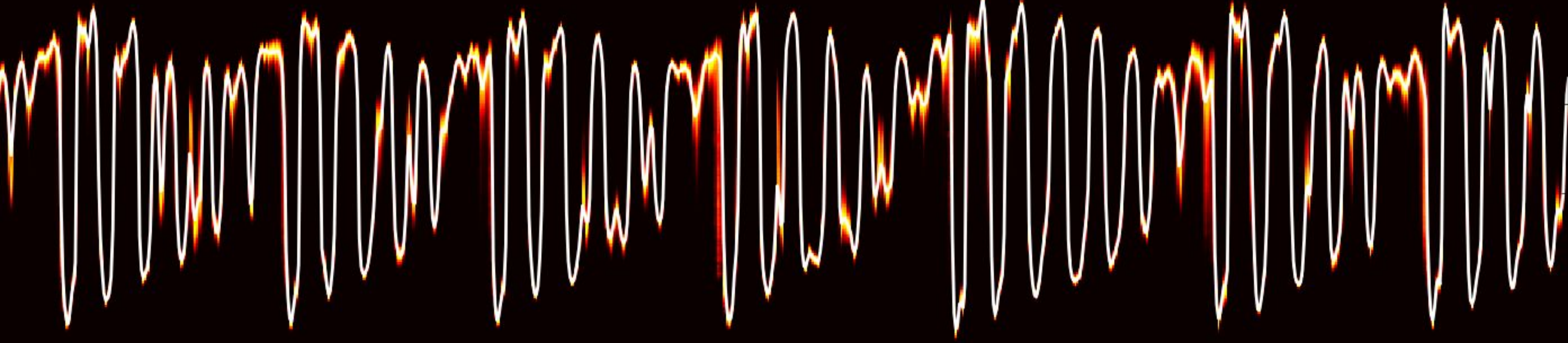


Neural Discrete Representation Learning

Aaron van den Oord, Oriol Vinyals, Koray Kavukcuoglu

Google DeepMind

Generative Models

Goal: Estimate the <u>probability distribution</u> of high-dimensional data

Such as images, audio, video, text, ...

Motivation:

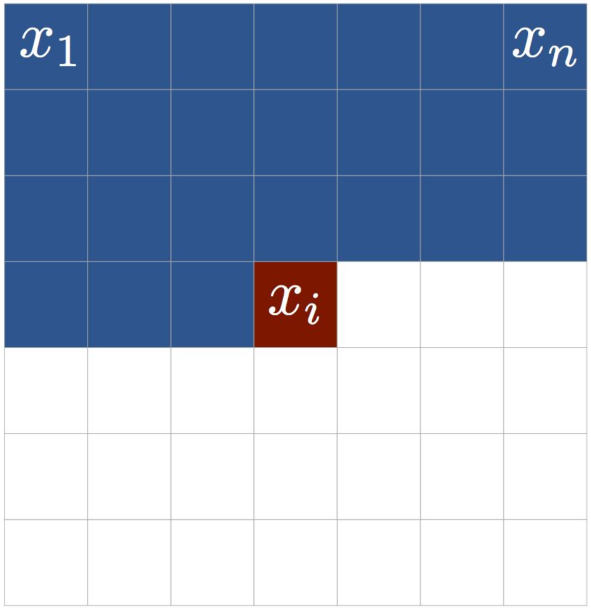
Learn the underlying structure in data.

Capture the dependencies between the variables.

Generate new data with similar properties.

Learn useful features from the data in an unsupervised fashion.

