Generating music in the raw audio domain

Sander Dieleman - October 18th, 2018

PhD student, Ghent University, Belgium (2010 - 2016)
“Learning feature hierarchies for musical audio signals”

Machine learning intern, Spotify, NYC (summer 2014)
Scaling up content-based music recommendation

Research Scientist, DeepMind, London UK (since July 2015)
AlphaGo, WaveNet, ...

http://benanne.github.io

sanderdieleman@gmail.com

@sedielem

Overview

Why model music in the raw audio domain?

WaveNet: an autoregressive model of raw audio

Generating music with WaveNets

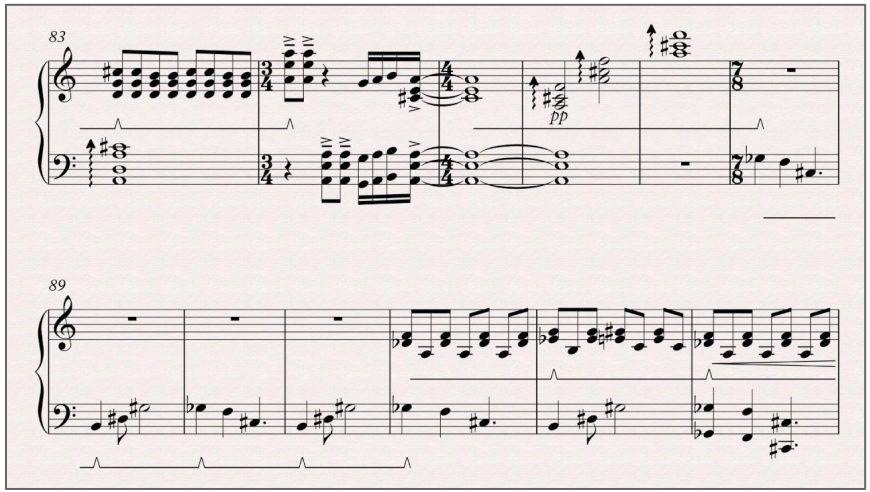
Modelling long-range correlations with WaveNets

Generating music with long-range consistency

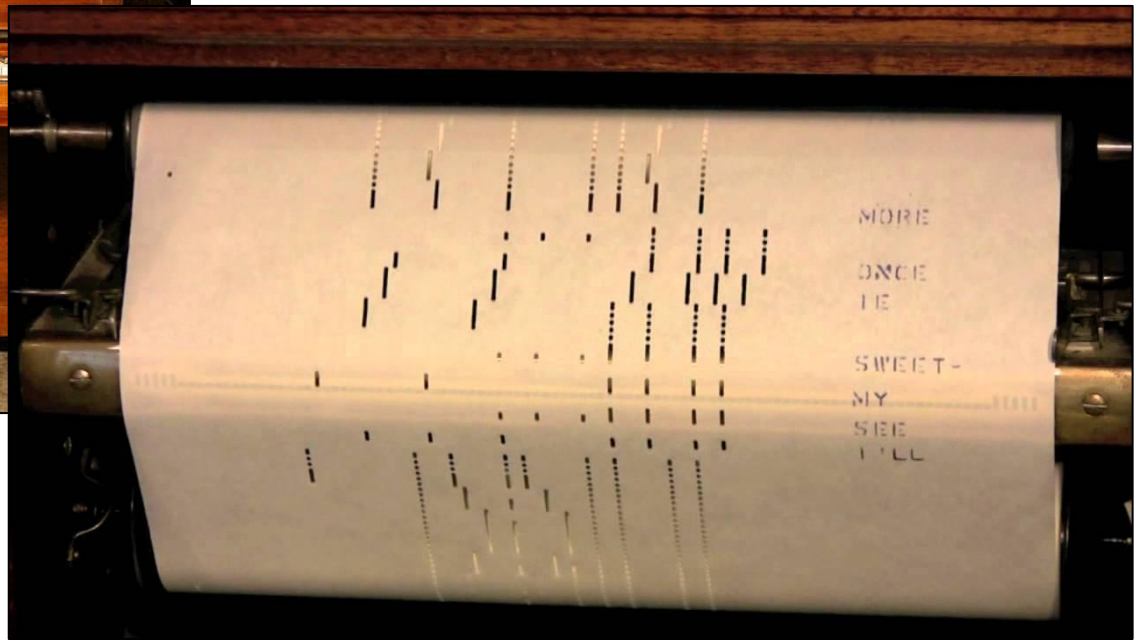
Why model music in the raw audio domain?

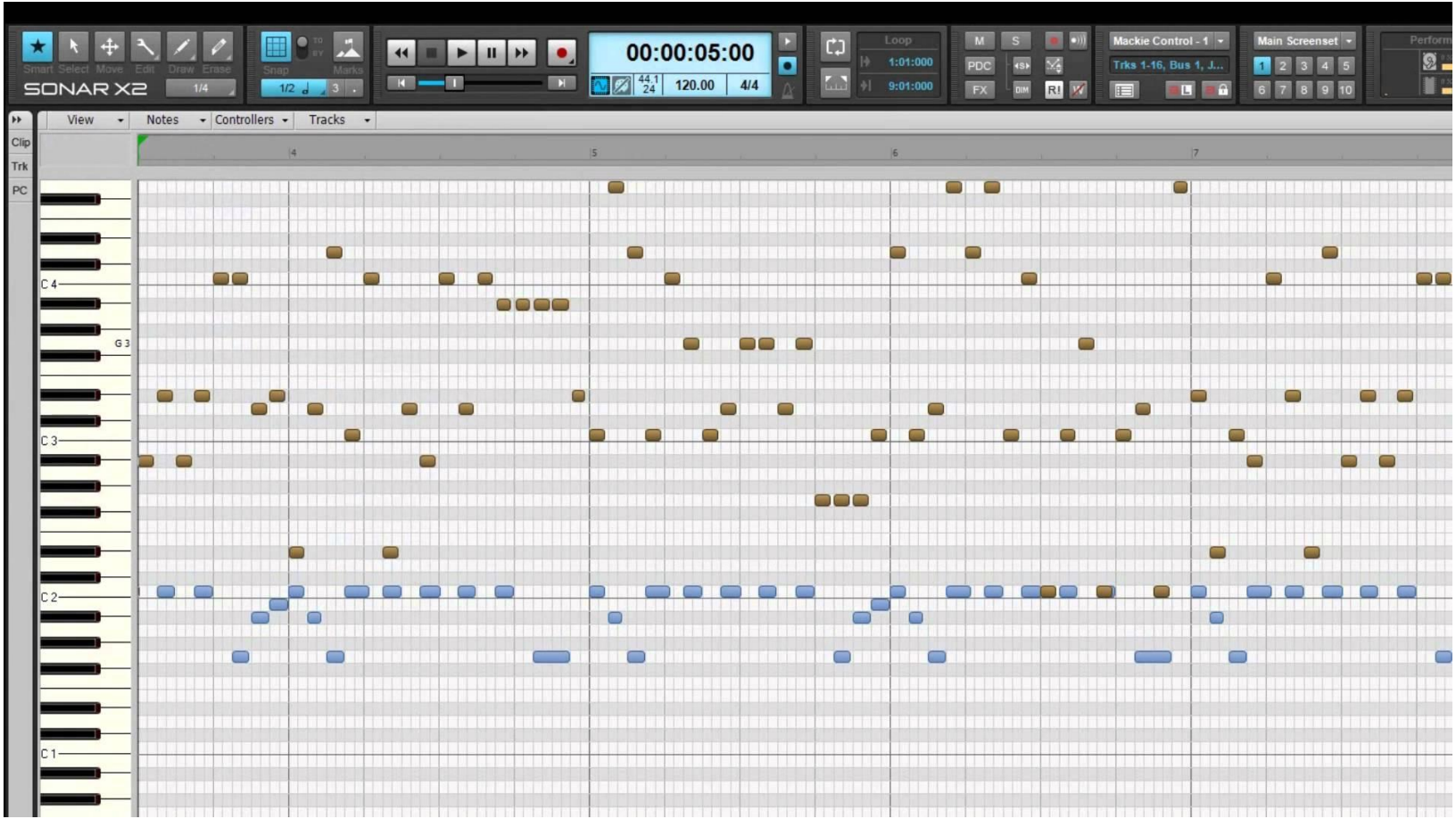
Why raw audio?

Music generation is typically studied in the symbolic domain




Steinway


MORE
ONCE
TE
SWEET-
MY
SEE


SONAR X2
Smart Select Move Edit Draw Erase
Snap Marks
00:00:05:00
120.00
4/4
Loop
1:01:000
9:01:000
Mackie Control - 1
Trks 1-16, Bus 1, J...
Main Screenset
View Notes Controllers Tracks
Clip
Trk
PC
C 4
G 3
C 3
C 2
C 1

Why raw audio?

