웨이브 데이터를 활용한 음악 생성 모델 톺아보기

박수철
Rubato Lab
scpark20@gmail.com

박수철

- 베이지안
- Autoregressive, VAE, Flow, GAN 등 생성모델 연구&개발

- 모두의 연구소 2년차
- 음성합성 등 딥러닝 3년차

- 안드로이드 앱 개발 3년
- 음악 DSP 모듈 개발 1년
- WEB 개발 2년

- 바하, 말러 덕후
- 달리기, 풀업 초보

모두의연구소
Rubato Lab

1. 웨이브 데이터란?

1.1 Wave 1.2 Spectrogram 1.3 Mel-Spectrogram

2. Autoregressive Models

2.1 Wavenet 2.2 VQ-VAE 2.3 Melnet 2.4 Sparse Transformer

3. Flow-Based Models

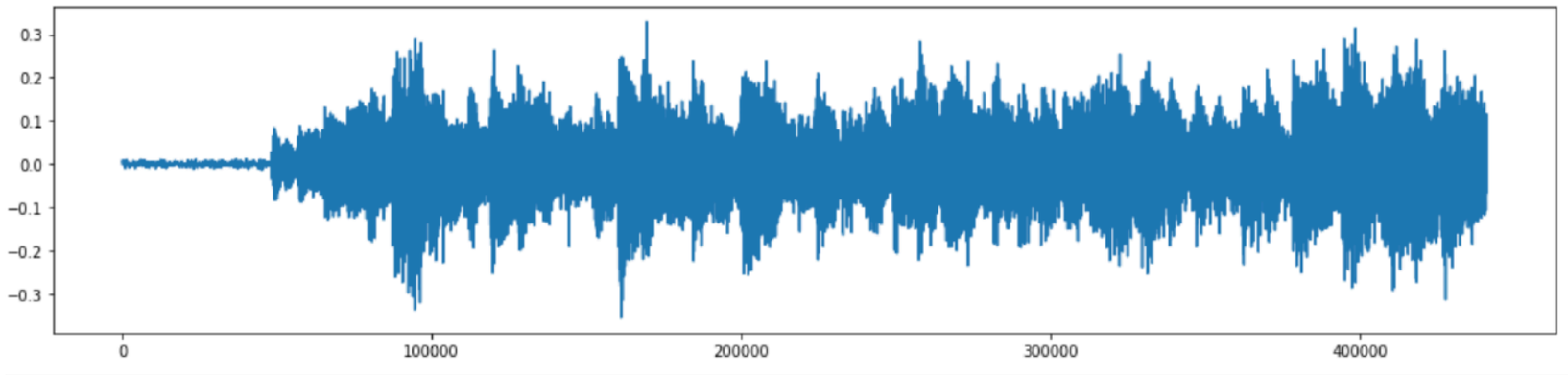
3.1 NICE&RealNVP&Glow 3.2 FlowSeq 3.3 MelFlow

4. 결론

웨이브 데이터란?

1. 웨이브 데이터란?