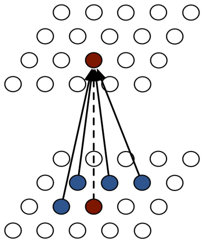
Autoregressive Models



$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, ..., x_{i-1})$$

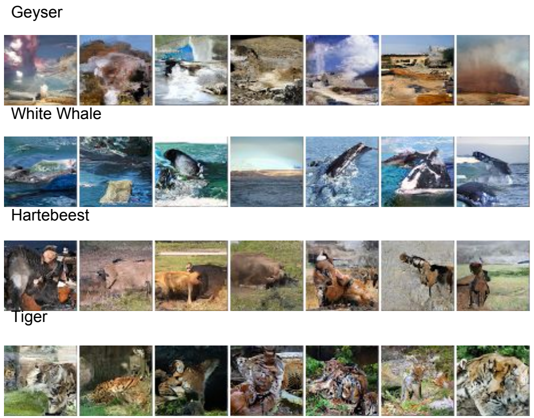
Google DeepMind

Recent Autoregressive models at DeepMind



PixelRNN

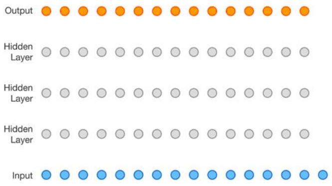
PixelCNN

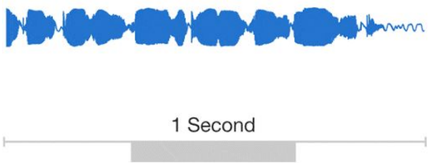


van den Oord et al, 2016ab

Video Pixel Networks

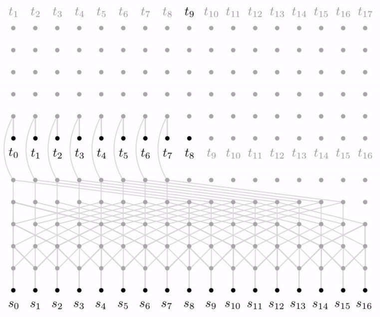
Kalchbrenner et al, 2016a





WaveNet

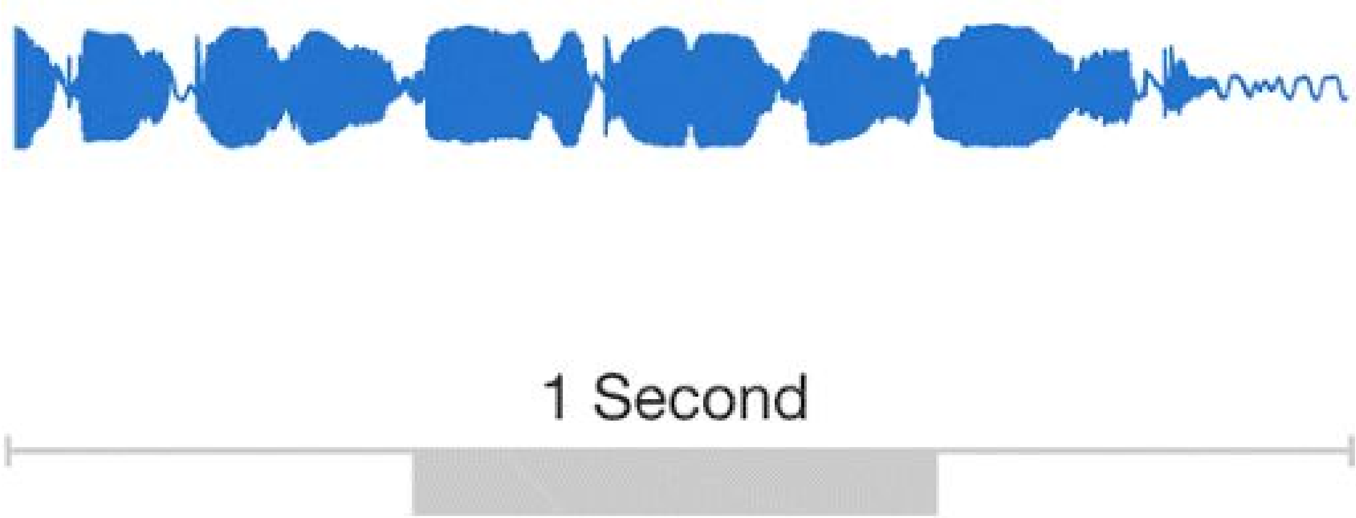
van den Oord et al, 2016c



ByteNet

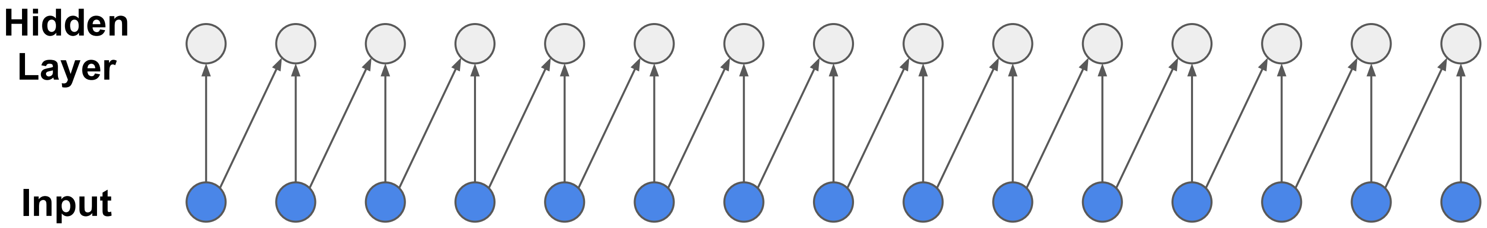
Kalchbrenner et al, 2016b

Modeling Audio

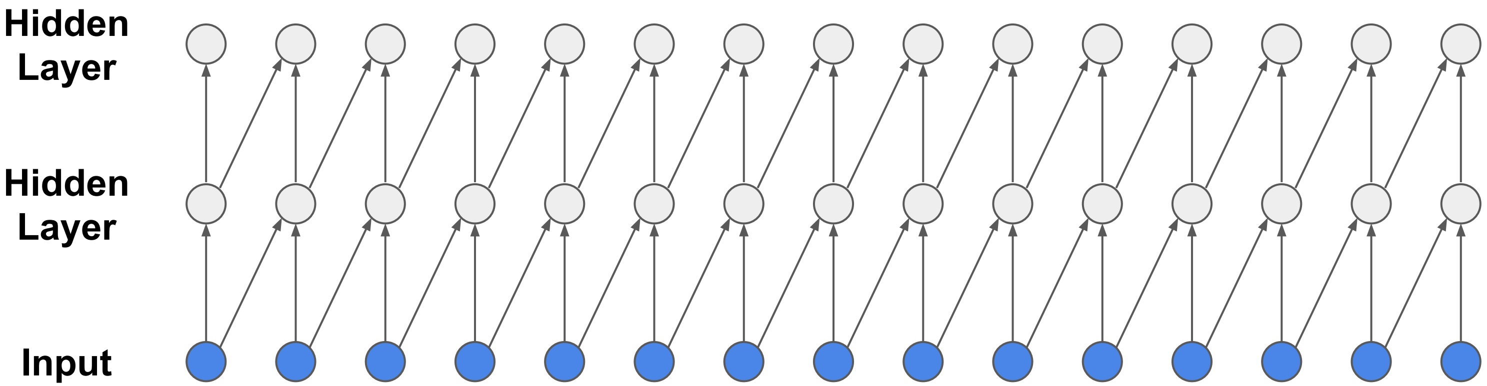


Google DeepMind

Causal Convolution

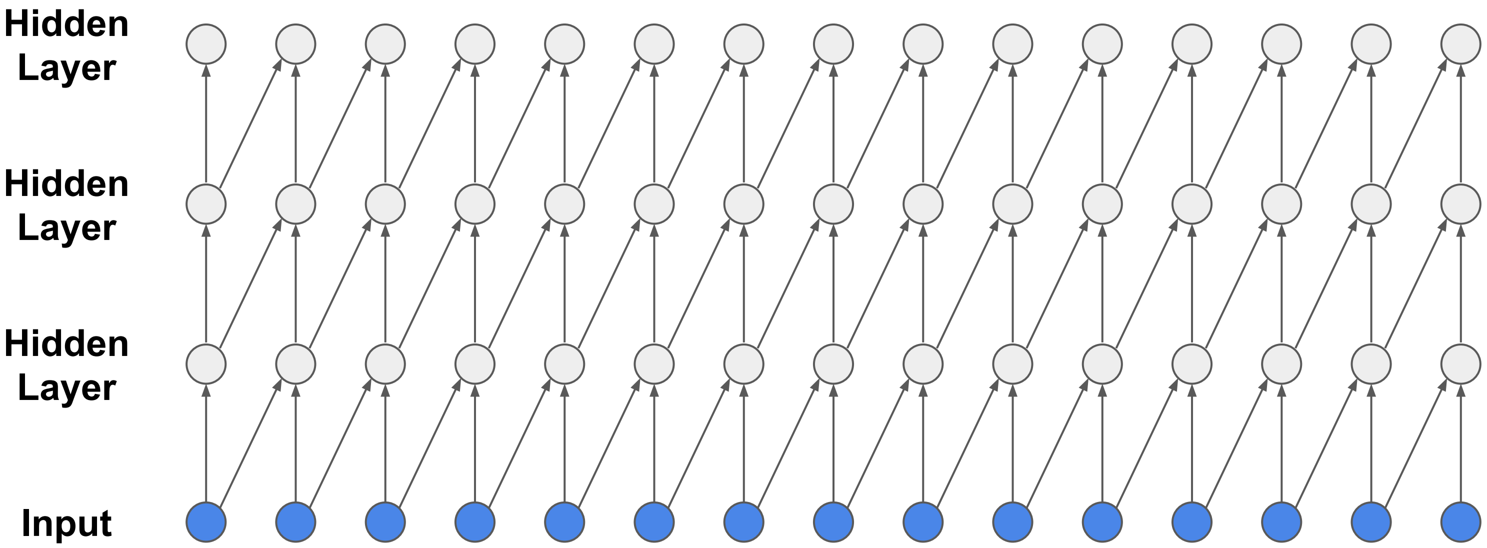


Causal Convolution

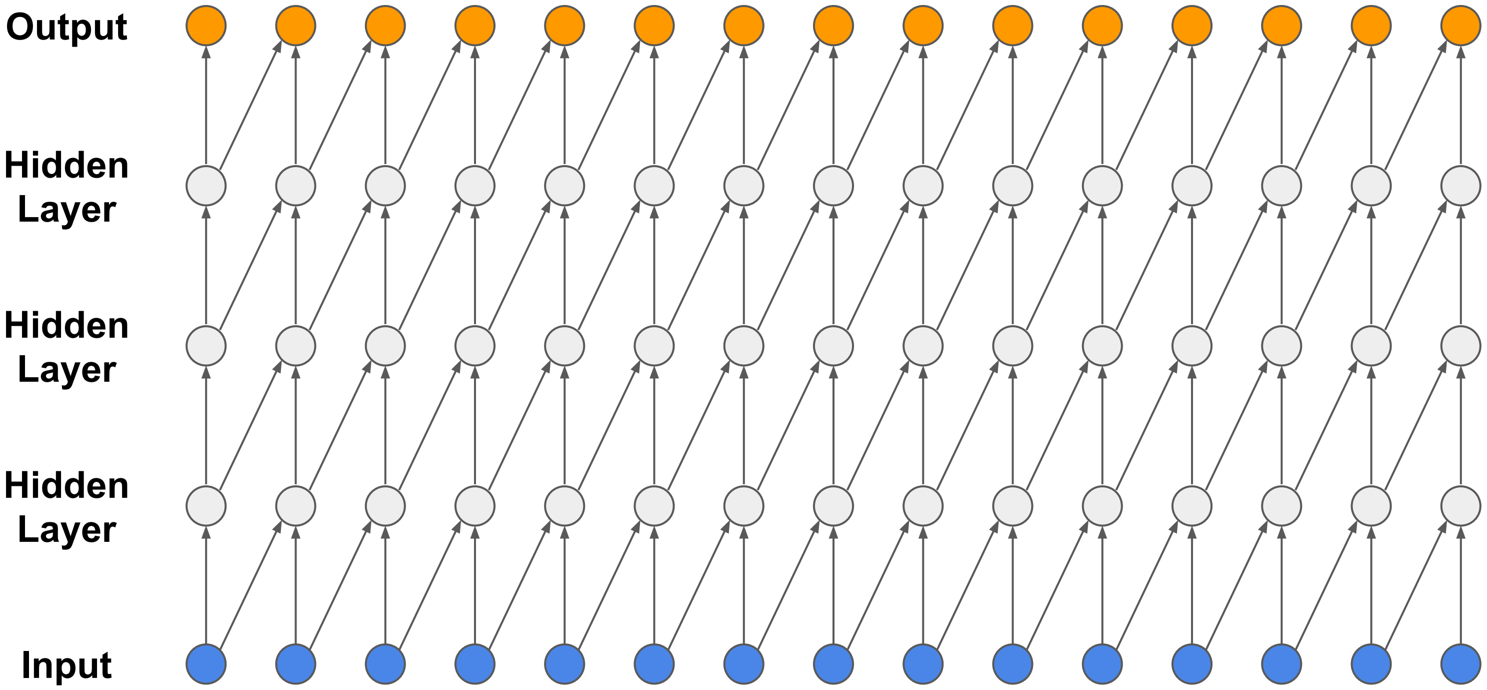


Google DeepMind

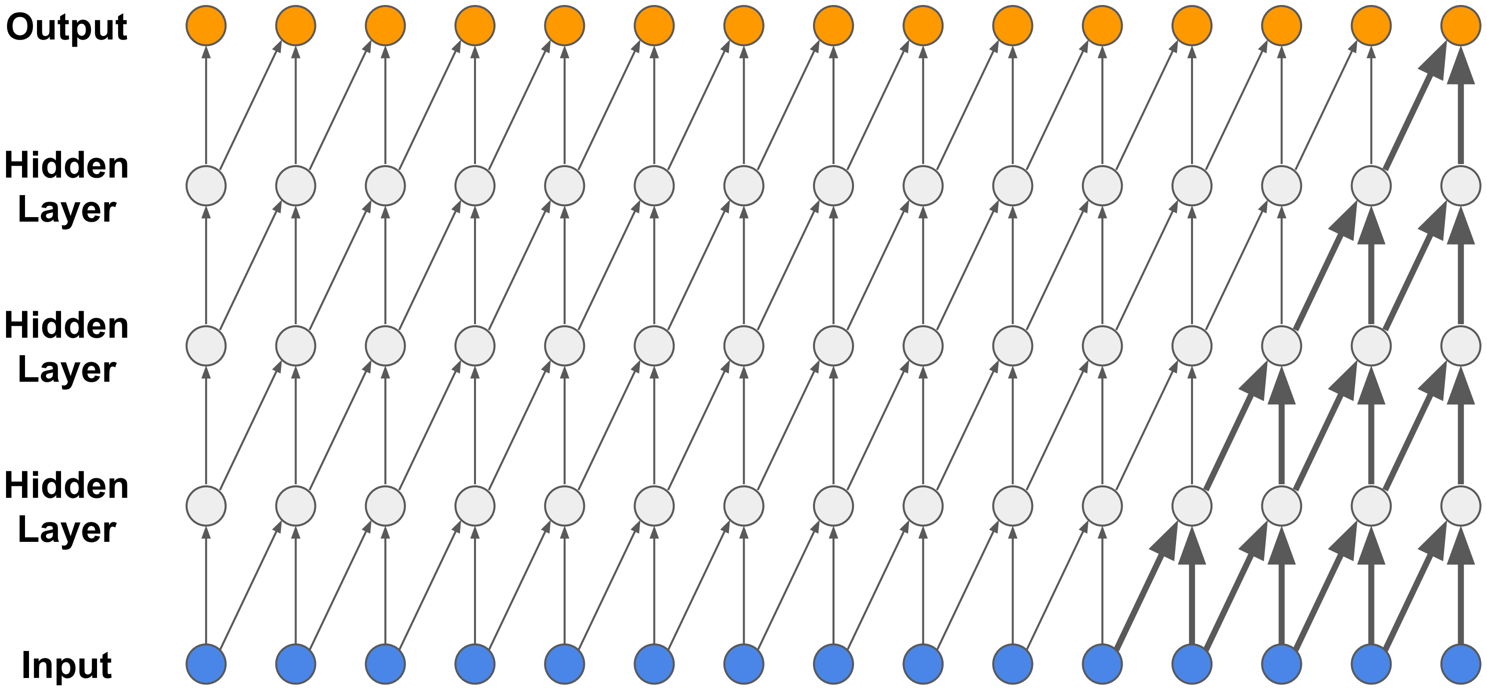
Causal Convolution



Causal Convolution



