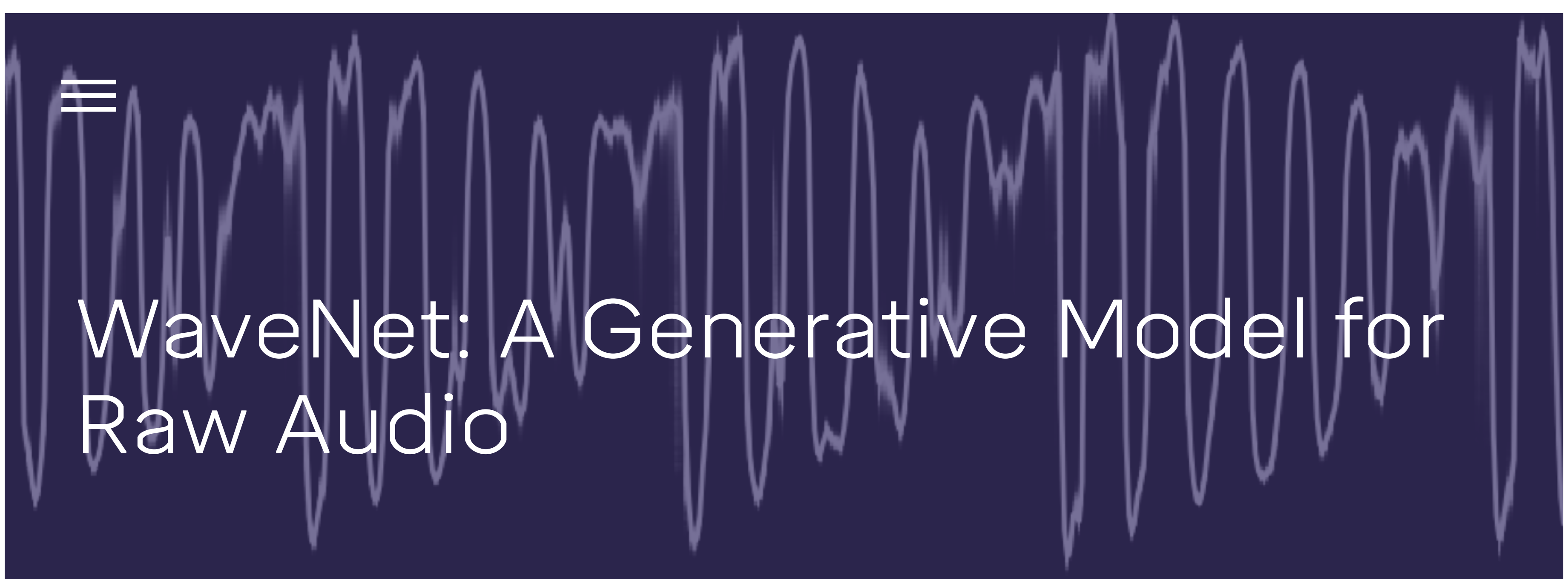
WaveNet

Steve Renals
Automatic Speech Recognition – ASR Lecture 19
30 March 2017

A van den Oord et al, “WaveNet: A Generative Model for Raw Audio”, arXiv:1609.03499



Motivation

“Researchers usually avoid modelling raw audio because it ticks so quickly: typically 16,000 samples per second or more, with important structure at many time-scales.

“Building a completely autoregressive model, in which the prediction for every one of those samples is influenced by all previous ones (in statistics-speak, each predictive distribution is conditioned on all previous observations), is clearly a challenging task.”