Many instruments have complex action spaces

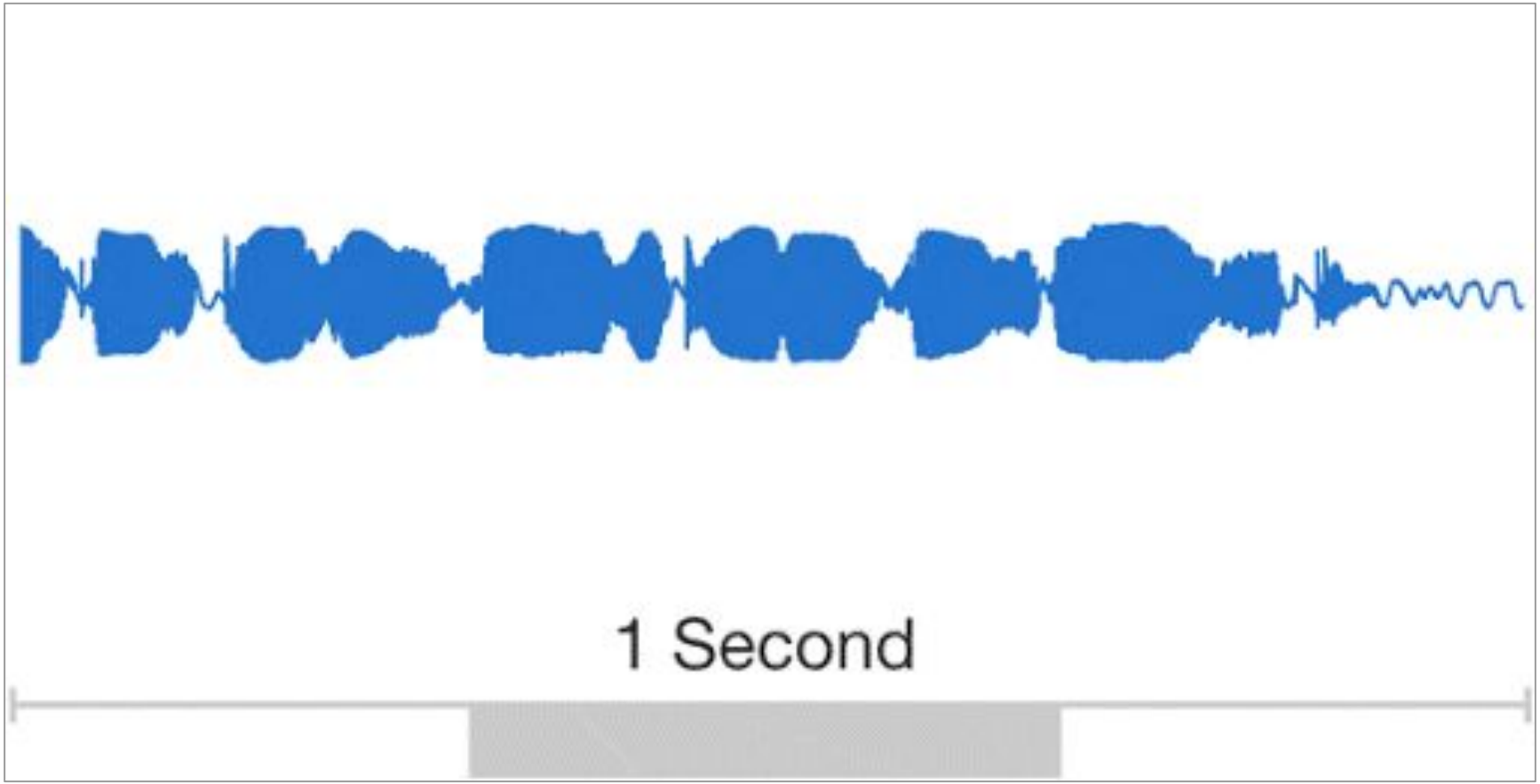
⇒ Rich palette of sounds and timbral variations

Guitar

- pick vs. finger
- picking position
- frets
- harmonics
- ...

WaveNet:
an autoregressive model of raw audio

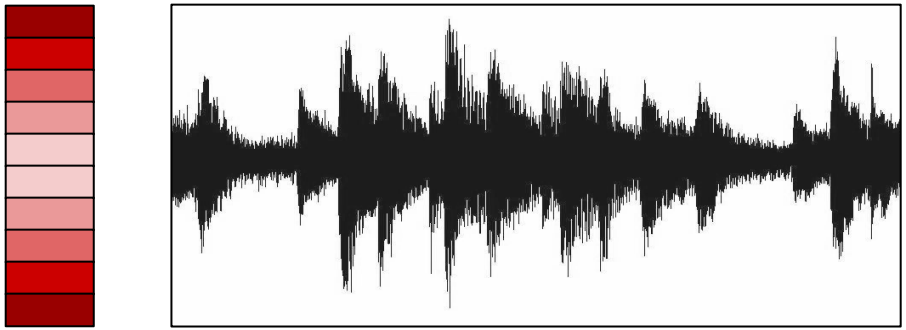
“WaveNet: A Generative Model for Raw Audio”, van den Oord et al. (2016)


1 Second

Modelling raw audio: μ-law encoding / quantisation

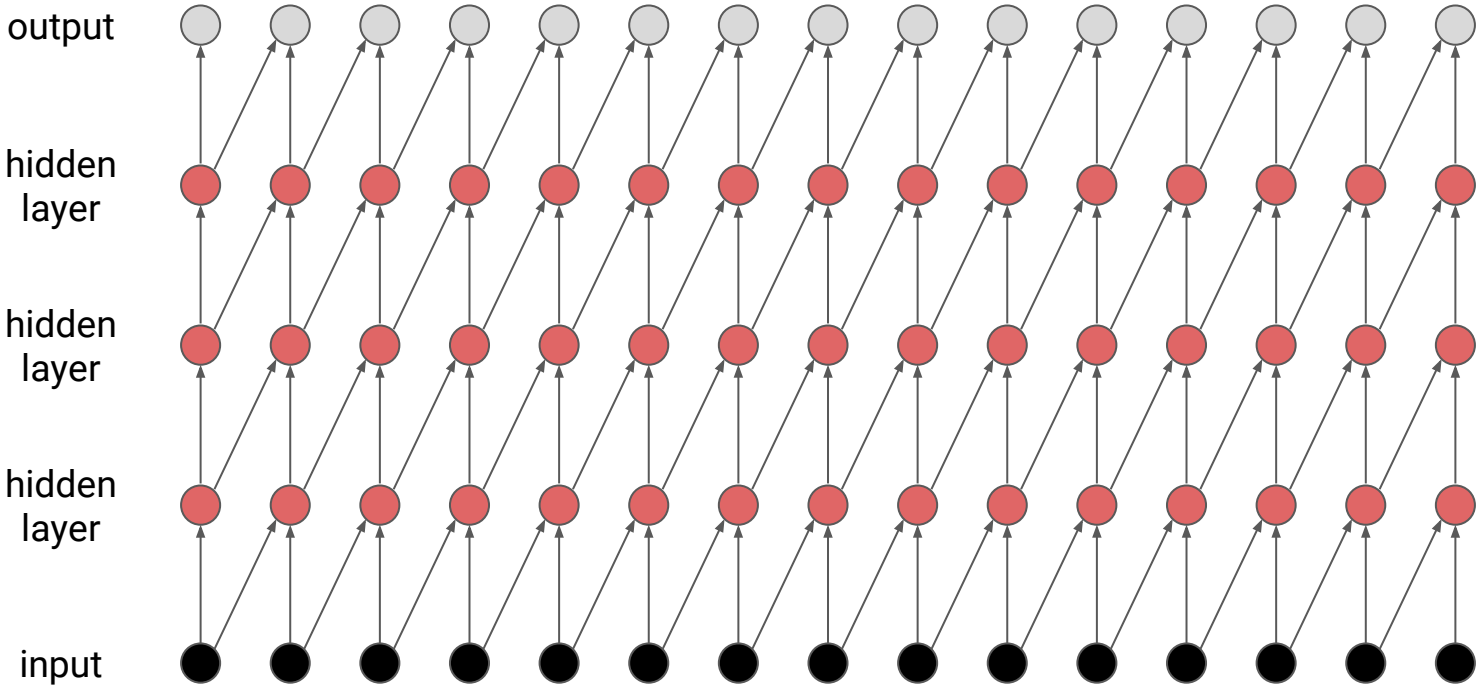
WaveNet models audio as a discrete time series