Causal Convolution

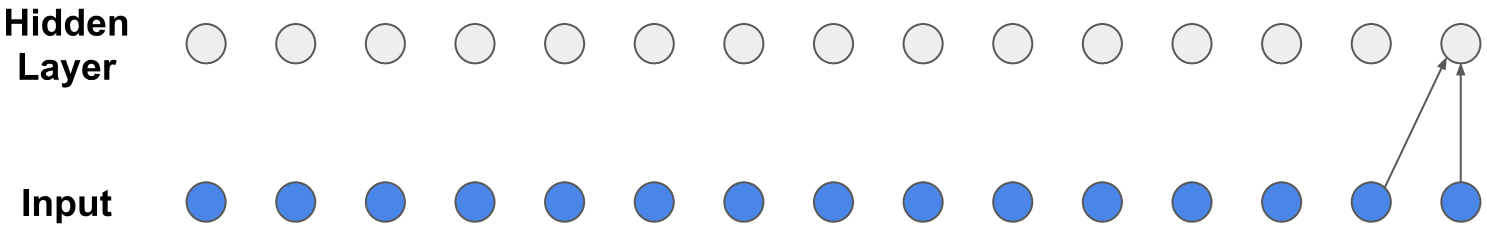


Google DeepMind

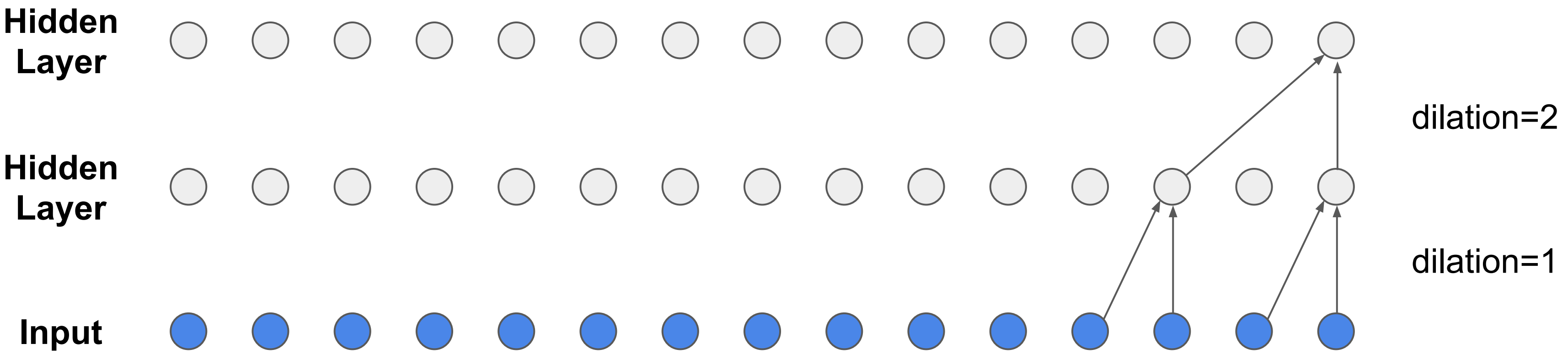
Causal Dilated Convolution

Input

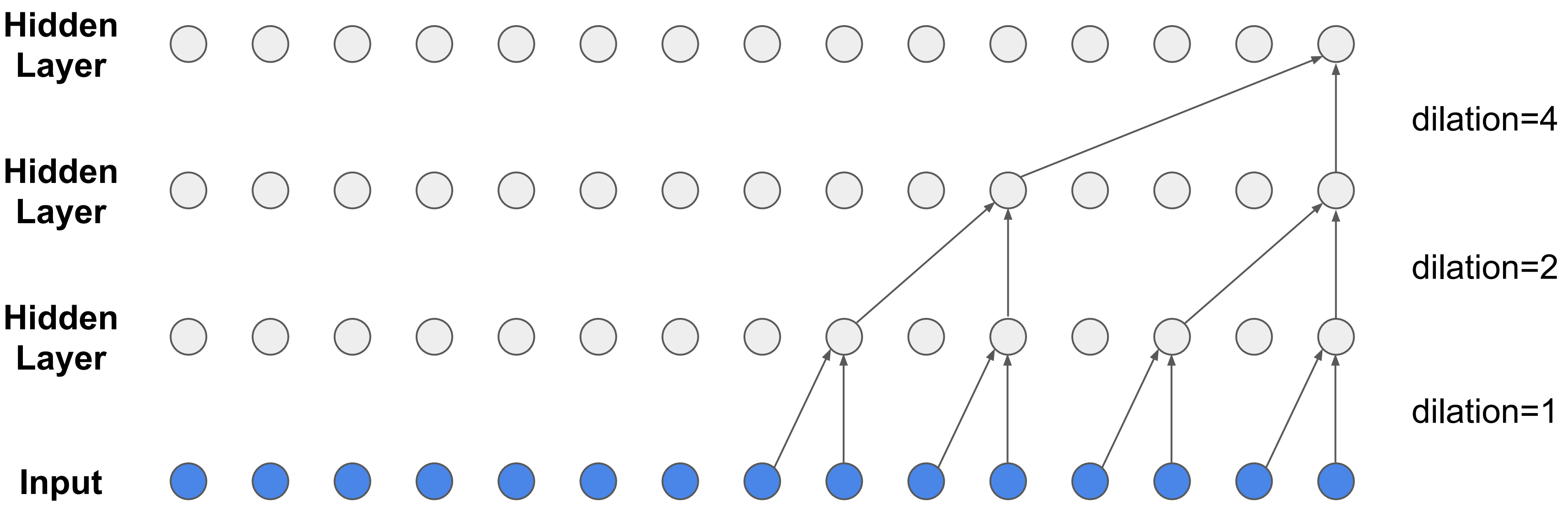
Causal Dilated Convolution



Causal Dilated Convolution

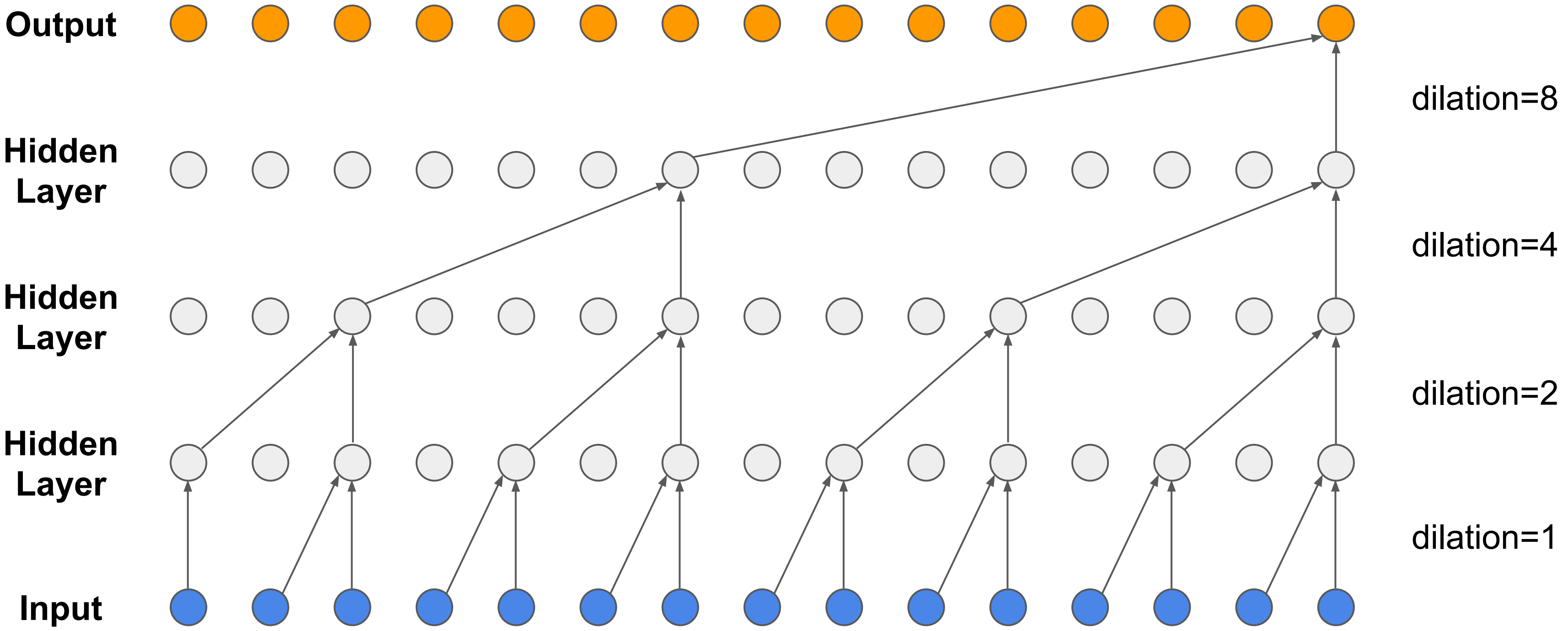


Causal Dilated Convolution



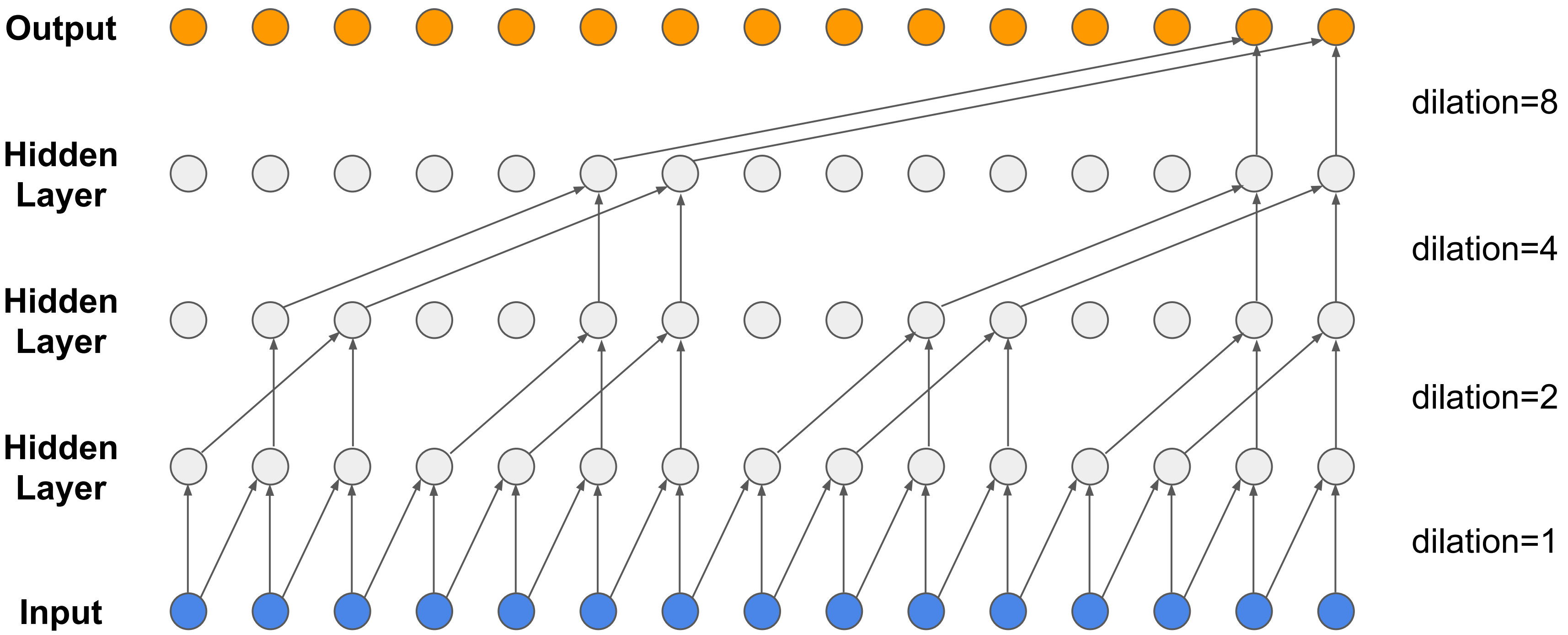
Google DeepMind

Causal Dilated Convolution



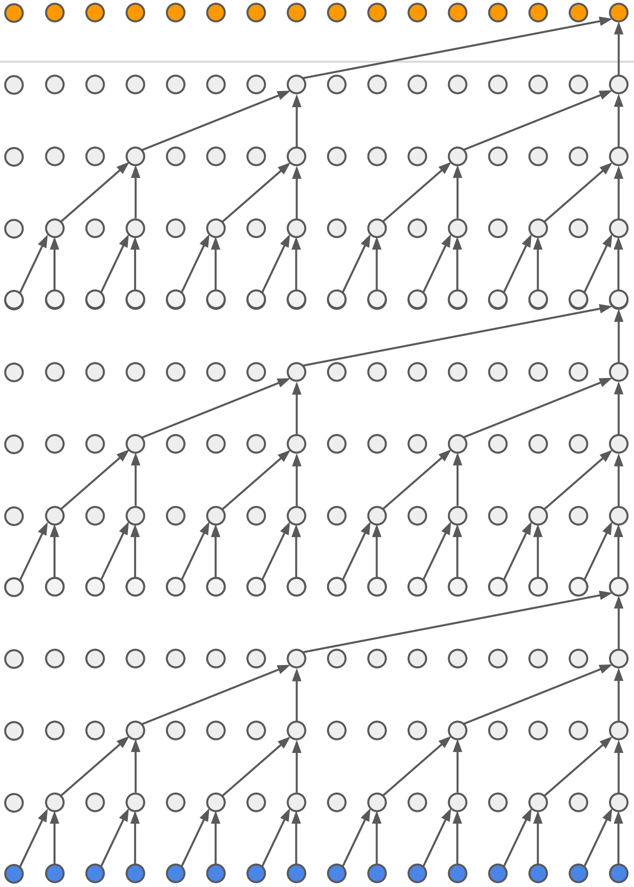
Google DeepMind

Causal Dilated Convolution

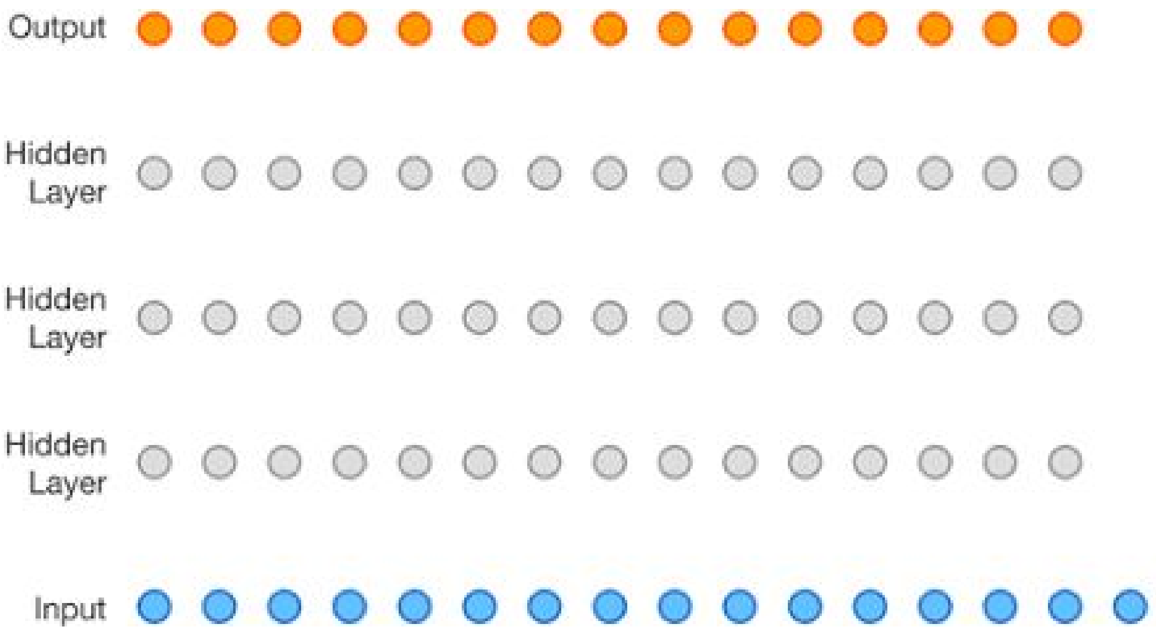


Google DeepMind

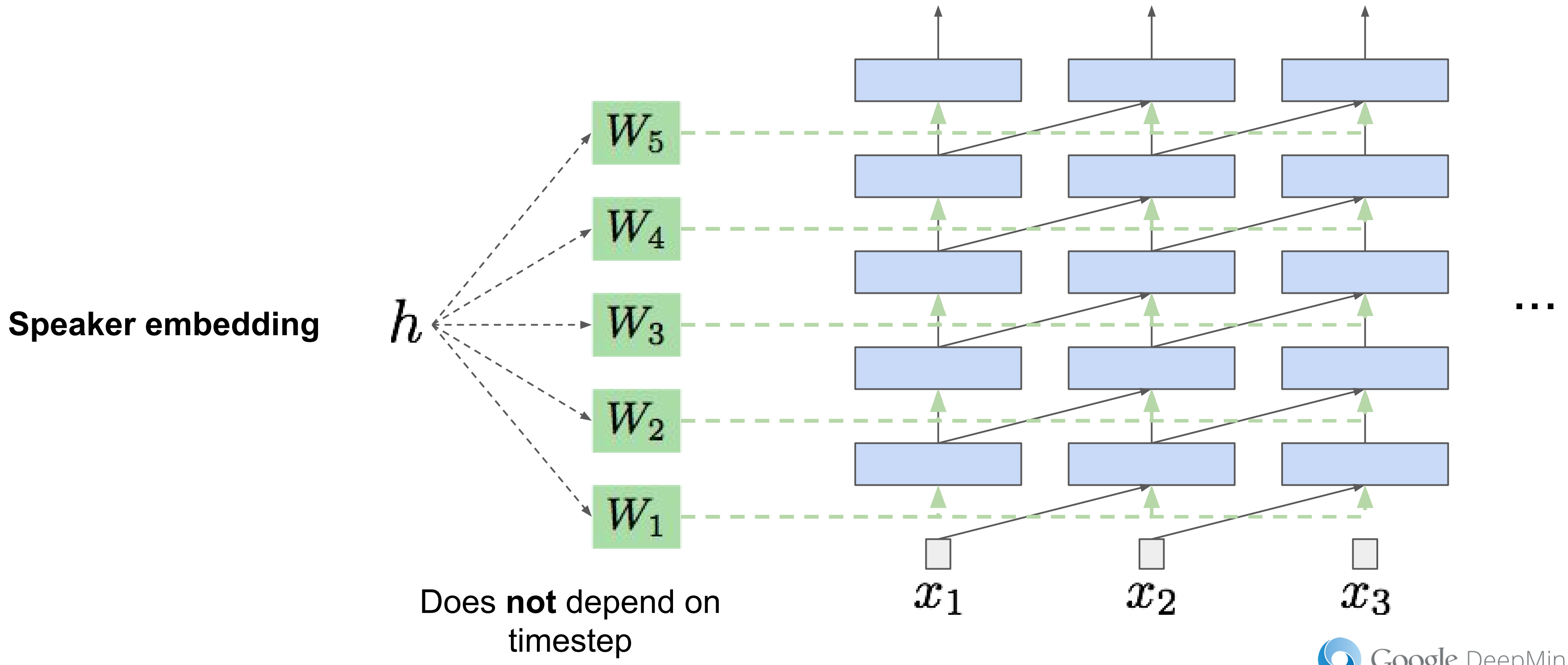
Multiple Stacks



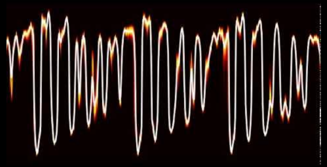
Sampling

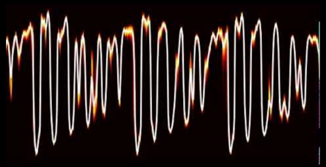


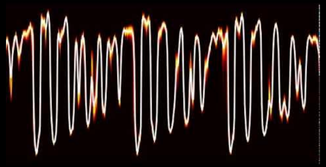
Speaker-conditional Generation

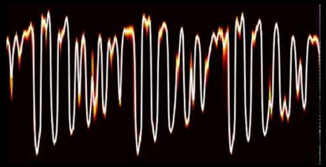


https://deepmind.com/blog/wavenet-generative-model-raw-audio/