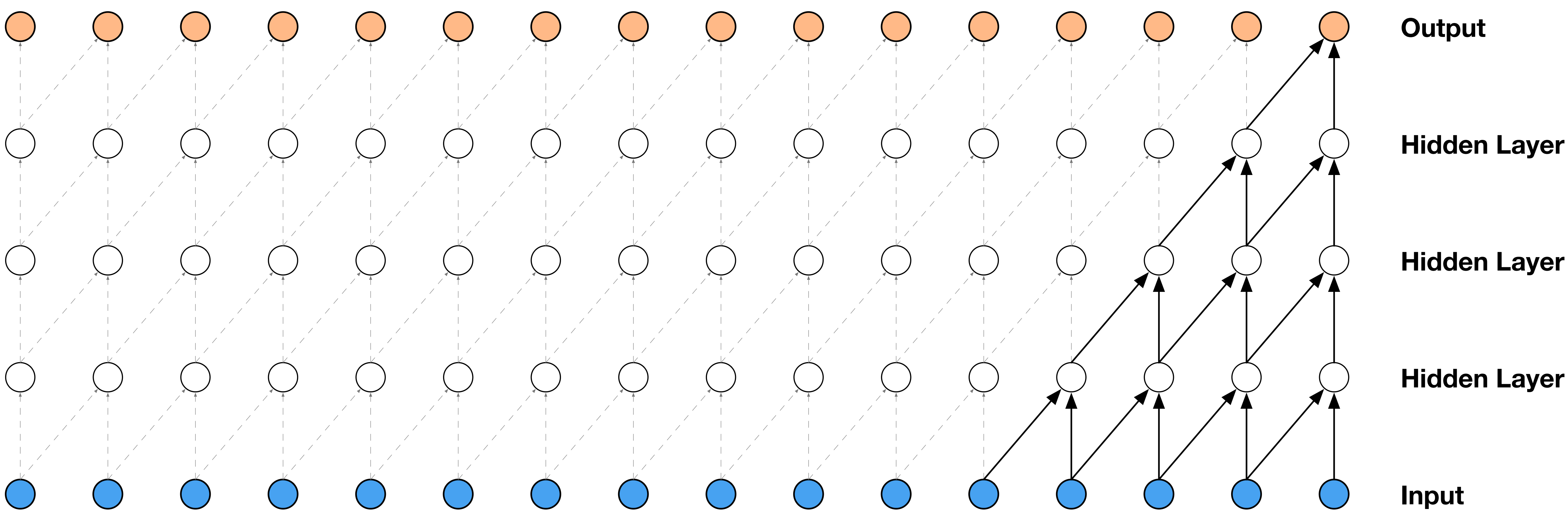
WaveNet Approach

- Generative model operating directly on the raw waveform

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

- WaveNet model is *probabilistic* and *autoregressive*
- Model using a deep stack of convolutional layers
- No pooling layers – output has same dimensionality as input

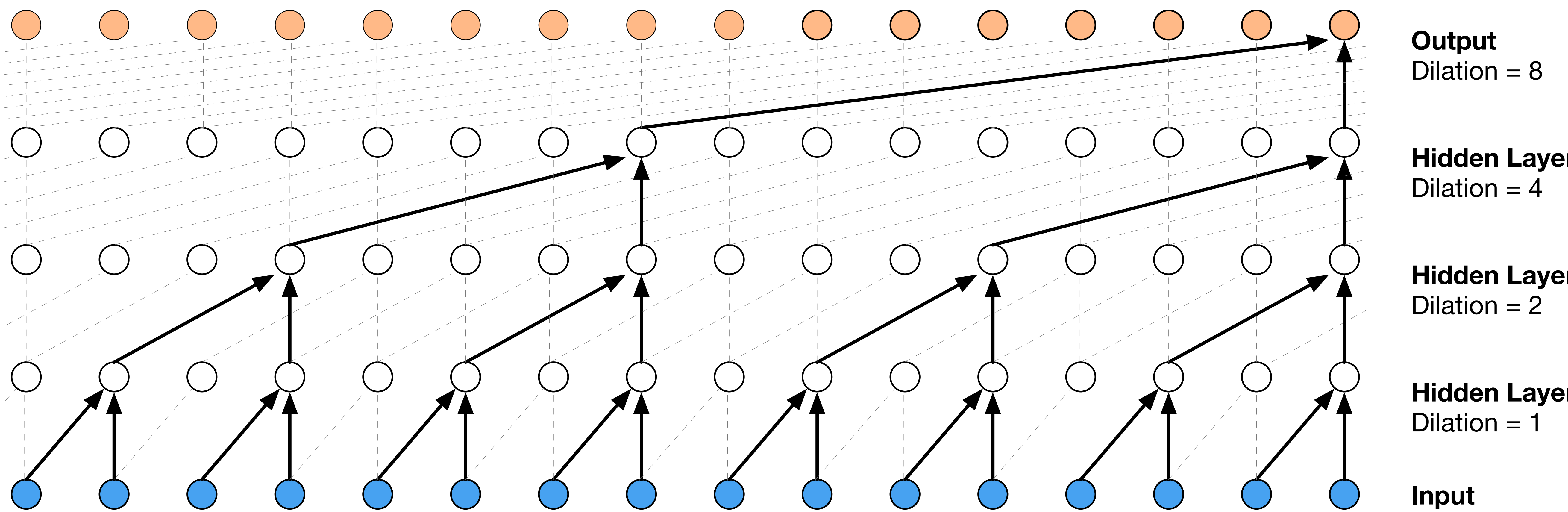
Causal convolutions



Efficiency

- Training: predictions can be made in parallel, because all timesteps of the ground truth training data **x** are known
- Generating: predictions are sequential, each predicted sample is used as part of the context for future samples
- Sequence modelling done by stacked convolutions
  - CNN more efficient that RNN (no backprop through time)
  - Many layers needed for long temporal context
  - Dilated convolutions increase the context