uniform quantisation



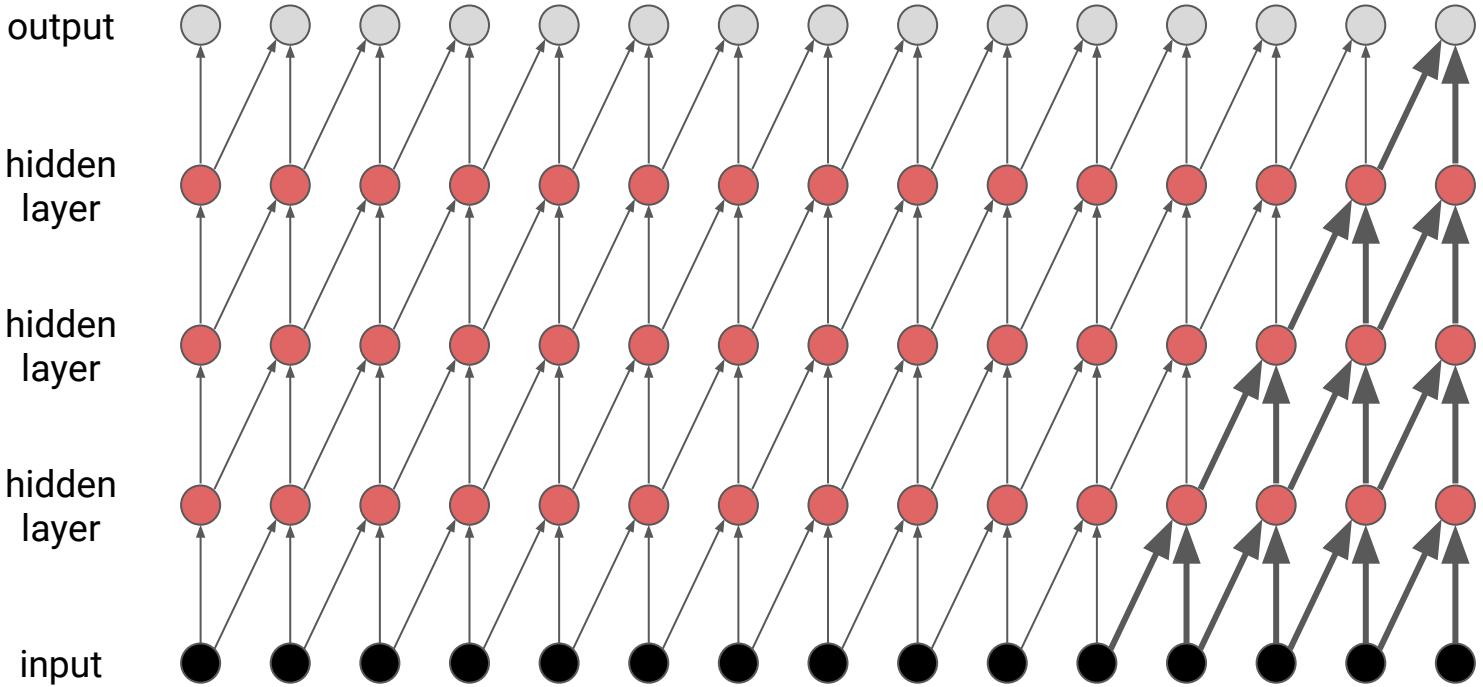
μ-law quantisation

WaveNet: an autoregressive model of raw audio

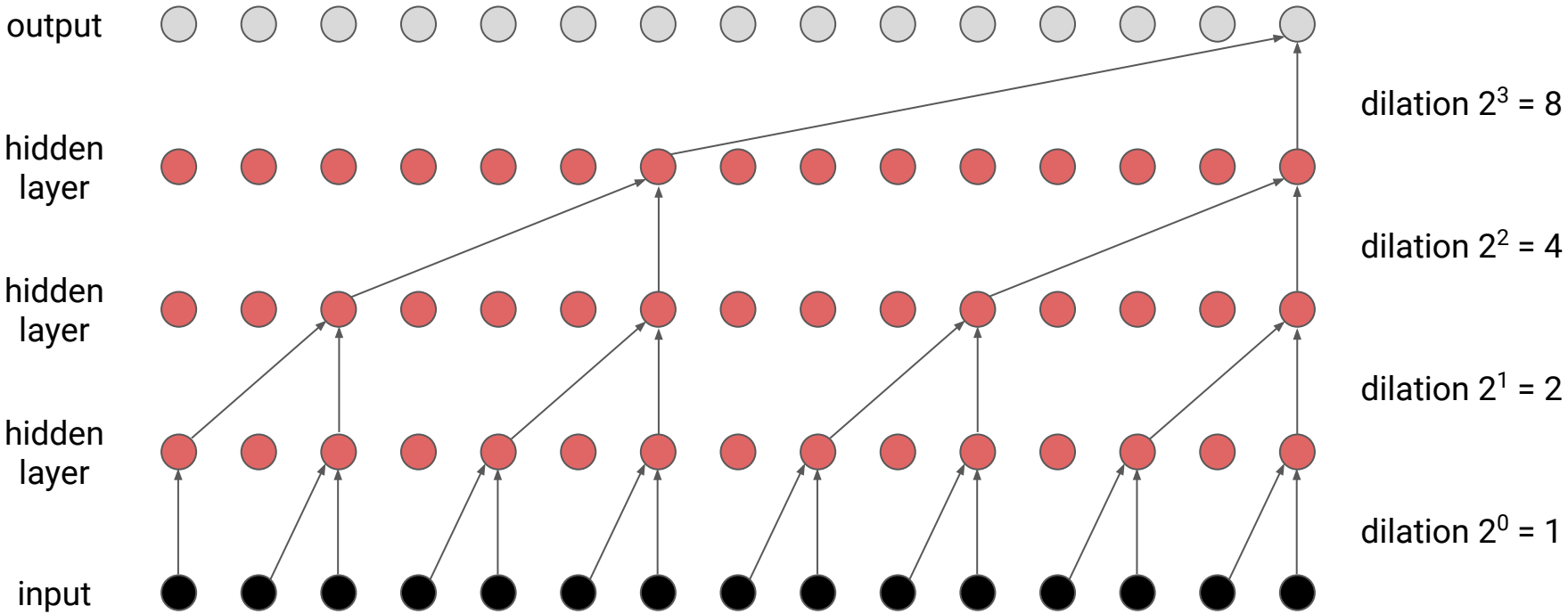


“WaveNet: A Generative Model for Raw Audio”, van den Oord et al. (2016)

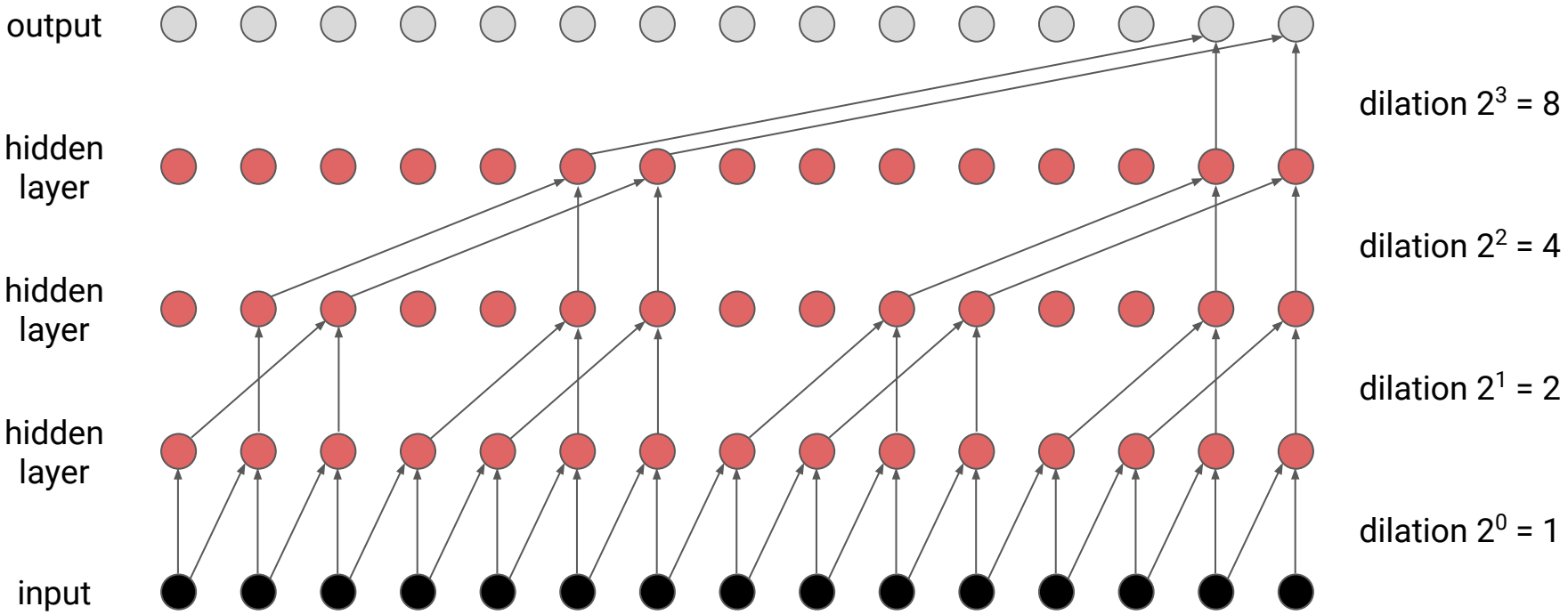
WaveNet: an autoregressive model of raw audio



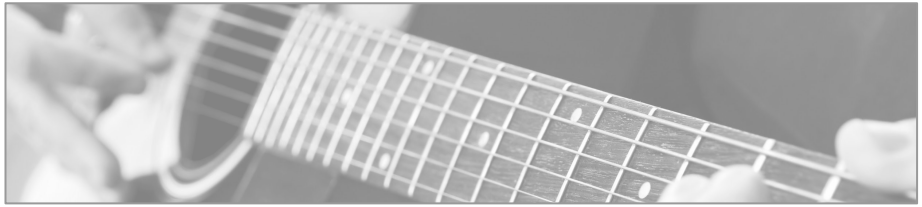
Dilated convolutions to enlarge receptive fields

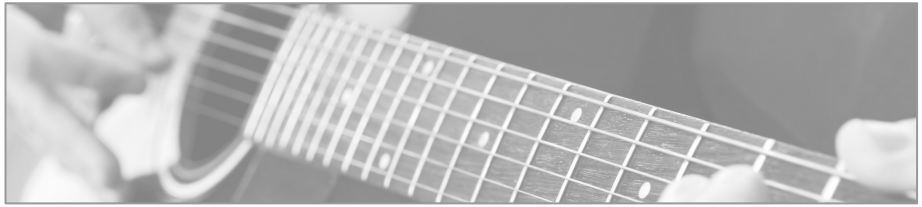


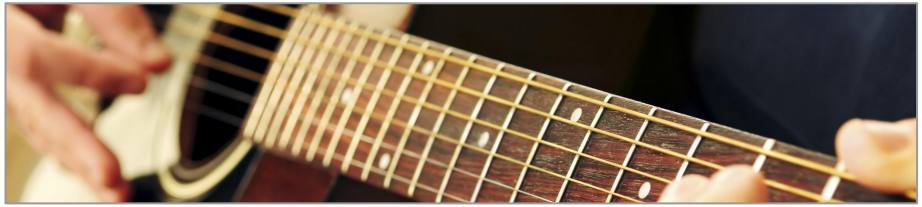
Dilated convolutions to enlarge receptive fields