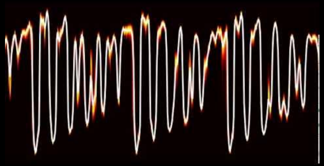


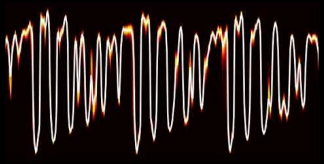


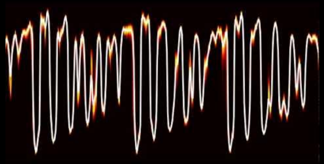
Text-To-Speech samples

Google DeepMind

https://deepmind.com/blog/wavenet-generative-model-raw-audio/





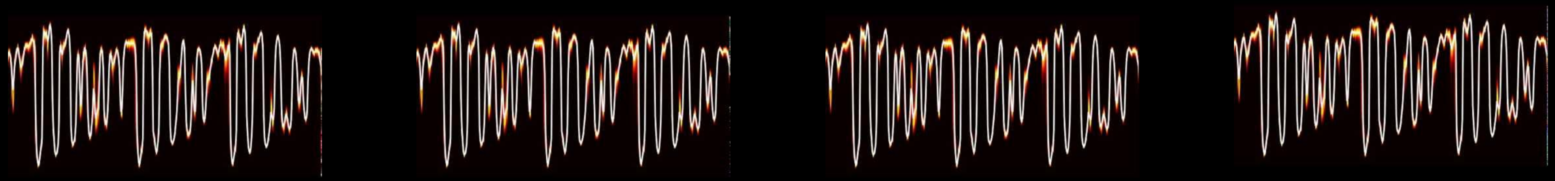


Speaker-conditional samples

(but not conditioned on text)

Google DeepMind

https://deepmind.com/blog/wavenet-generative-model-raw-audio/



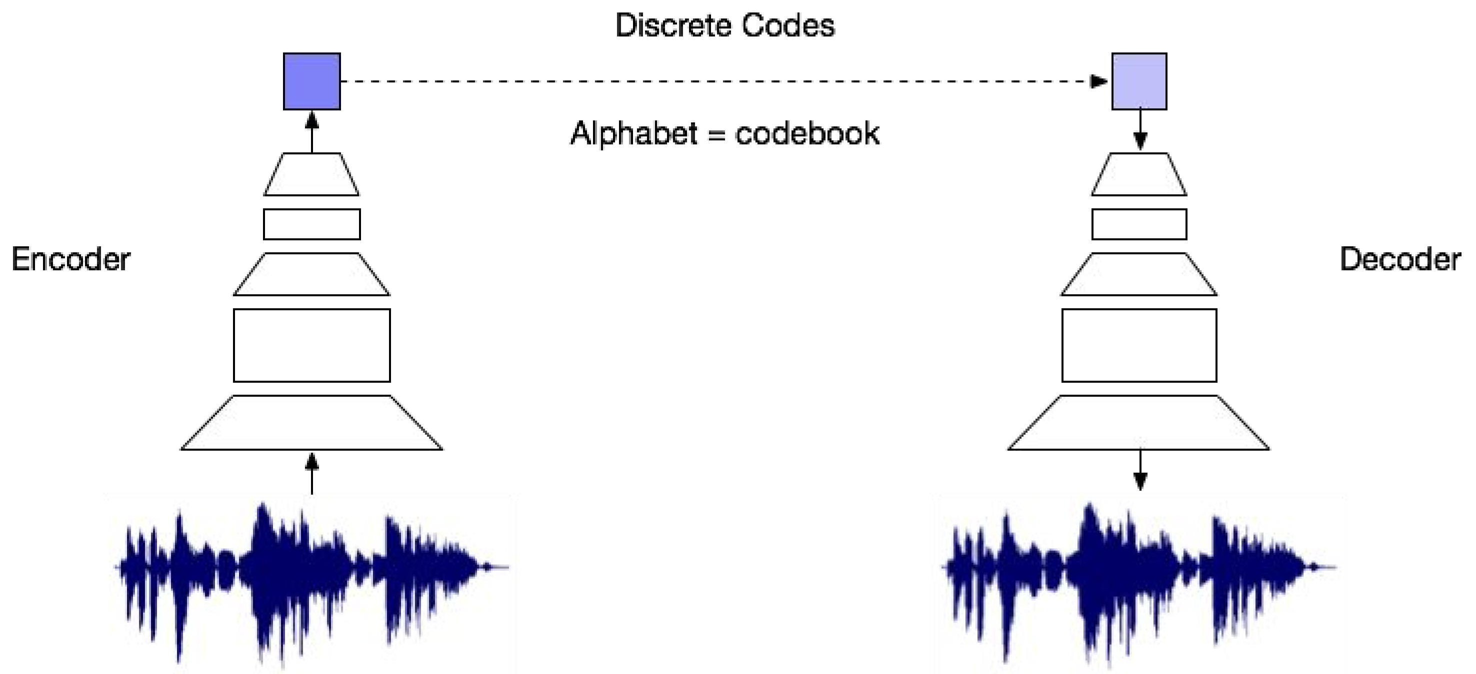
Piano Music samples

VQ-VAE

- Towards modeling a latent space
    - Learn meaningful representations.
    - Abstract away noise and details.
    - Model what’s important in a compressed latent representation.

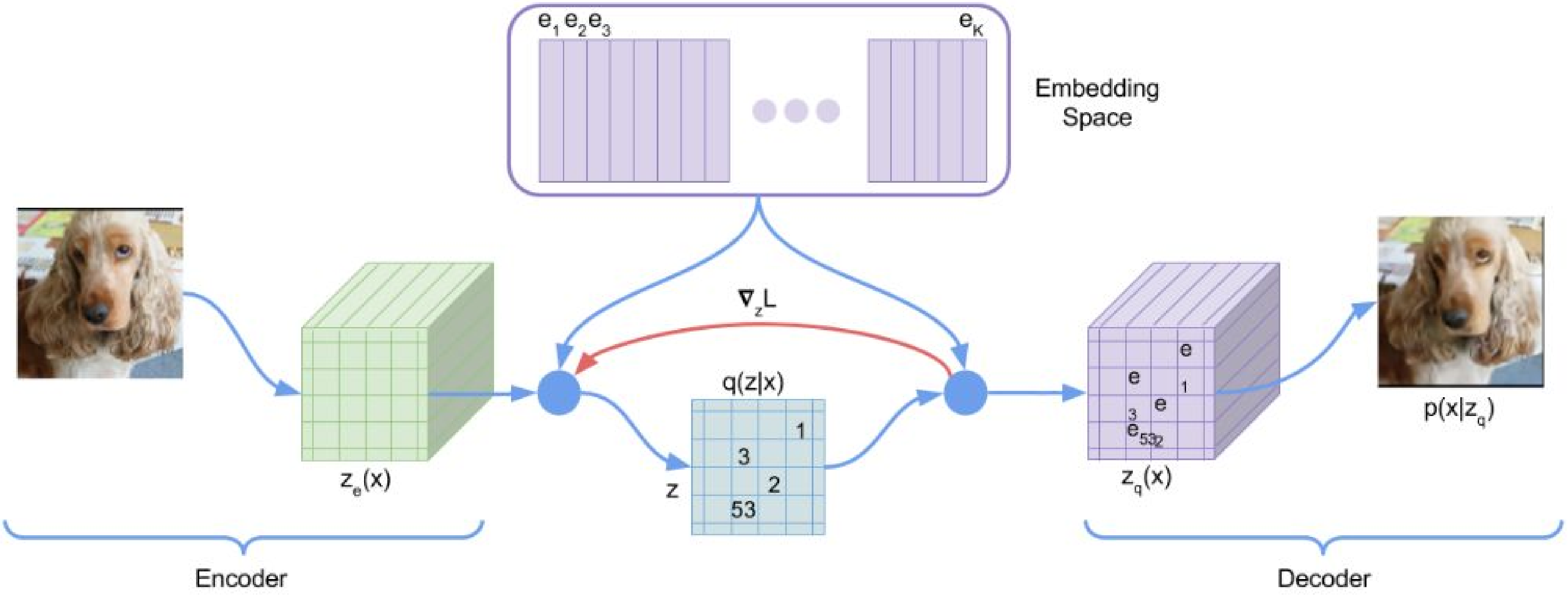
- Why discrete?
    - Many important real-world things are discrete.
    - Arguably easier to model for the prior (e.g., softmax vs RNADE)
    - Continuous representations are often inherently discretized by encoder/decoder.