Dilated causal convolutions



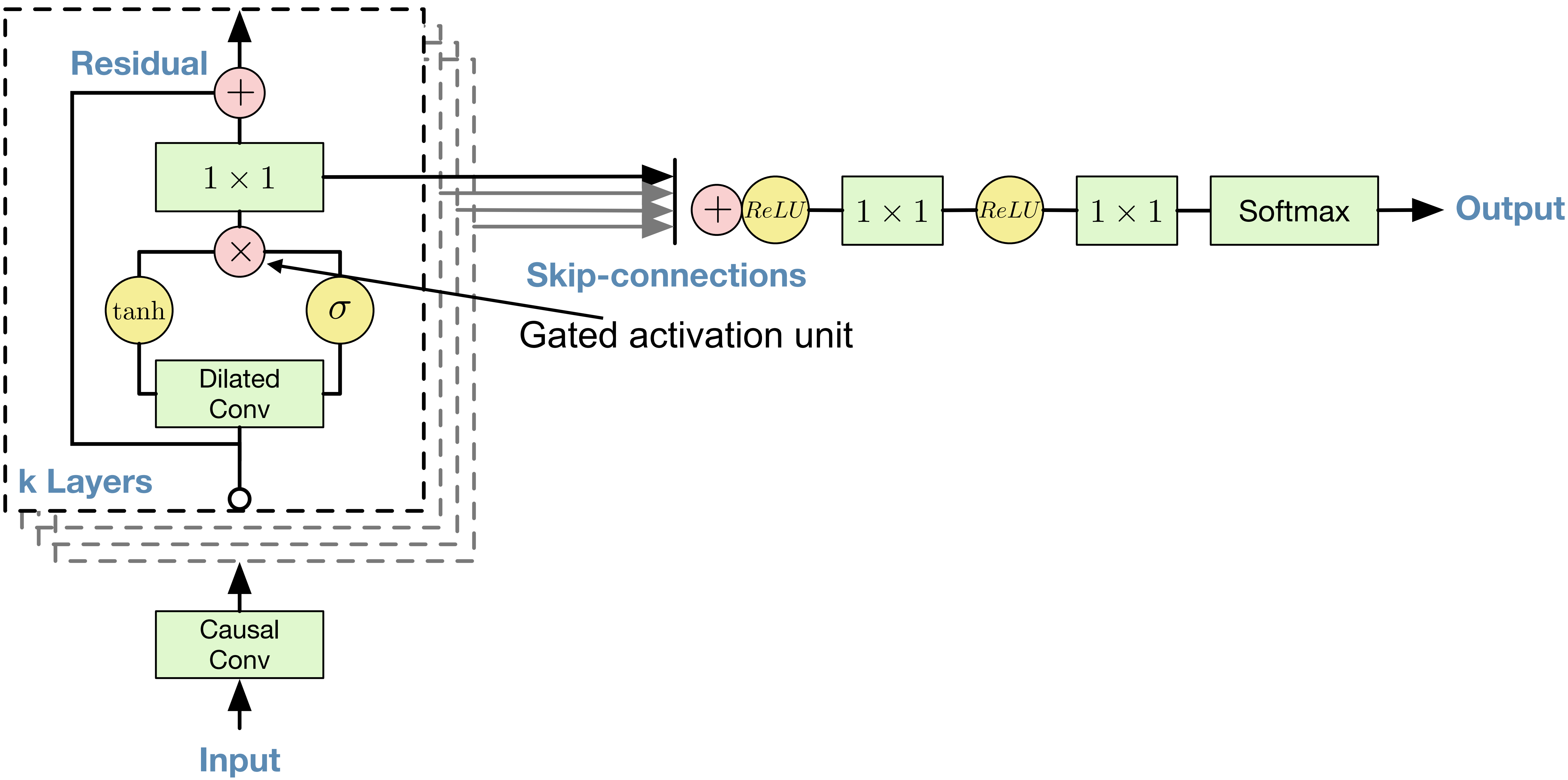
In WaveNet dilations increase to a limit, then repeated:
1,2,4,…,512,1,2,4,…,512,1,2,4,…512

Each 1,2,4,…,512 block has a context of 1024
– more efficient and discriminative than a 1024-convolution

WaveNet Output

- Use a softmax distribution to model the outputs – but if sample is x is 16 bits, then we would have 65,536 outputs
  - 8-bit sample coding using μ-law compression
  - 256 outputs
- This is like a “language model” for audio samples

Residual/skip connections



Control: Conditional WaveNets

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}, \mathbf{h})$$

- By conditioning the model on other variables can control the characteristics of generated audio
  - crucial for speech synthesis
  - for multi-speaker modelling, **h** could encode speaker identity

WaveNet Generation

- Free-form speech generation
  - WaveNet conditioned on speaker identity
  - Trained on 44h speech from 109 speakers
- Text-to-speech synthesis
  - locally conditioned on linguistic features and log F0
  - trained on multispeaker data, conditioned on speaker identity

WaveNet for Speech Recognition

- Use WaveNet as learned front end to ASR neural network

- Mean pooling layer after the dilated convolutions

    - aggregate to 10ms frames (mean-pooling)

    - followed by a “few non-causal convolutions”

    - multi-task training to simultaneously predict the next sample and classify the frame

    - 18.6% PER on TIMIT

The End.