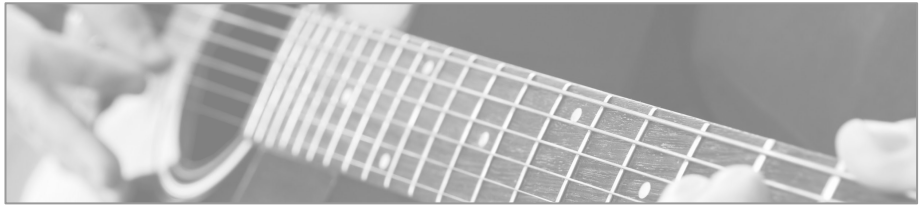


Generating music with WaveNets



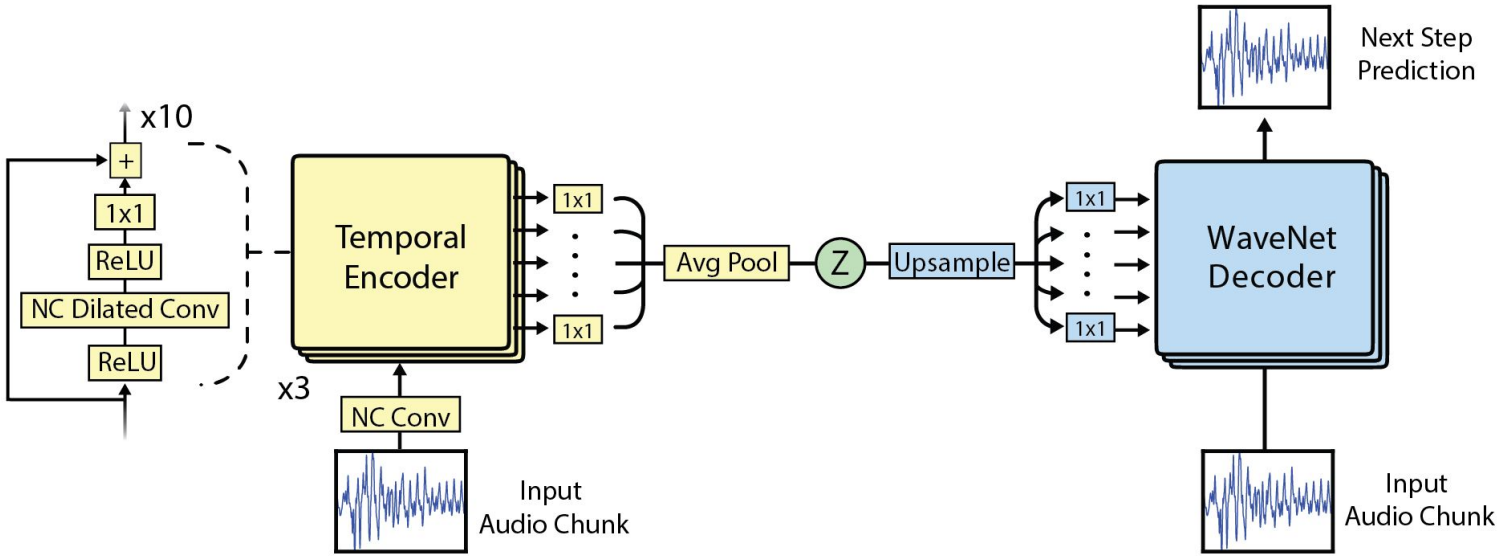






Aside: NSynth

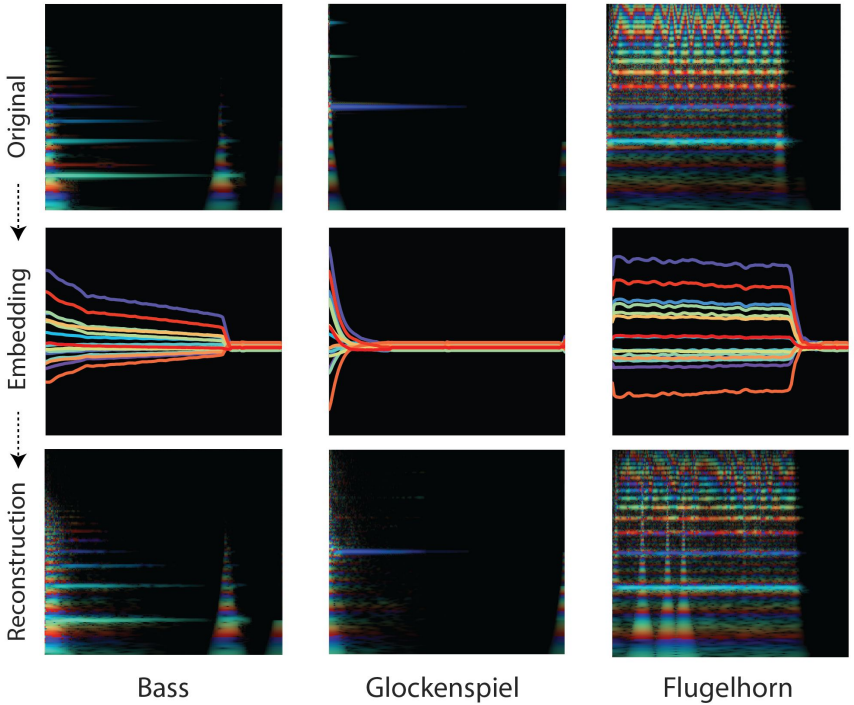
NSynth: WaveNet autoencoders for timbre modelling



Collaboration with magenta

“Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders”, Engel et al. (ICML 2017)

NSynth: WaveNet autoencoders for timbre modelling



Blog post: `https://magenta.tensorflow.org/nsynth`

Modelling long-range correlations with WaveNets

“The challenge of realistic music generation: modelling raw audio at scale”, Dieleman et al. (2018)

The challenge of realistic music generation: modelling raw audio at scale

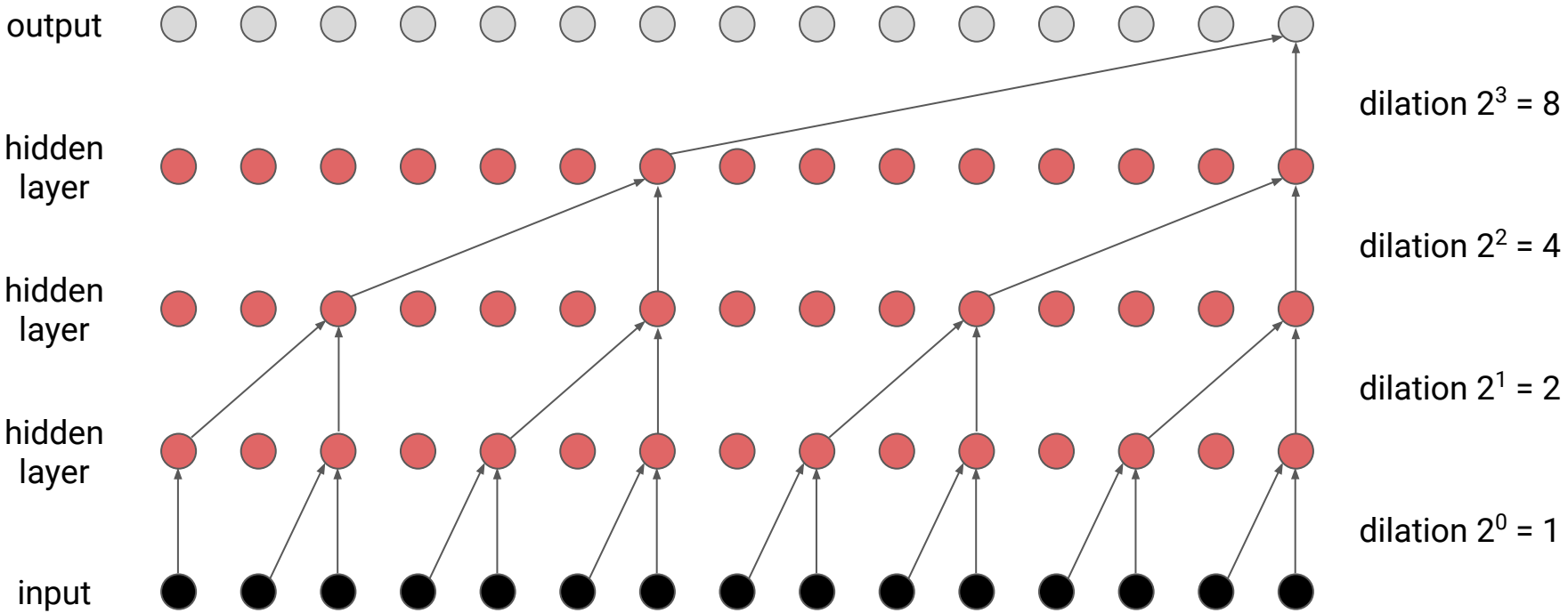
Sander Dieleman **Aäron van den Oord** **Karen Simonyan**
DeepMind
London, UK
{sedielem,avdnoord,simonyan}@google.com

Abstract

Realistic music generation is a challenging task. When building generative models of music that are learnt from data, typically high-level representations such as scores or MIDI are used that abstract away the idiosyncrasies of a particular performance. But these nuances are very important for our perception of musicality and realism, so in this work we embark on modelling music in the *raw audio* domain. It has been shown that autoregressive models excel at generating raw audio waveforms of speech, but when applied to music, we find them biased towards capturing local signal structure at the expense of modelling long-range correlations. This is problematic because music exhibits structure at many different timescales. In this work, we explore autoregressive discrete autoencoders (ADAs) as a means to enable autoregressive models to capture long-range correlations in waveforms. We find that they allow us to unconditionally generate piano music directly in the raw audio domain, which shows stylistic consistency across tens of seconds.

https://arxiv.org/abs/1806.10474

Dilation: #layers ~ log(receptive field length)