VQ-VAE



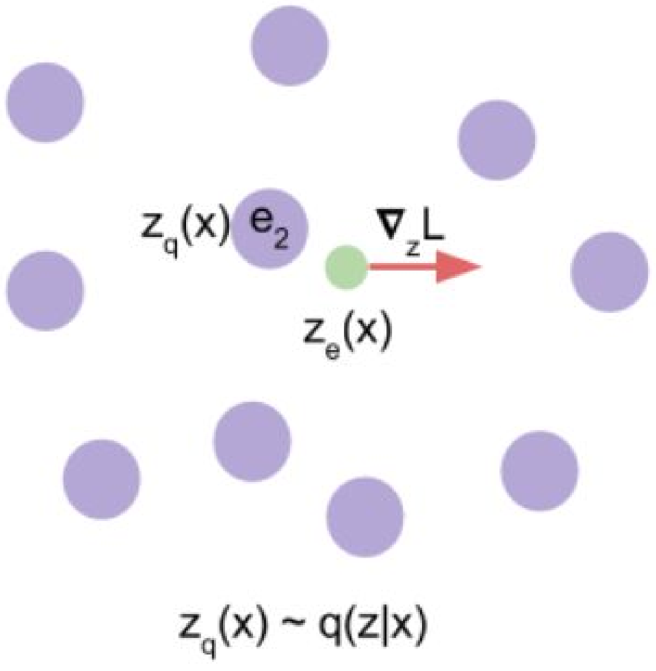
Related work:
PixelVAE (Gulrajani et al, 2016)
Variational Lossy AutoEncoder (Chen et al, 2016)

VQ-VAE



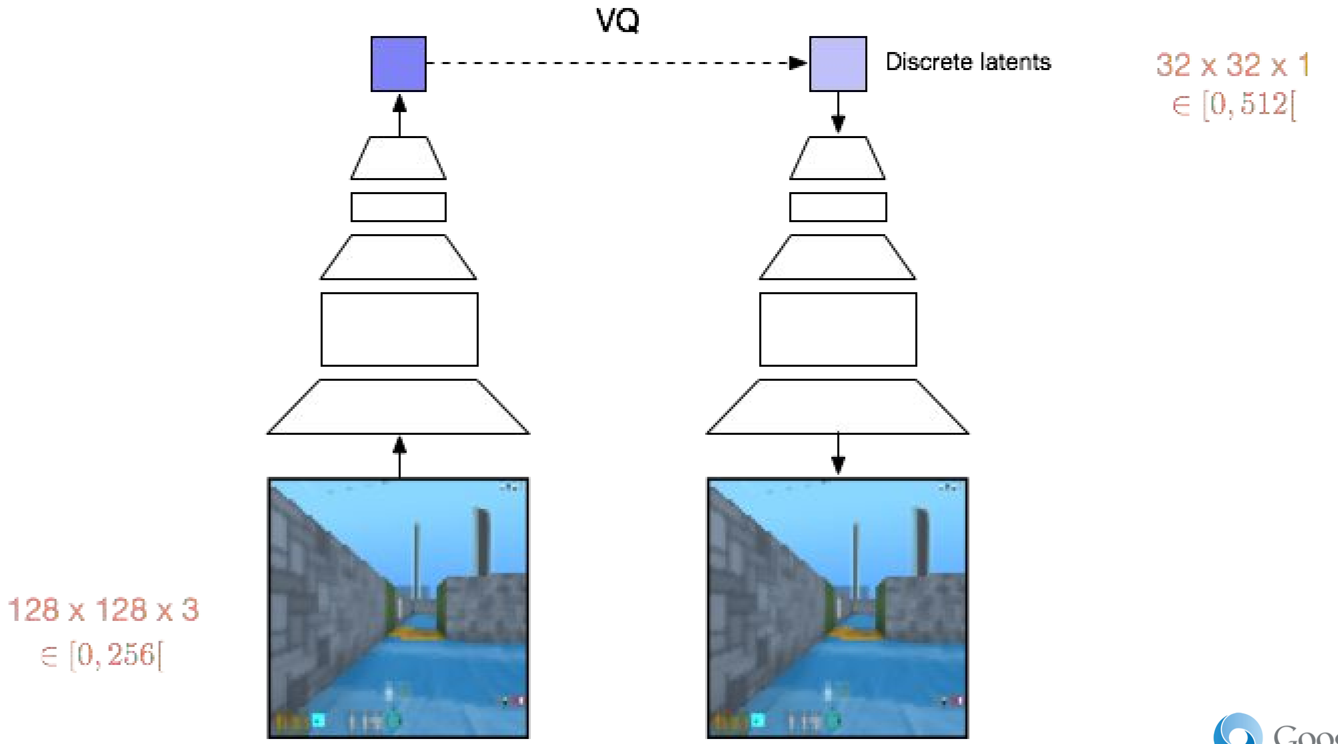
e1 e2 e3
eK
Embedding Space
∇zL
q(z|x)
1
3
2
53
z
ze(x)
zq(x)
p(x|zq)
Encoder
Decoder

