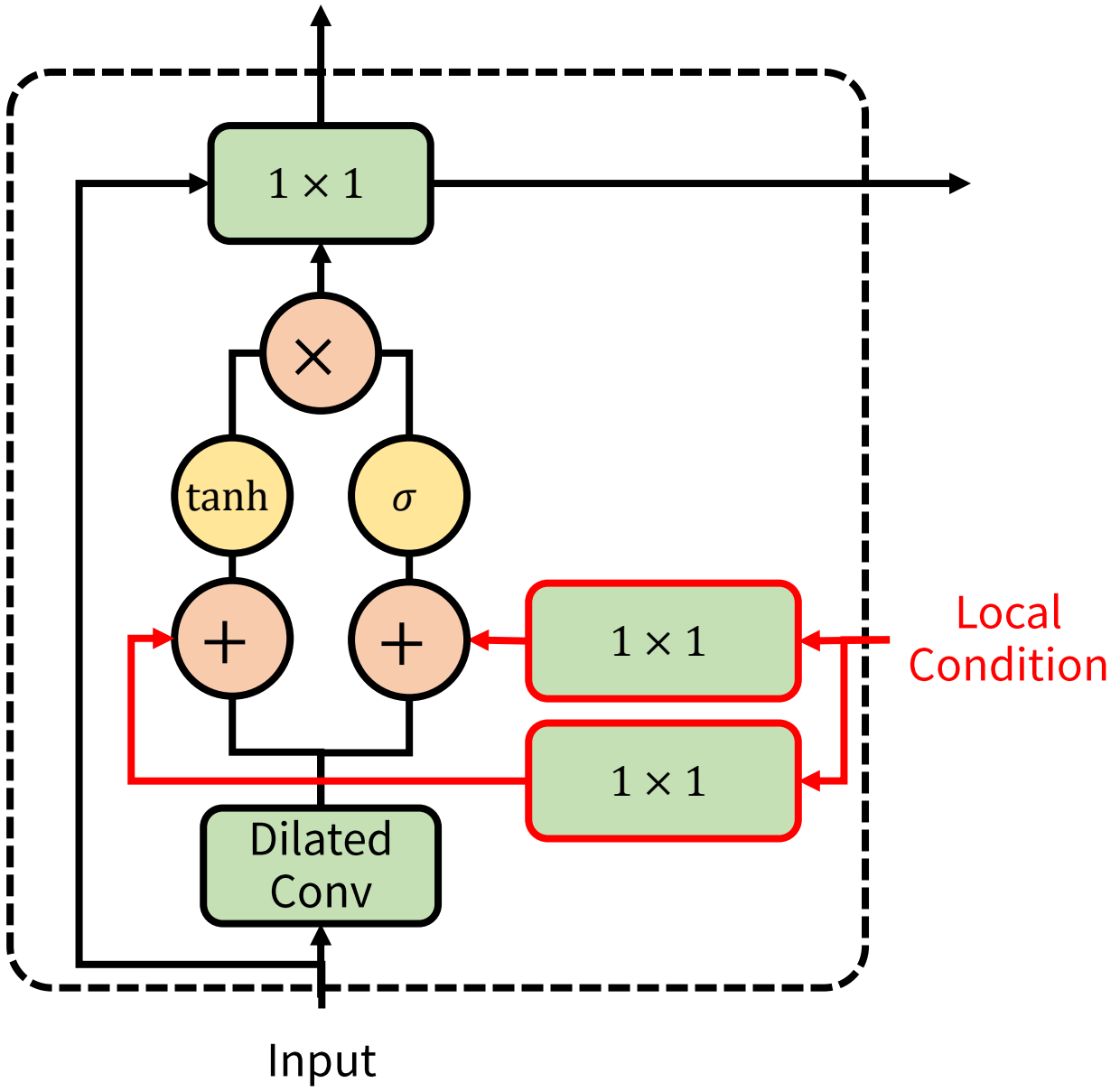




화자, 감정상태 등 한 문장을 생성하는 Global Condition을 설정할 수 있다.

Local Condition





문장의 Text 정보와 같은 Local Condition을 지정하여 학습할 수 있다.