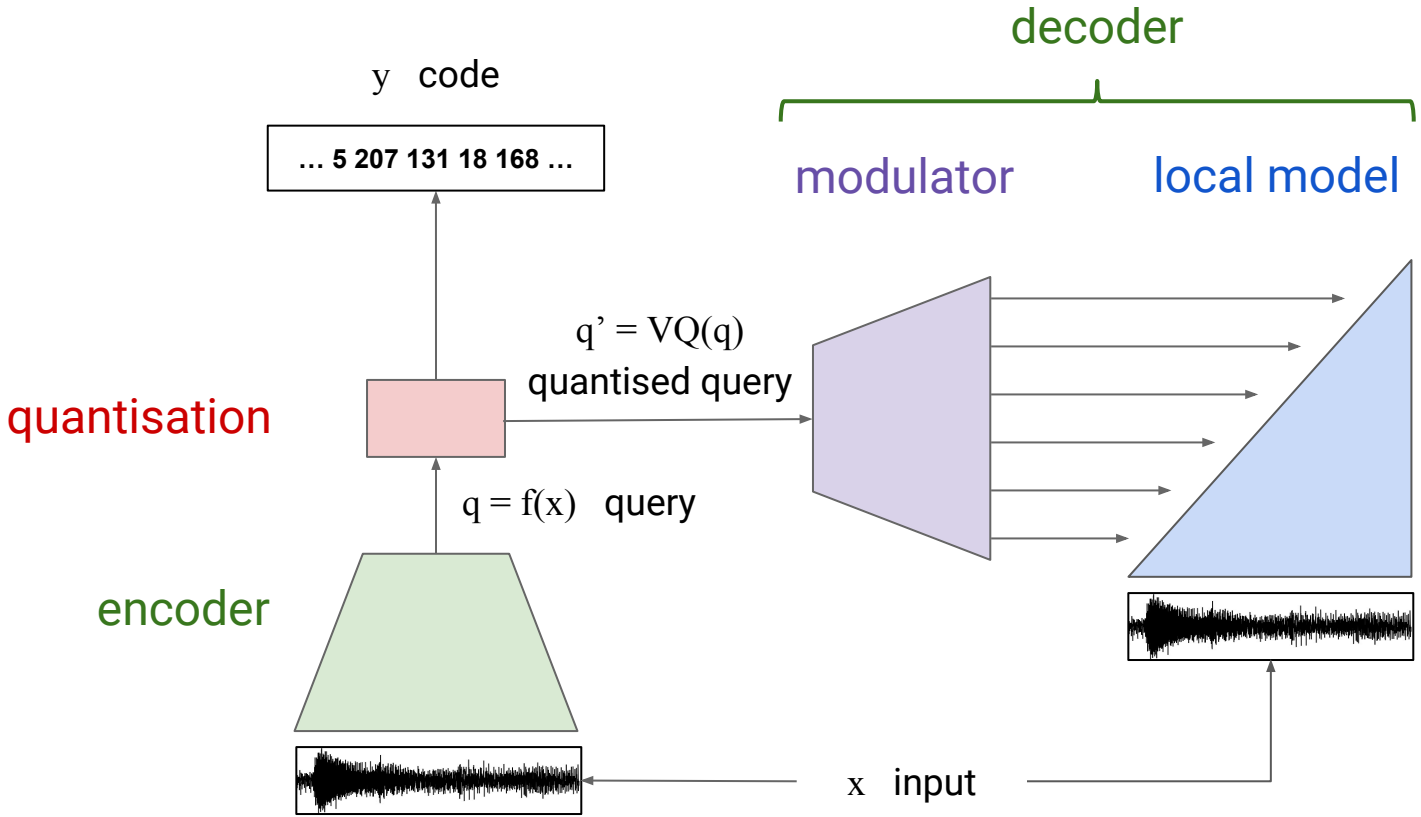
… but memory usage ~ receptive field length

Required model depth is **logarithmic** in the desired receptive field length

Required memory usage during training is still **linear** in the desired receptive field length!

⇒ We cannot scale indefinitely using dilation

Autoregressive discrete autoencoders (ADAs)



Autoregressive discrete autoencoders (ADAs)

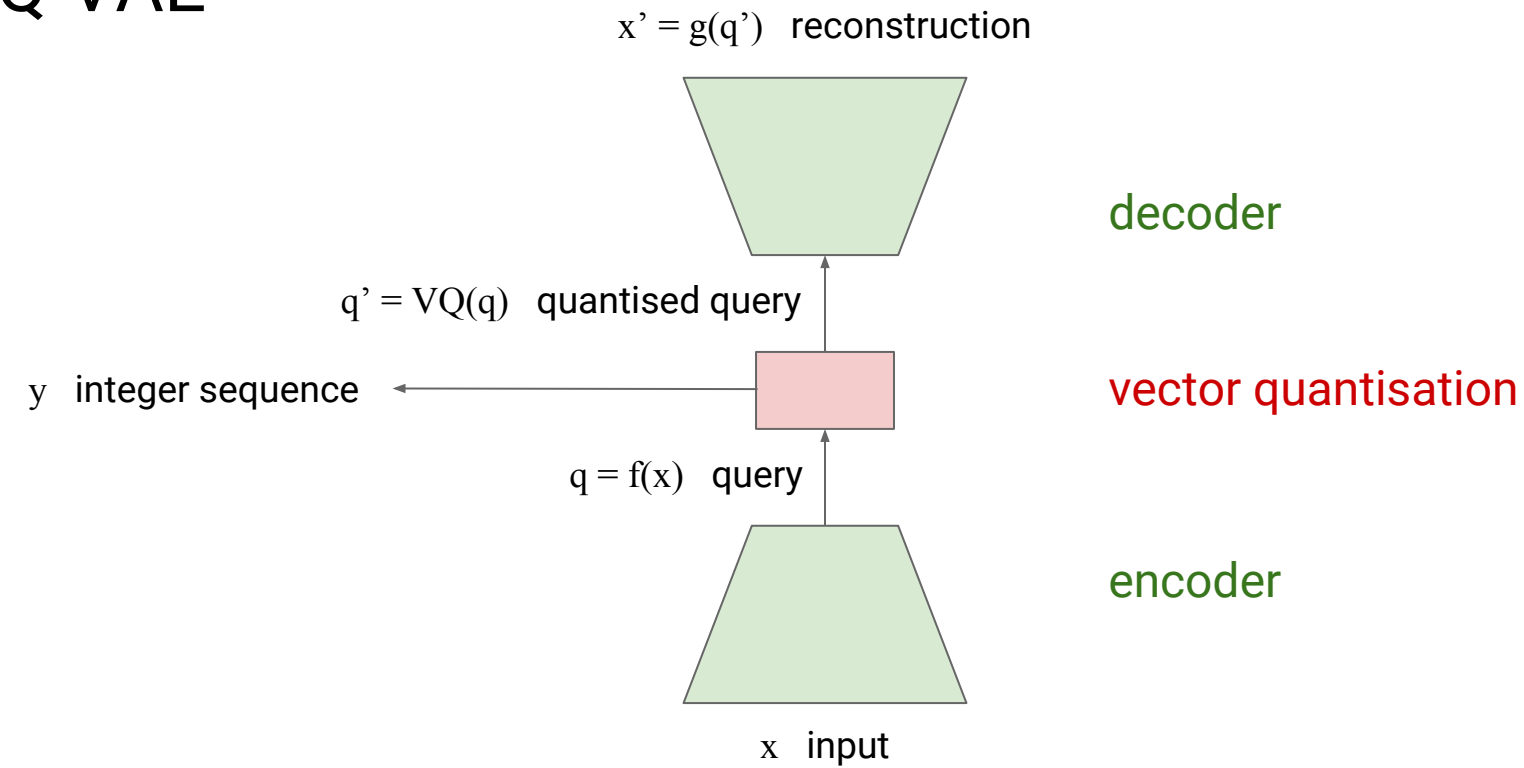
Two variants:

- **VQ-VAE:** vector quantisation variational autoencoder
  learn a codebook, **quantise** the **queries** to elements of this codebook

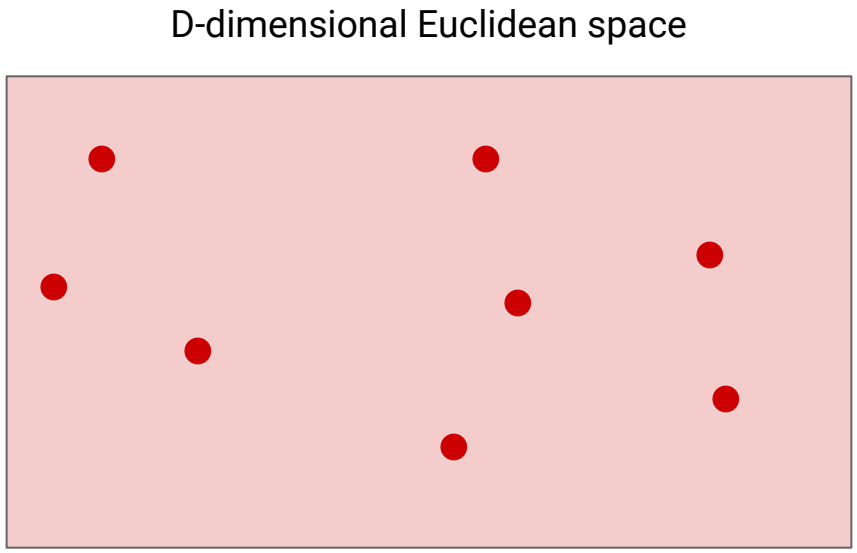
- **AMAE:** argmax autoencoder
  encourage the **queries** to be close to **one-hot** (e.g. use softmax), quantise them on the simplex

“Neural Discrete Representation Learning”, van den Oord et al. (2017)
“The challenge of realistic music generation: modelling raw audio at scale”, Dieleman et al. (2018)

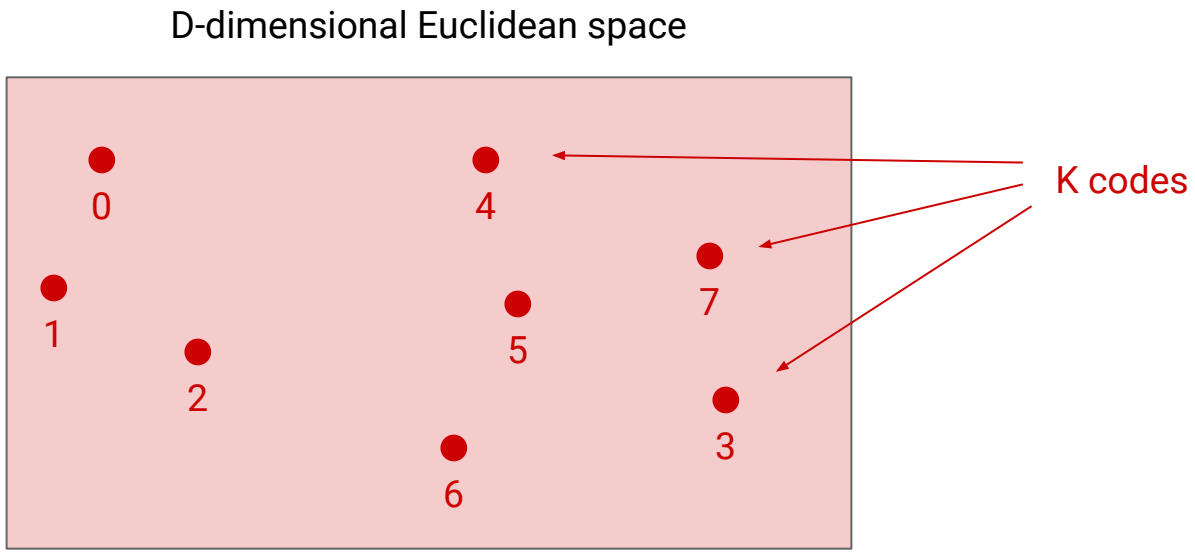
VQ-VAE


x' = g(q') reconstruction
decoder
q' = VQ(q) quantised query
y integer sequence
vector quantisation
q = f(x) query
encoder
x input