VQ-VAE



Google DeepMind

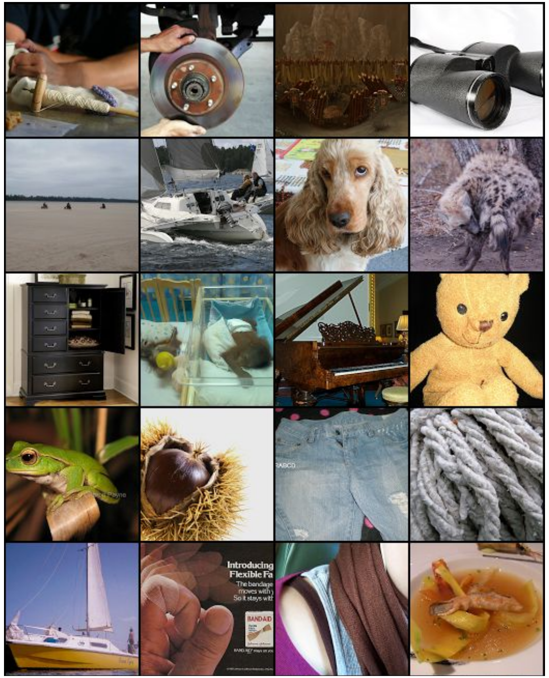
Images



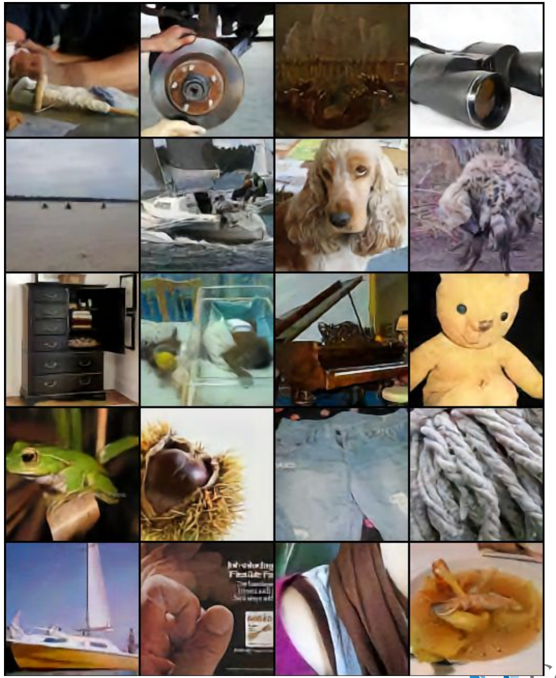
Google DeepMind

ImageNet reconstructions

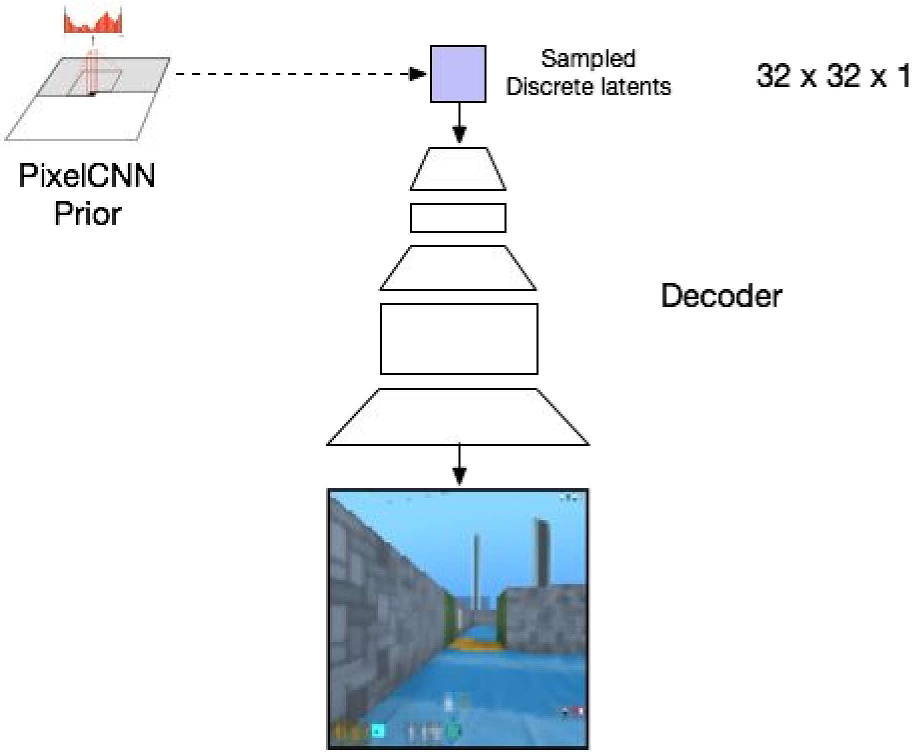
Original 128x128 images



Reconstructions

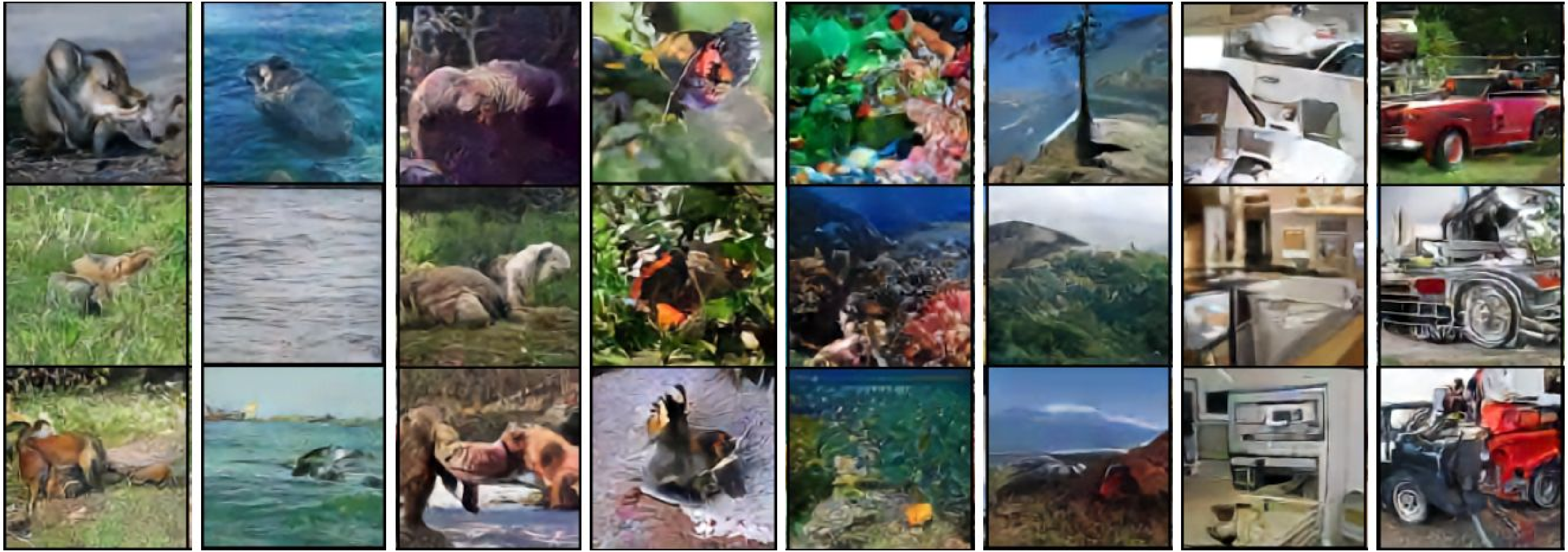


VQ-VAE - Sample



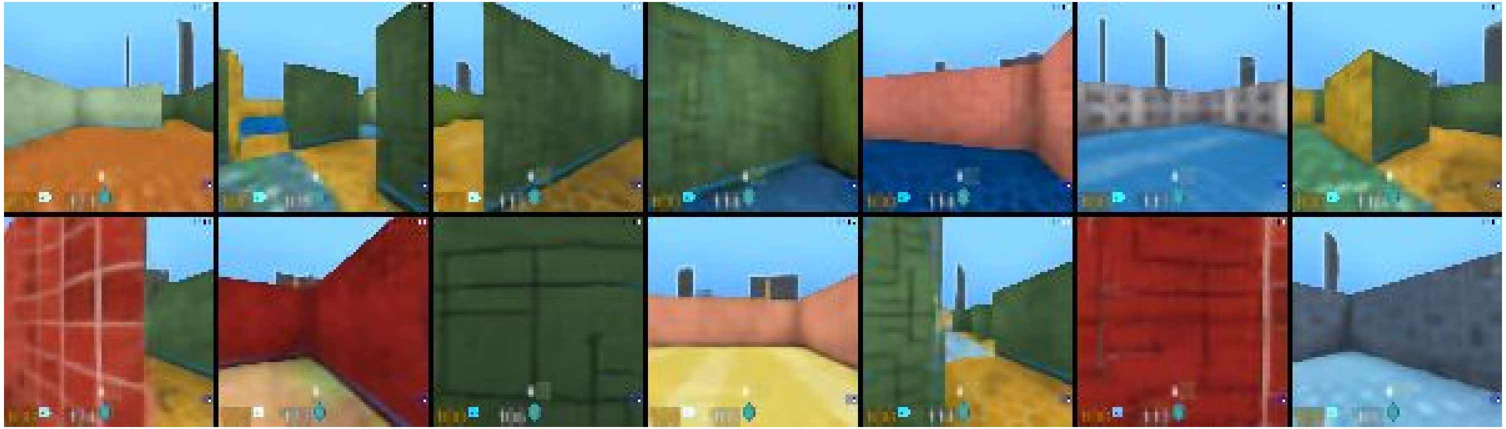
Google DeepMind

ImageNet samples



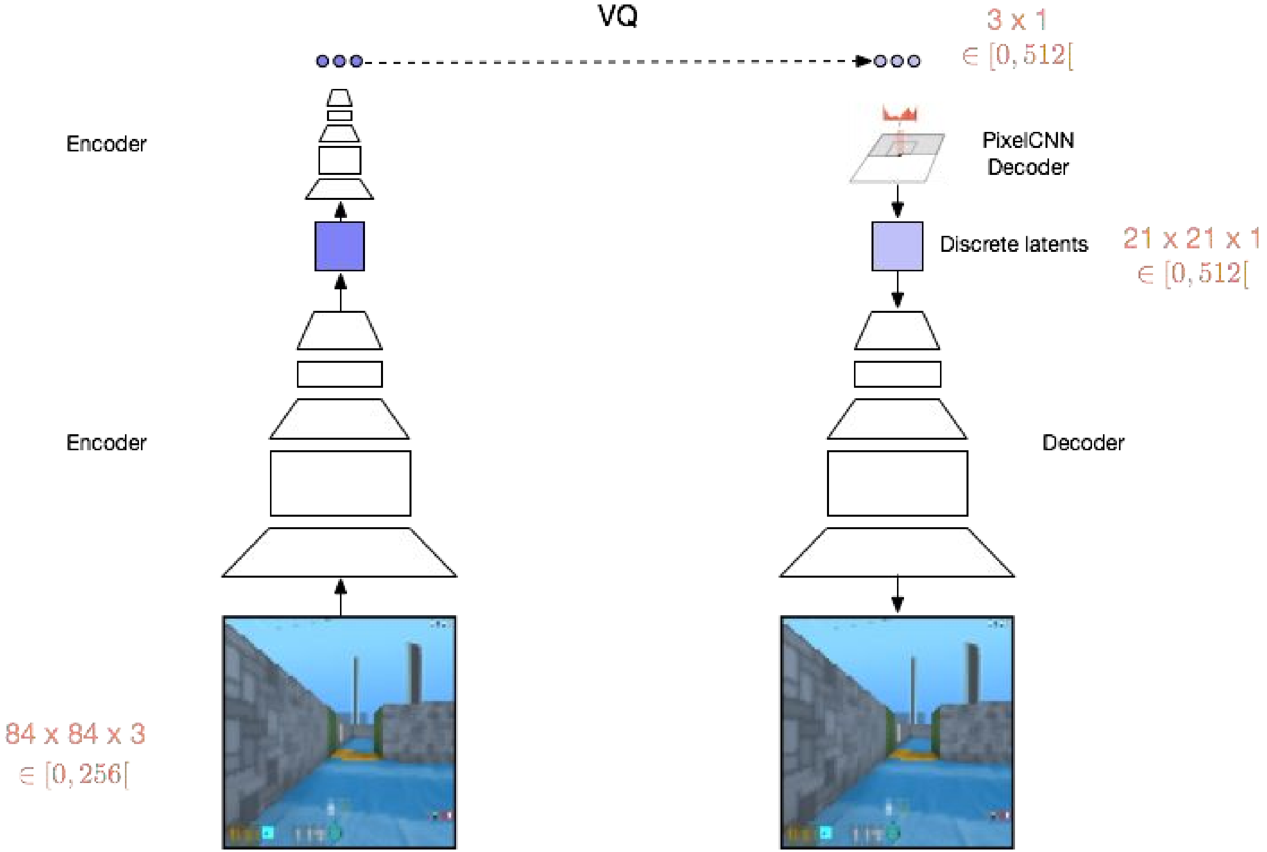
Google DeepMind

DM-Lab Samples



Google DeepMind

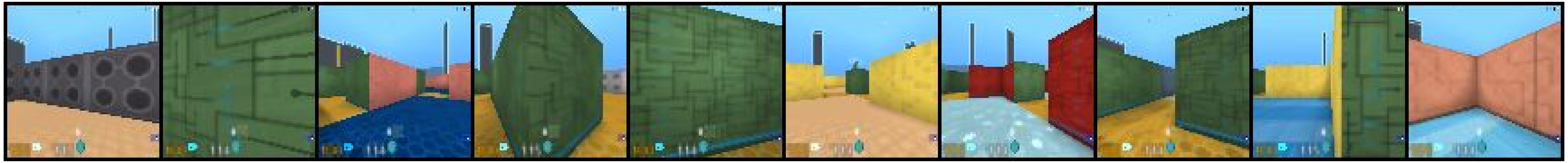
3 Global Latents Reconstruction



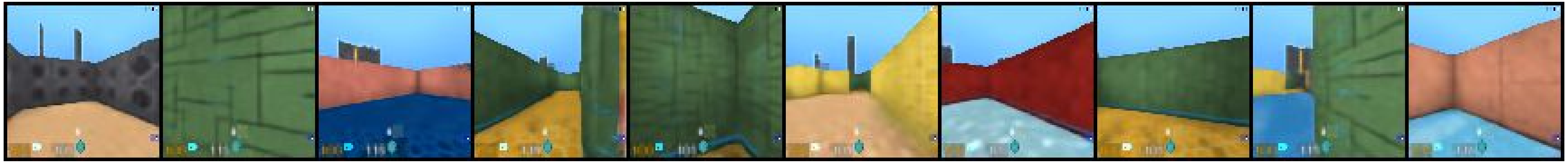
Google DeepMind

3 Global Latents Reconstruction

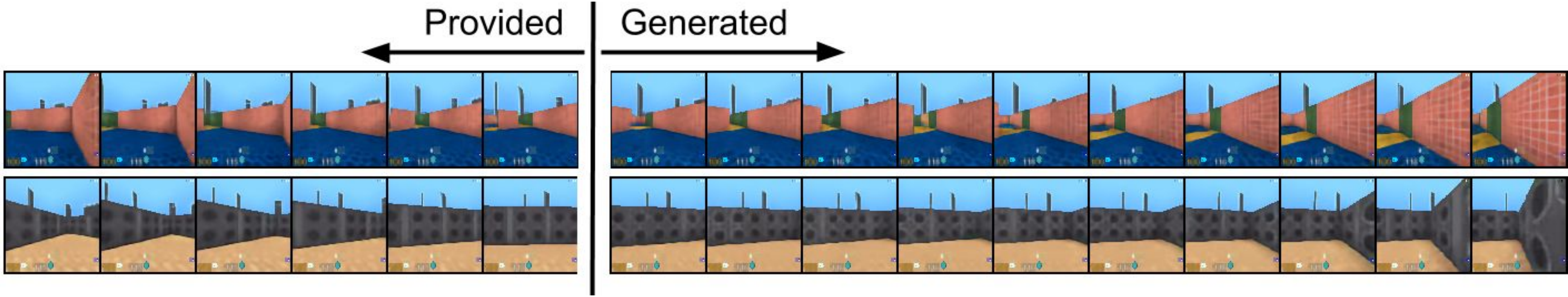
Originals



Reconstructions from compressed representations (27 bits per image).