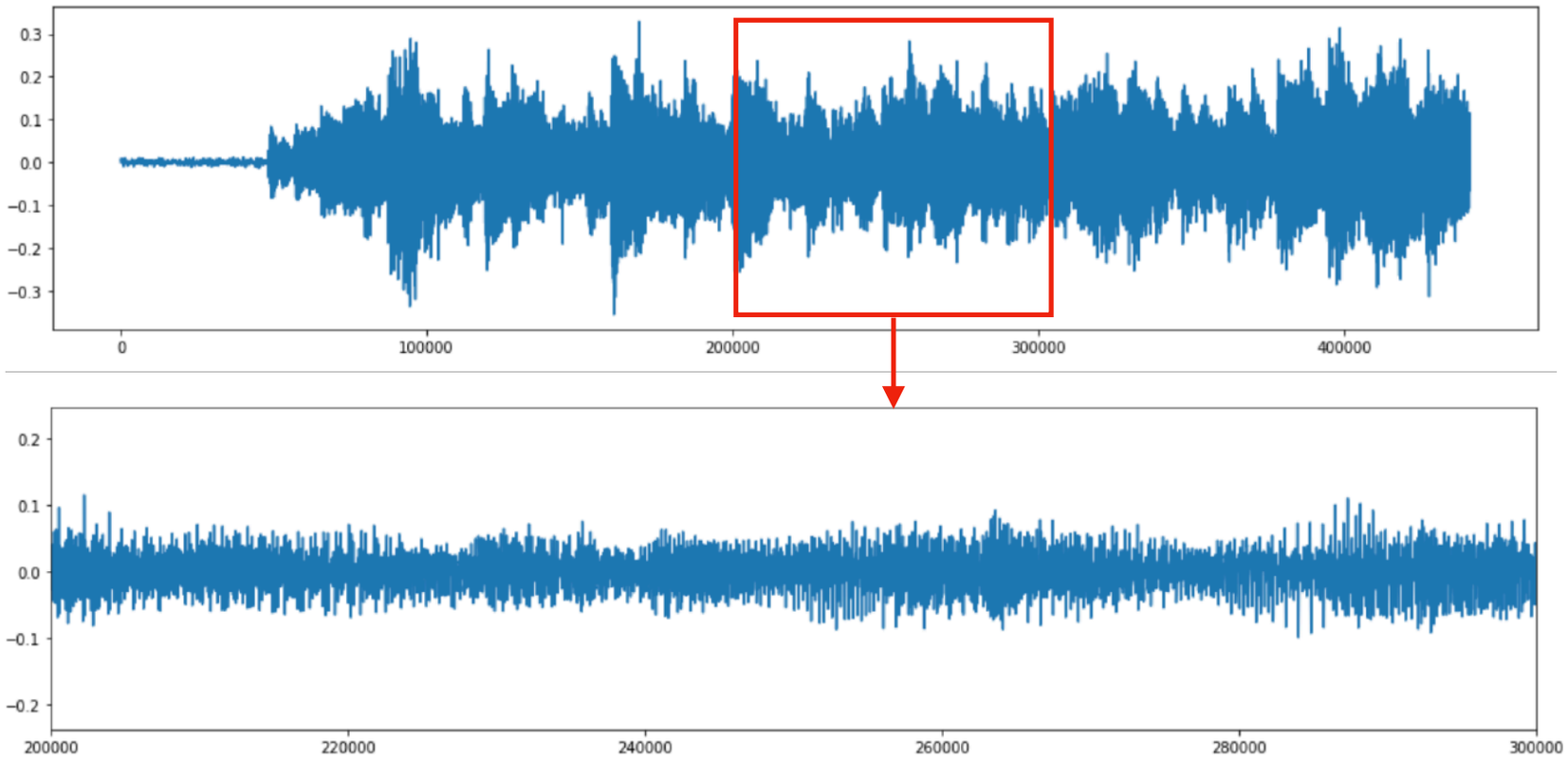
소리를 디지털화하여 기록한 것



Maestro Dataset, Chopin Etude Op.25 No.1

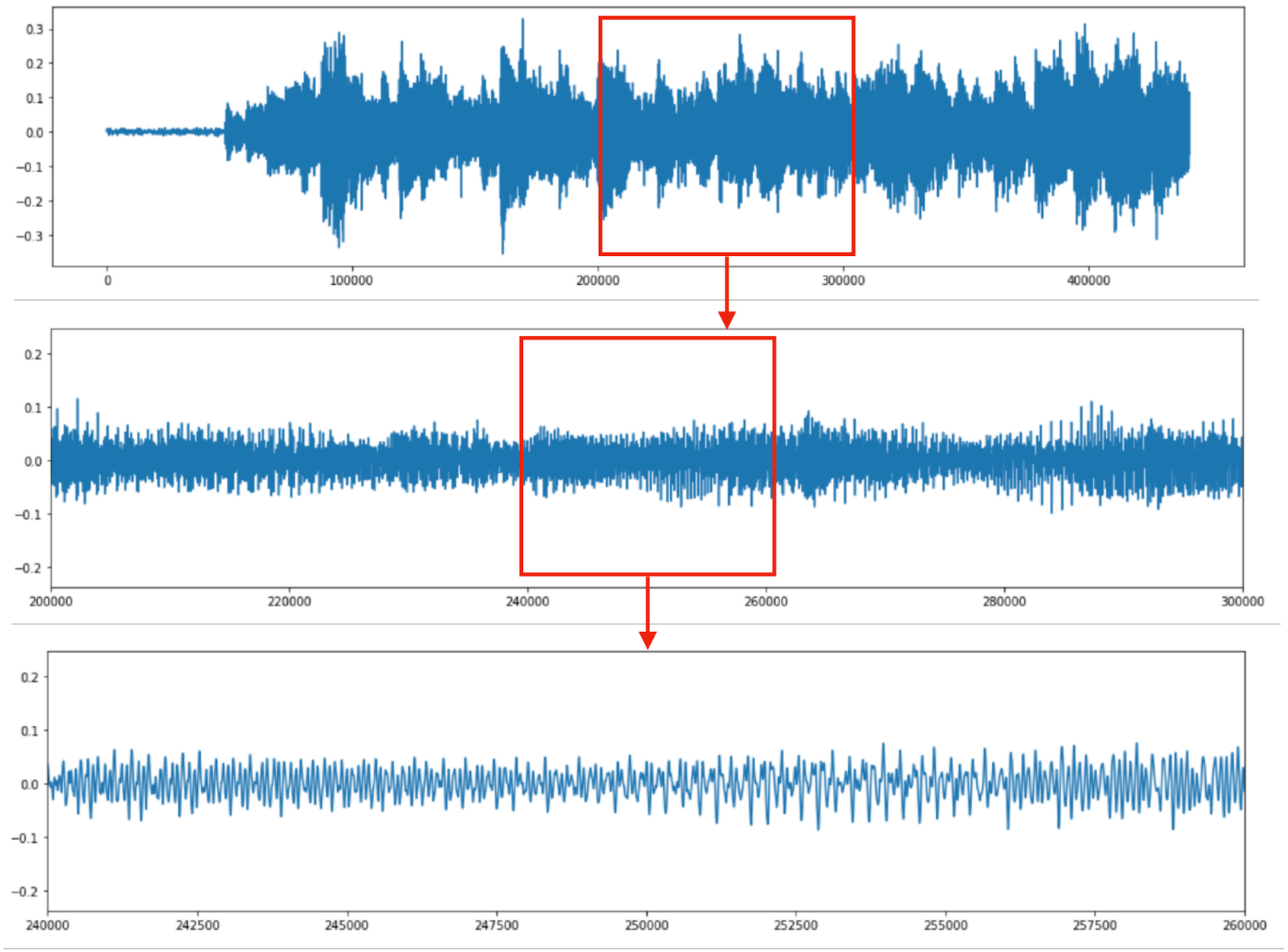