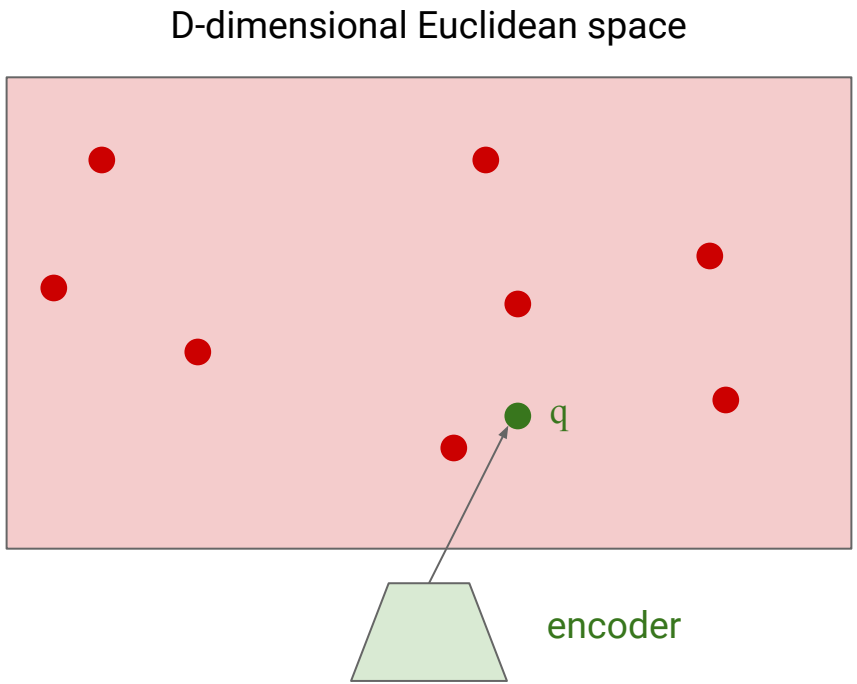
Vector quantisation



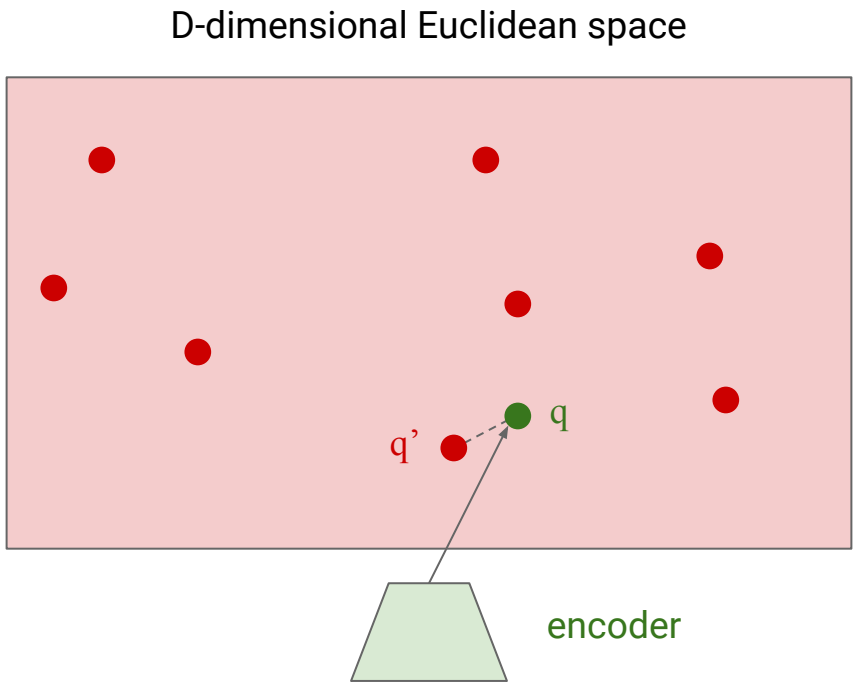
Vector quantisation



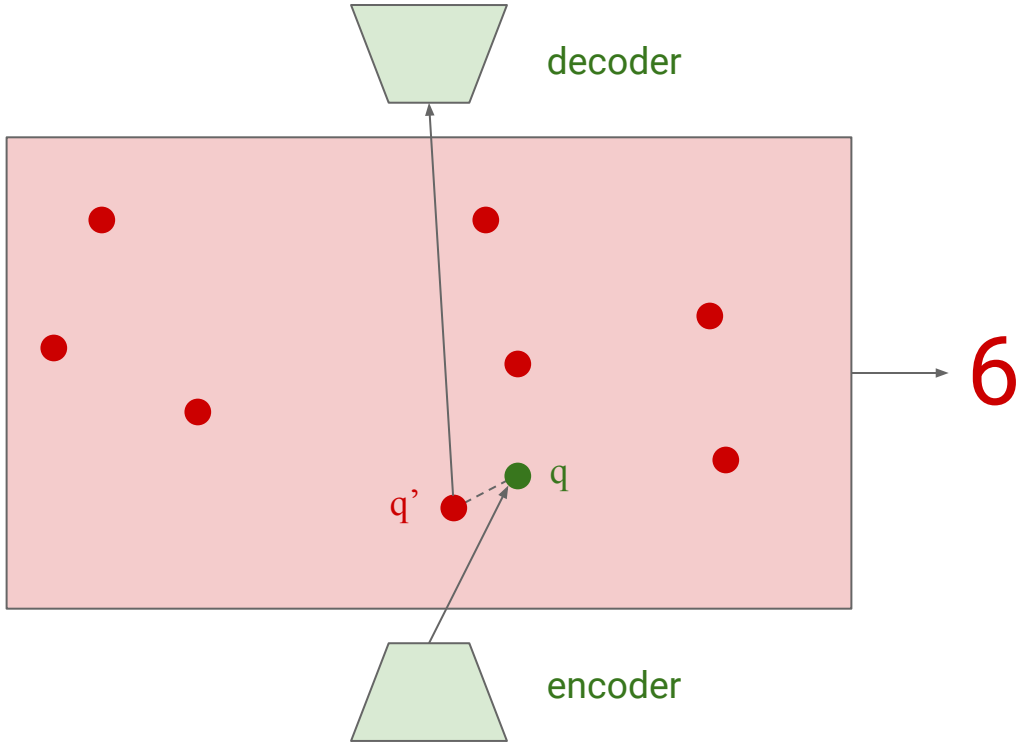
Vector quantisation



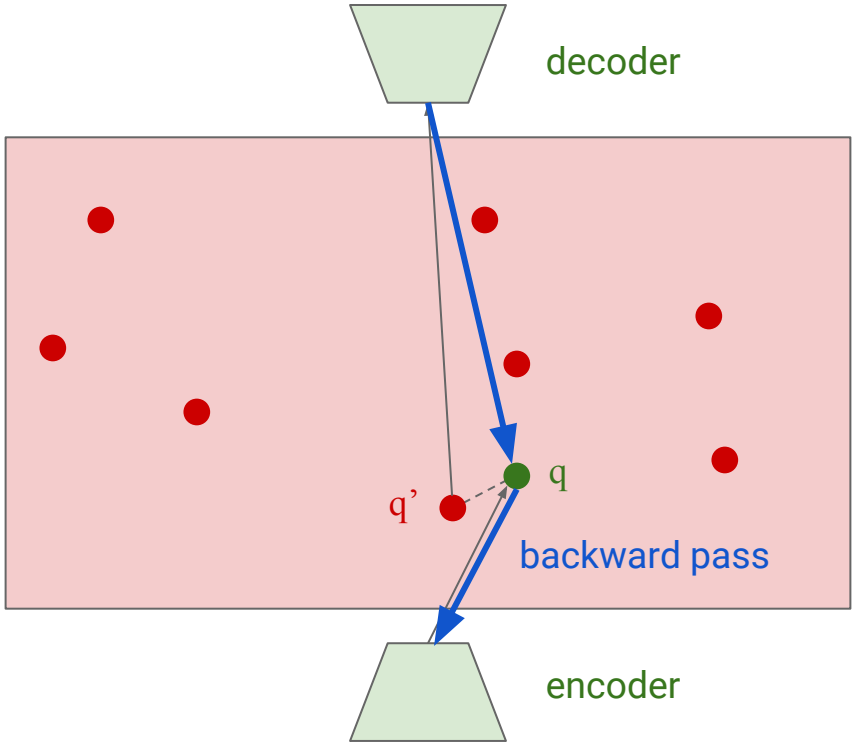
Vector quantisation



Vector quantisation



Straight-through estimation for VQ



Training VQ autoencoders

$$\mathcal{L} = -\log p(x|q')$$

Reconstruction loss (NLL)

Training VQ autoencoders

$$\mathcal{L} = -\log p(x|q')$$ Reconstruction loss (NLL)

$$+ \alpha \cdot (q' - [q])^2$$ Codebook loss

[...] = tf.stop_gradient(...)