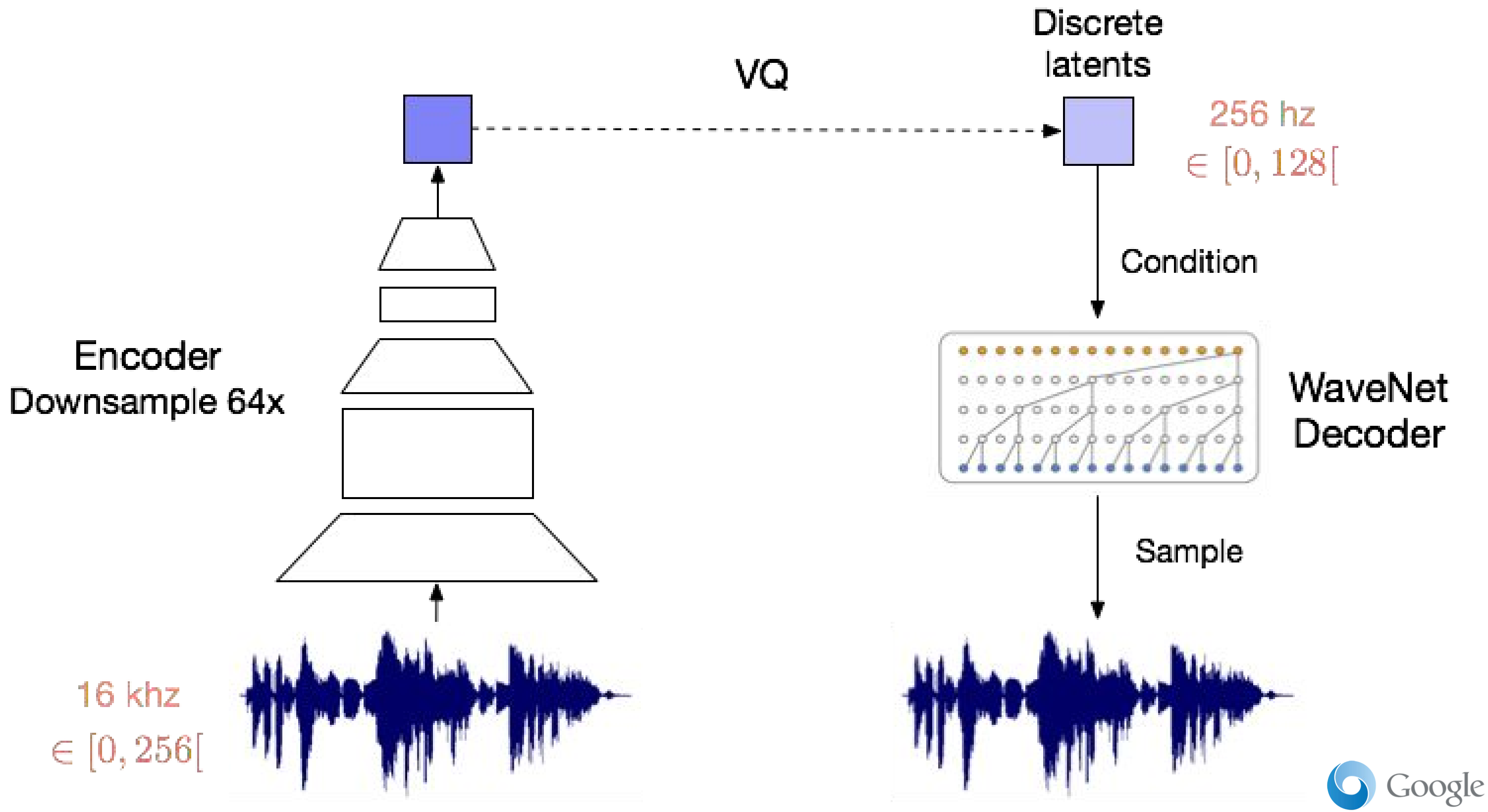


Video Generation in the latent space



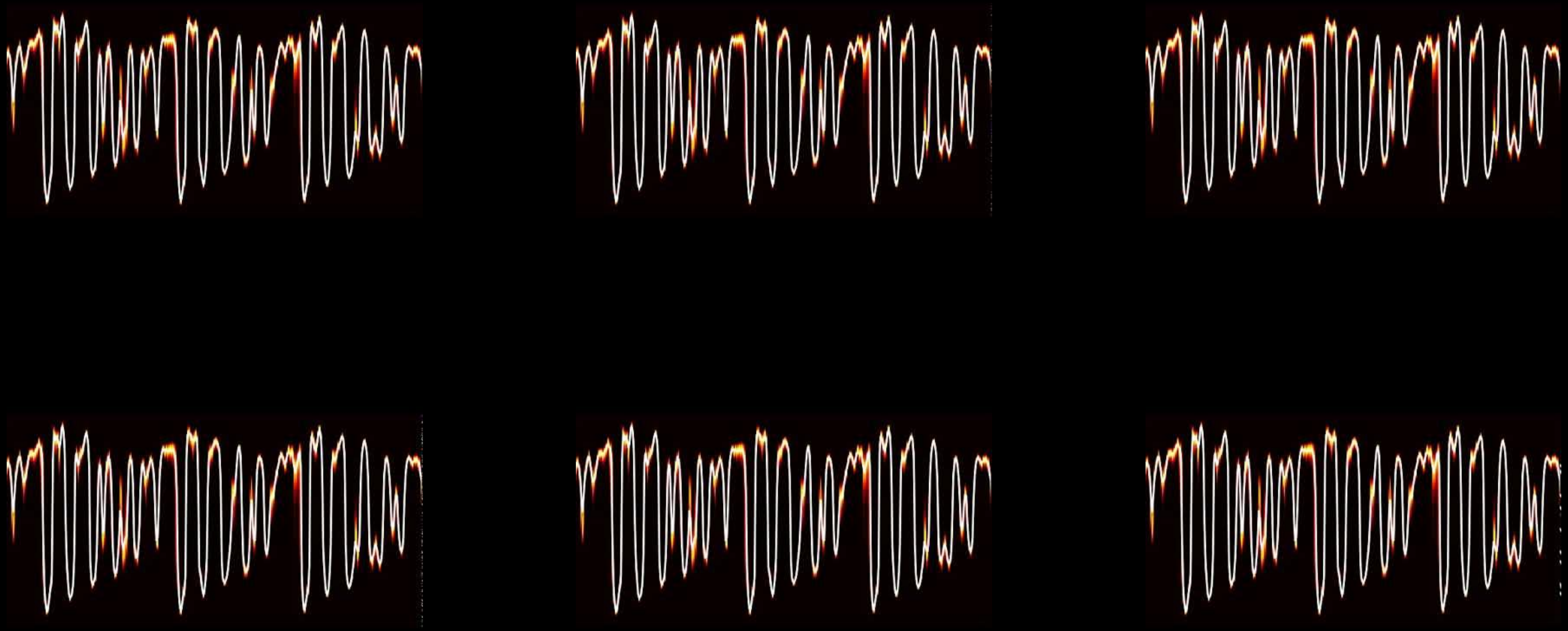
Google DeepMind

Speech



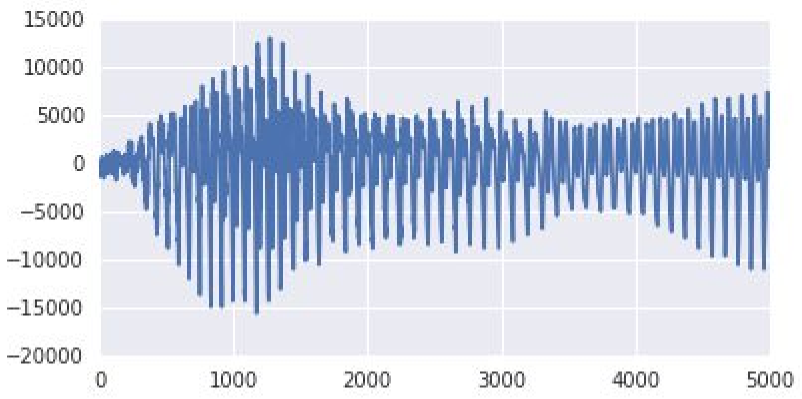
VQ
Discrete latents
256 hz
$\in [0, 128[$
Condition
WaveNet Decoder
Sample
Encoder
Downsample 64x
16 khz
$\in [0, 256[$