Training VQ autoencoders

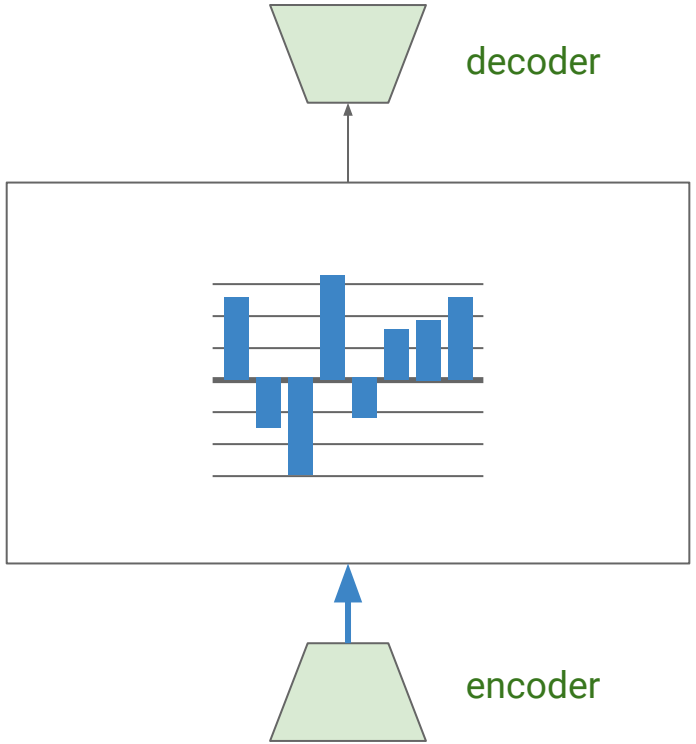
$$\mathcal{L} = -\log p(x|q')$$ Reconstruction loss (NLL)

$$+ \alpha \cdot (q' - [q])^2$$ Codebook loss

$$+ \beta \cdot ([q'] - q)^2$$ Commitment loss

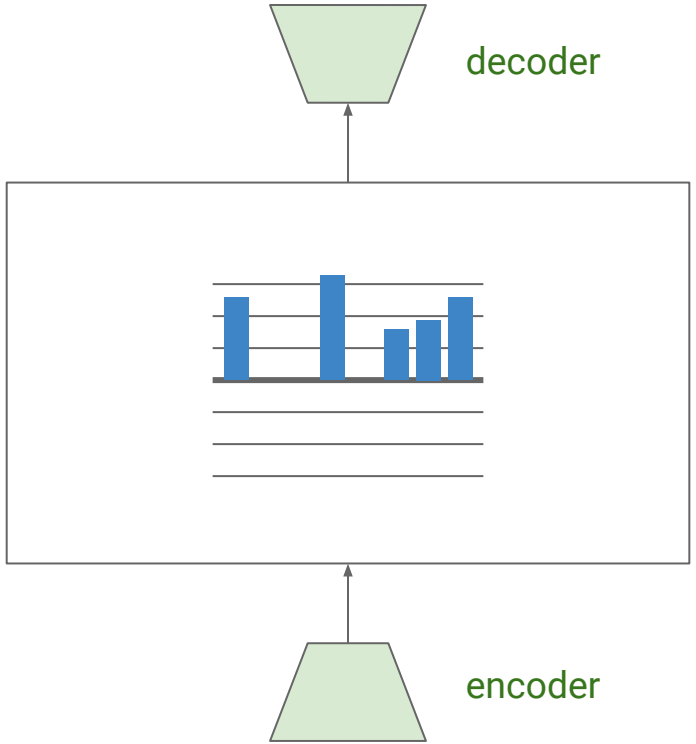
[...] = tf.stop_gradient(...)

Argmax autoencoder (AMAE)



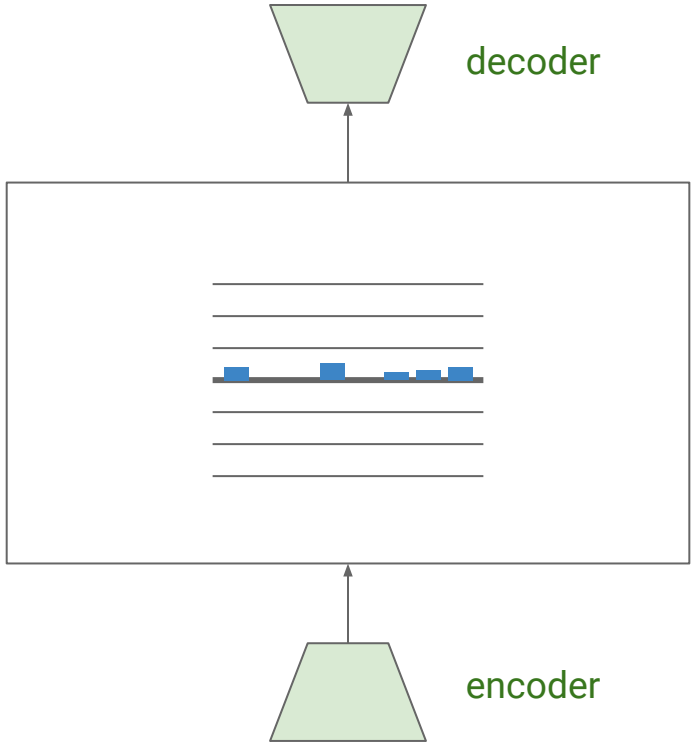
Encoder outputs real values

Argmax autoencoder (AMAE)



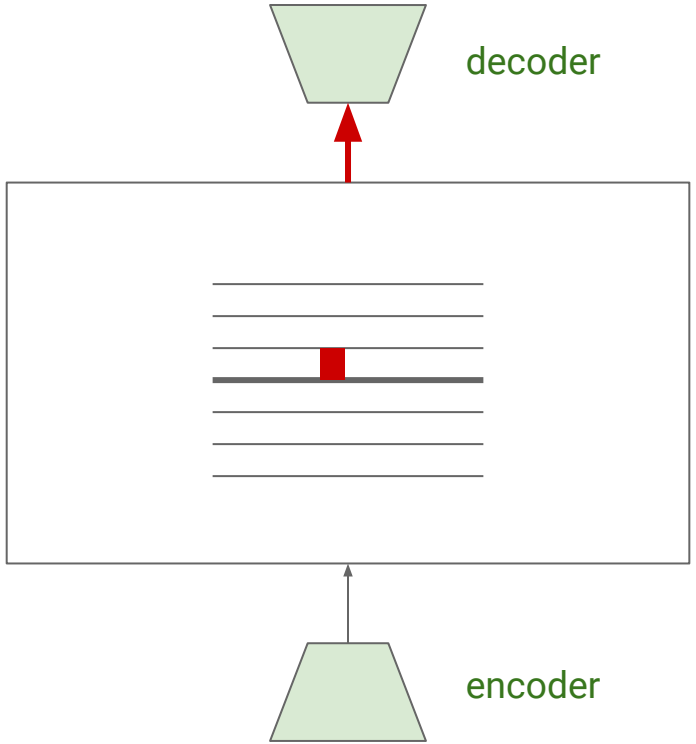
apply ReLU

Argmax autoencoder (AMAE)



Normalise so that activations sum to 1

Argmax autoencoder (AMAE)



Perform argmax
and pass to decoder

AMAE loss function

$$\mathcal{L} = -\log p(x|q')$$ Reconstruction loss (NLL)

$$+ \beta \cdot (q' - q)^2$$ Commitment loss

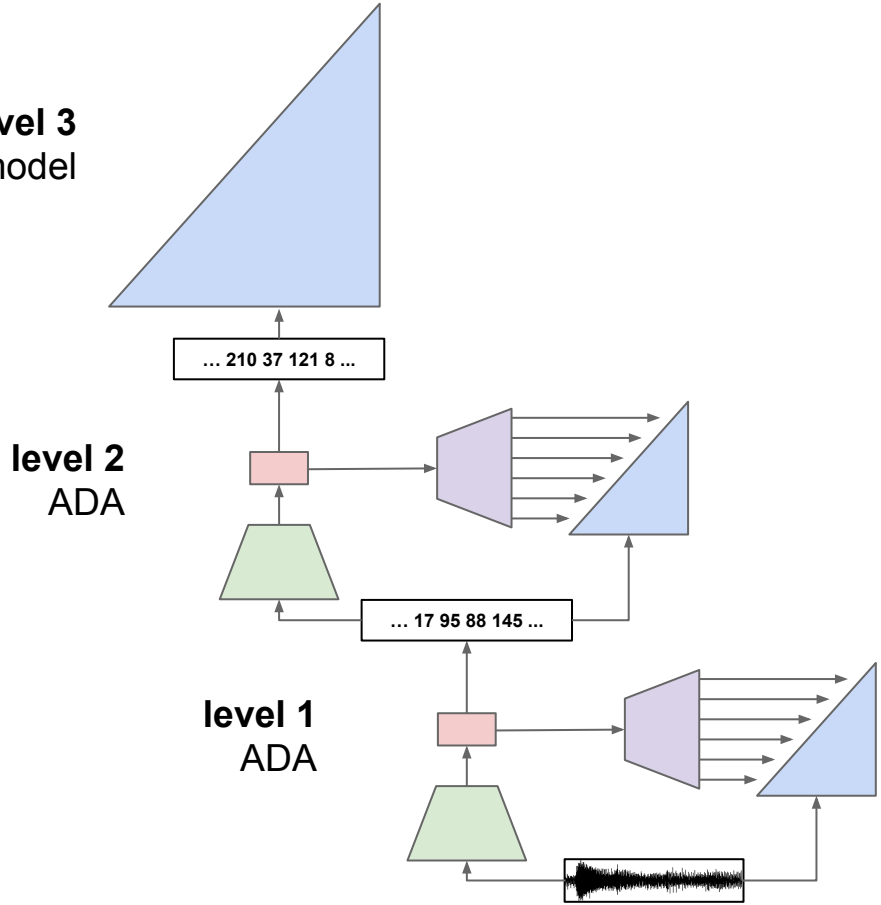
$$+ \nu \cdot (E_t[q] - C^{-1})^2$$ Diversity loss