https://avdnoord.github.io/homepage/vqvae/

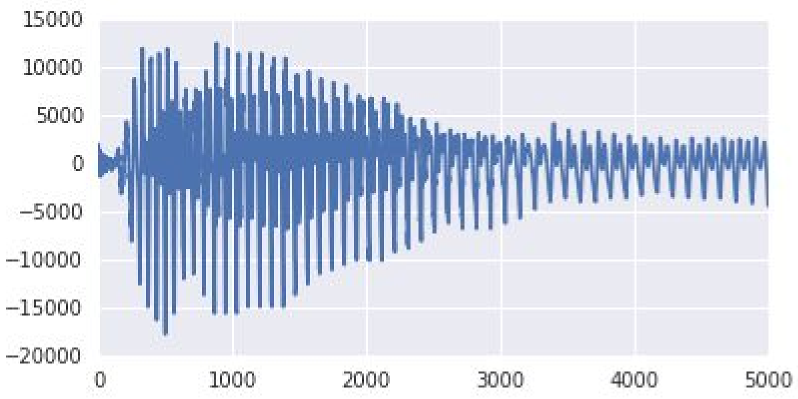


Google DeepMind

Speech - reconstruction



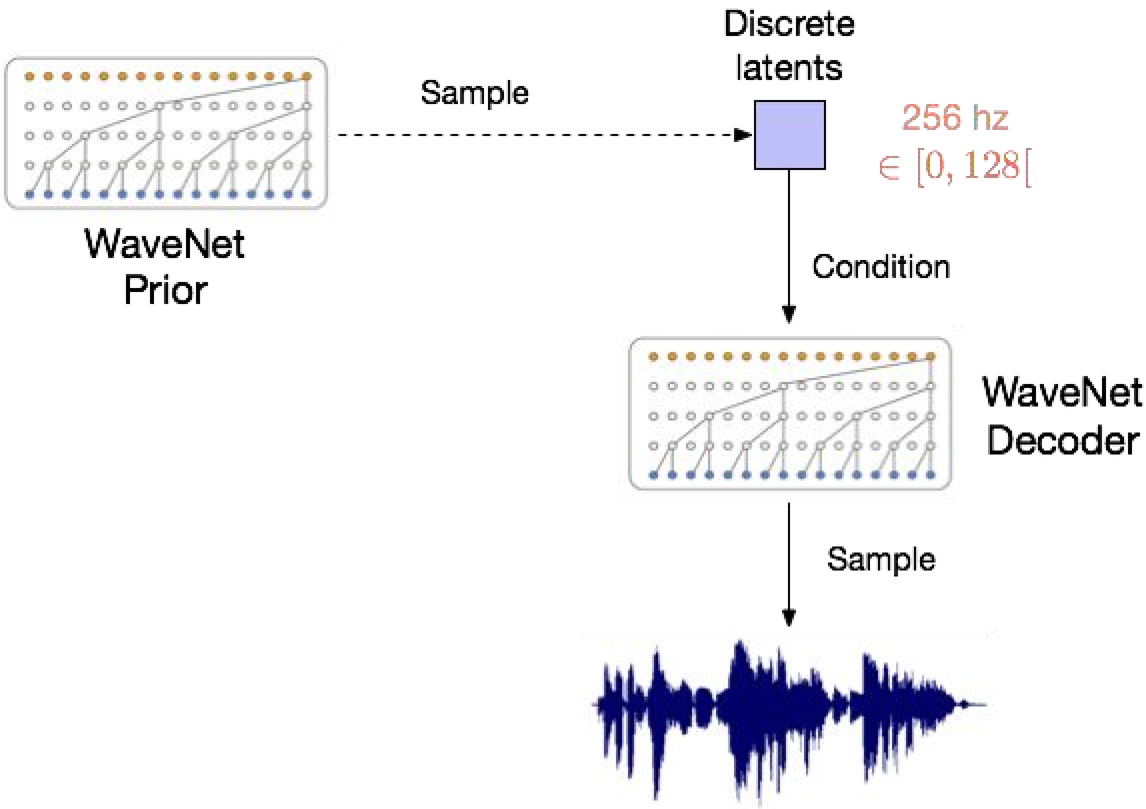
Original



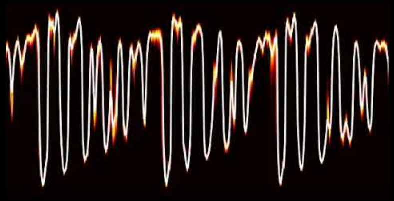
Reconstruction

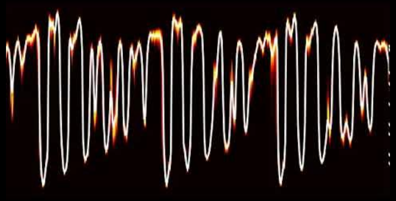
Google DeepMind

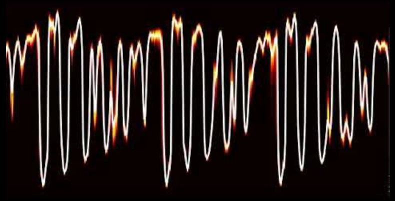
Speech - Sample from prior



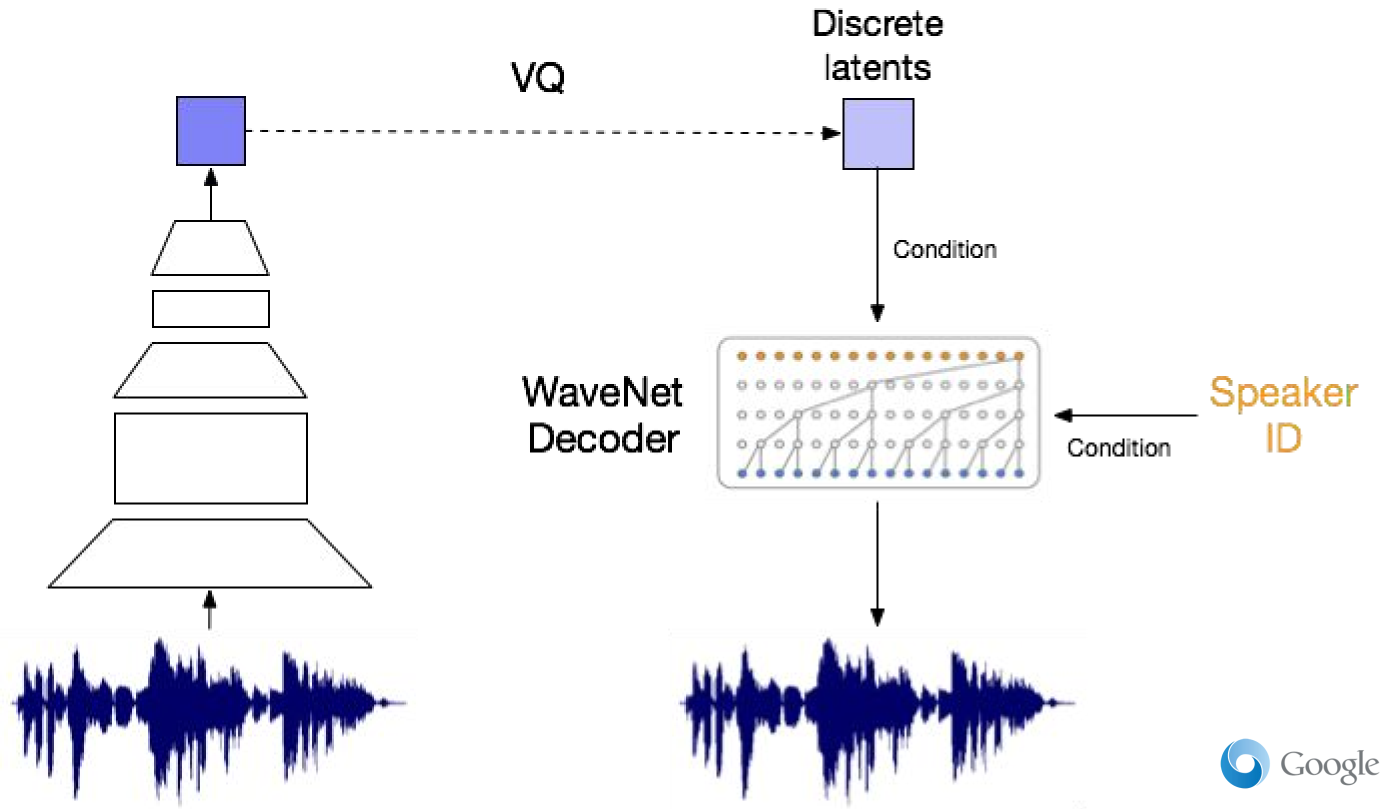
https://avdnoord.github.io/homepage/vqvae/



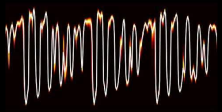
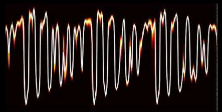
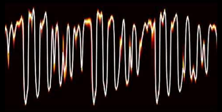
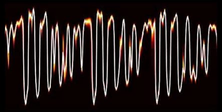
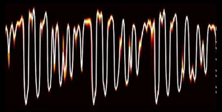
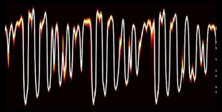


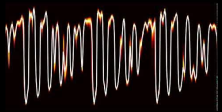
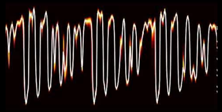
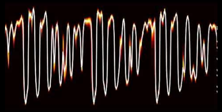
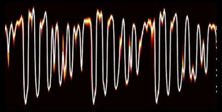
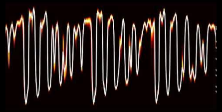
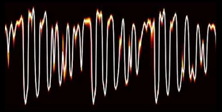


Speech - speaker conditional



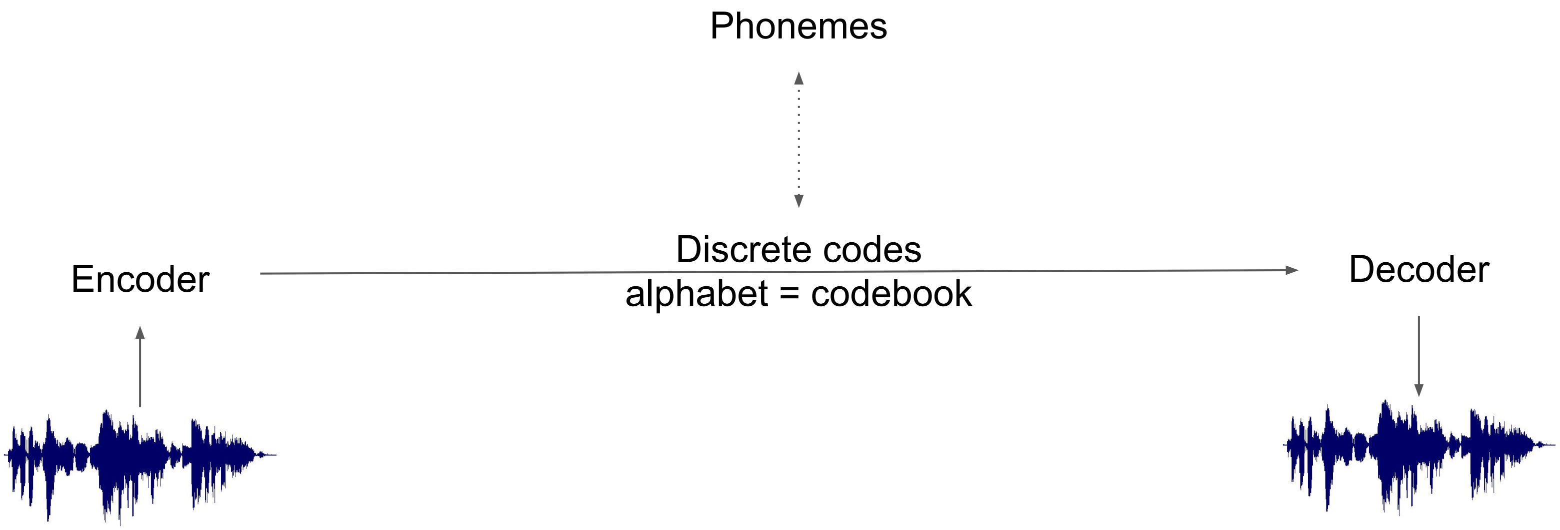
https://avdnoord.github.io/homepage/vqvae/

Google DeepMind

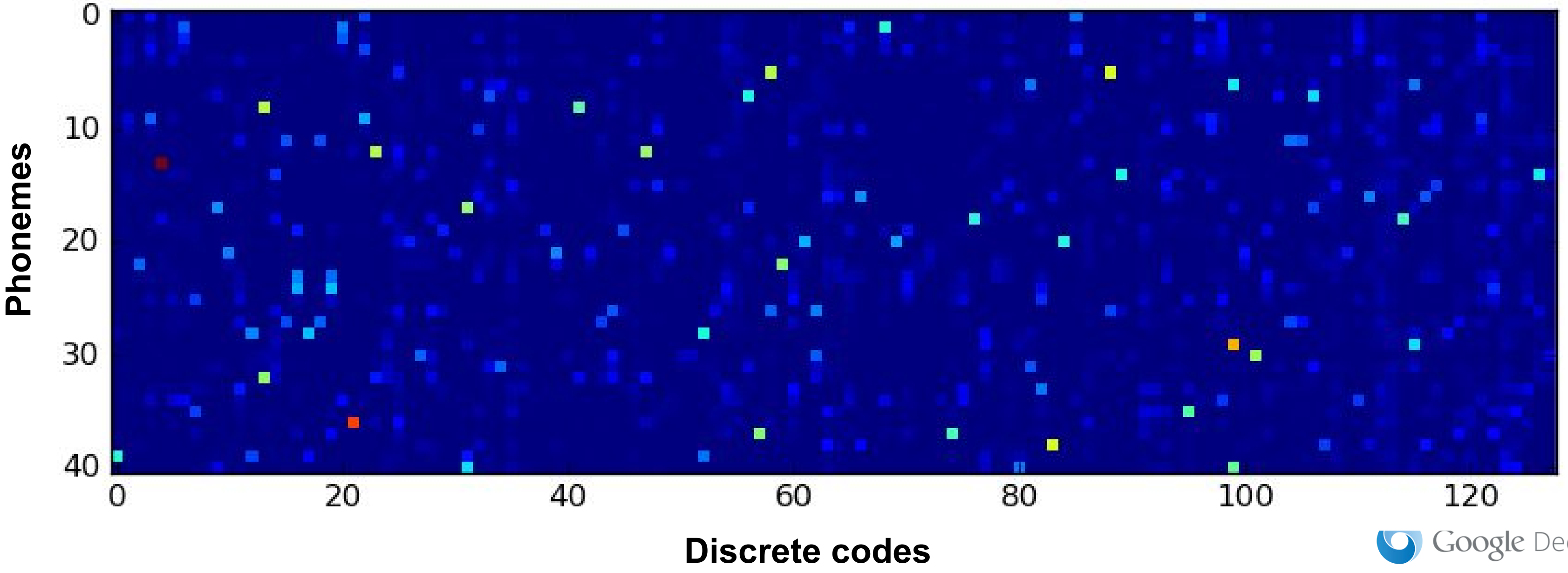
Unsupervised Learning of phonemes



Unsupervised Learning of phonemes

41-way classification

49.3**%** accuracy **fully unsupervised**



Google DeepMind

References and related work

Pixel Recurrent Neural Networks - van den Oord et al, ICML 2016

Conditional Image Generation with PixelCNN Decoders - van den Oord et al, NIPS 2016

WaveNet: A Generative Model For Raw Audio - van den Oord et al, Arxiv 2016

Neural Machine Translation in Linear Time - Kalchbrenner et al, Arxiv 2016

Video Pixel Networks - Kalchbrenner et al, ICML 2017

Neural Discrete Representation Learning - van den Oord et al, NIPS 2017

Related work:

The Neural Autoregressive Distribution Estimator - Larochelle et al, AISTATS 2011

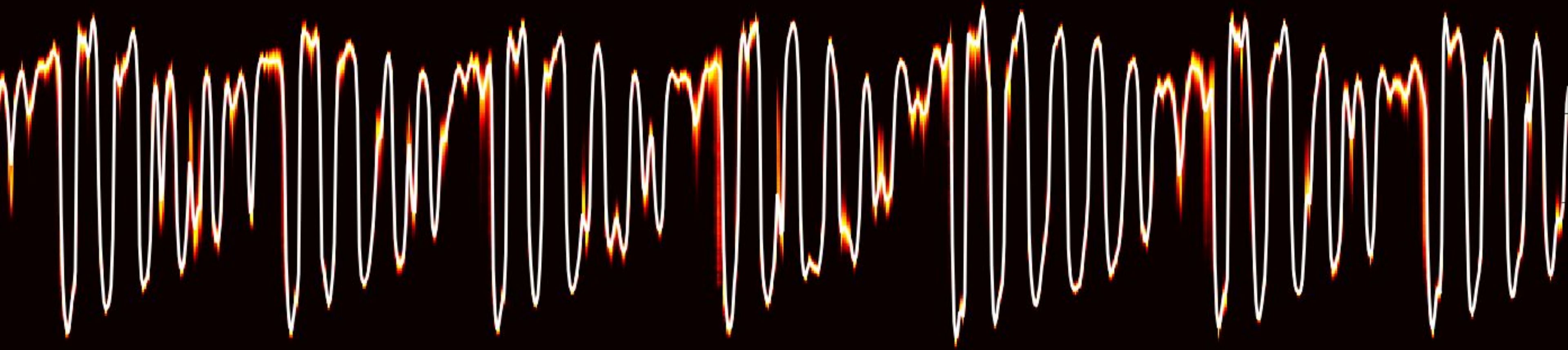
Generative image modeling using spatial LSTMs - Theis et al, NIPS 2015

SampleRNN: An Unconditional End-to-End Neural Audio Generation Model - Mehri et al, ICLR 2017

PixelVAE: A Latent Variable Model for Natural Images - Gulrajani et al, ICLR 2017

Variational Lossy Autoencoder - Chen et al, ICLR 2017

Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations - Agustsson et al, NIPS 2017



Thank you!

Google DeepMind