AMAE: alternative interpretation

AMAE = **VQ-VAE with a fixed codebook**, consisting of one-hot vectors

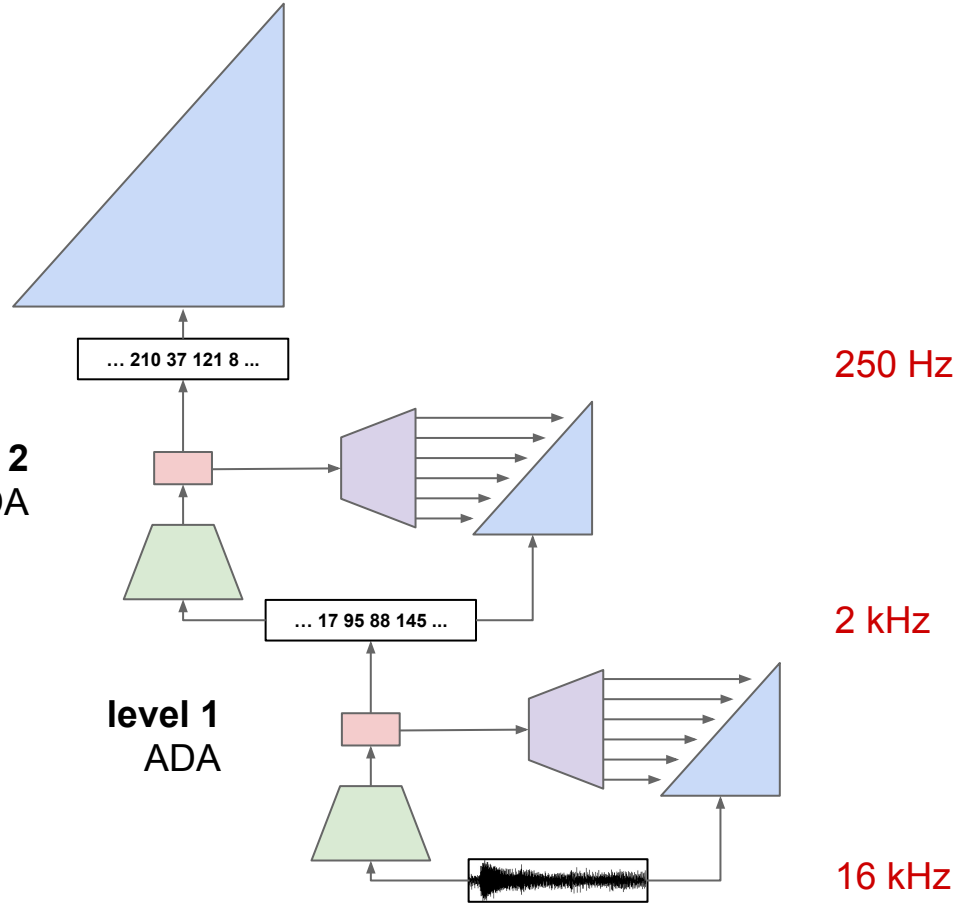
Stacking ADAs

level 3
unconditional model



Stacking ADAs

level 3
unconditional model



level 2
ADA

level 1
ADA

250 Hz

2 kHz

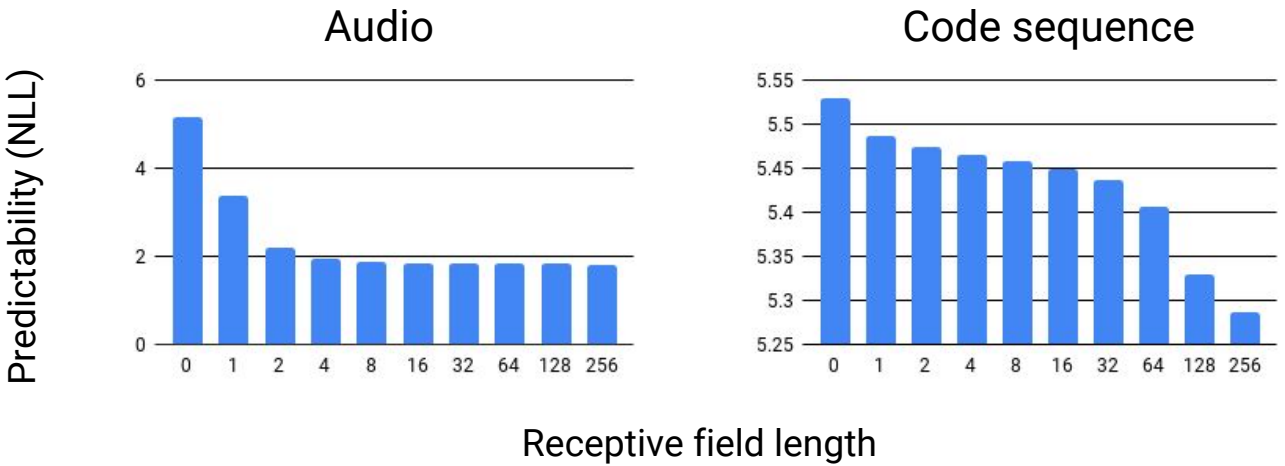
16 kHz

Code sequences are harder to model than audio

Training second level ADAs is much harder

- Train VQ-VAE with population-based training (PBT)
- Train AMAE, which is less unstable

"Population Based Training of Neural Networks", Jaderberg et al. (2017)



Generating music
with long-range consistency

Some results from a human evaluation

Model	Num. levels	RF	Human evaluation Fidelity	Musicality
Large WaveNet	1	384 ms	3.82 ± 0.18	2.43 ± 0.14
Very large WaveNet	1	768 ms	3.82 ± 0.20	2.89 ± 0.17
Thin WaveNet with large RF	1	3072 ms	2.43 ± 0.17	1.71 ± 0.18
hop-8 VQ-VAE + large WaveNet	2	3072 ms	3.79 ± 0.16	3.04 ± 0.16
hop-64 VQ-VAE + large WaveNet	2	24576 ms	3.54 ± 0.18	3.07 ± 0.17
VQ-VAE + PBT-VQ-VAE + large WaveNet	3	24576 ms	3.71 ± 0.18	4.04 ± 0.14
VQ-VAE + AMAE + large WaveNet	3	24576 ms	3.93 ± 0.18	3.46 ± 0.15

Samples

https://bit.ly/2IPXoDu

Machine Learning for Creativity and Design

NIPS 2018 Workshop, Montreal, Canada

Saturday December 8th 8:30 — 18:00

https://nips2018creativity.github.io/

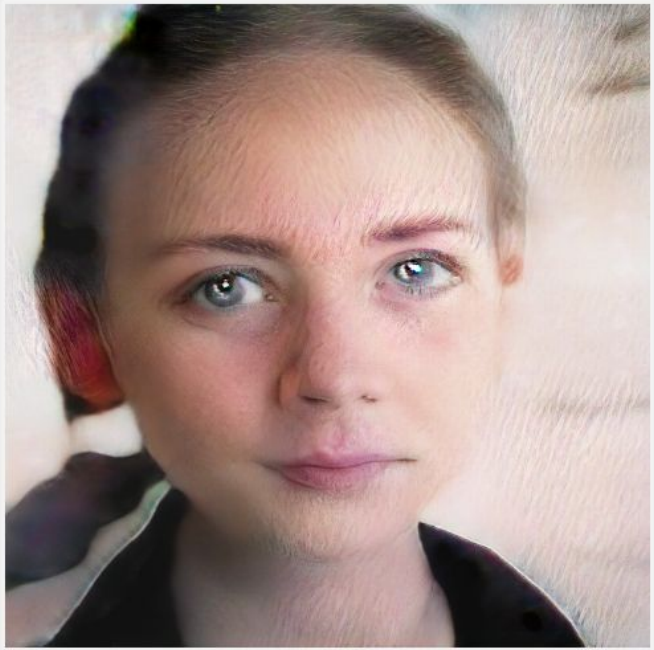


Image credit: Mike Tyka's Portraits of Imaginary People from the NIPS 2017 creativity art gallery.

Keynote speakers

Kenneth Stanley, University of Central Florida

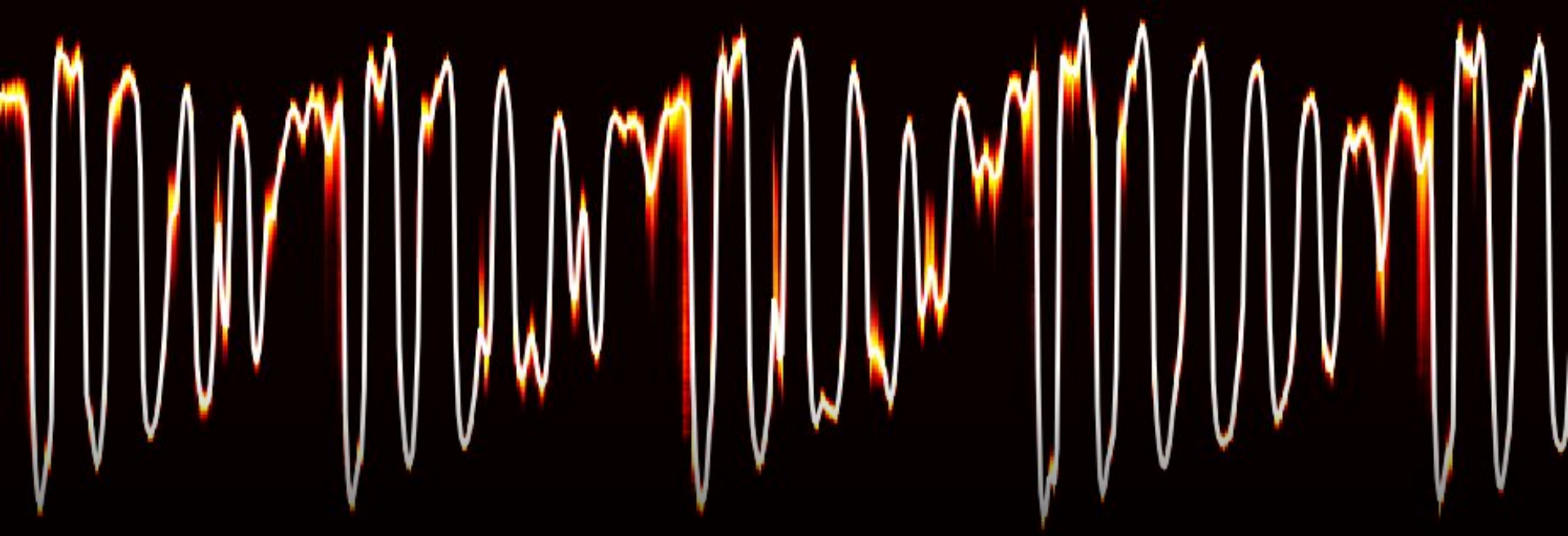
David Ha, Google Brain

Allison Parrish, NYU ITP

Yaroslav Ganin, DeepMind

Yaniv Taigman, Facebook AI Research

Submission deadline (papers & art): October 28th



WaveNet blog post:

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

@sedielem