1. 웨이브 데이터란?

소리를 디지털화하여 기록한 것



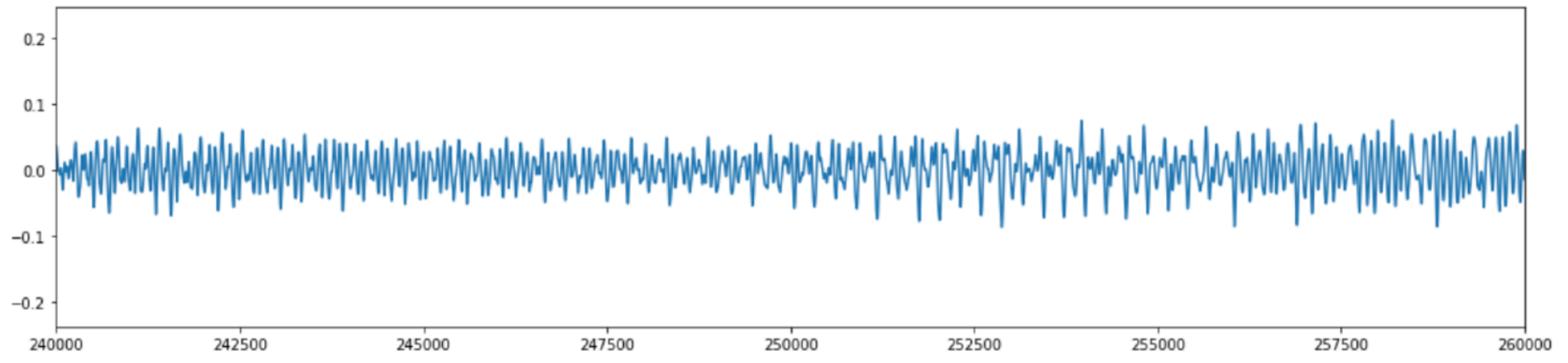
1. 웨이브 데이터란?

소리를 디지털화하여 기록한 것



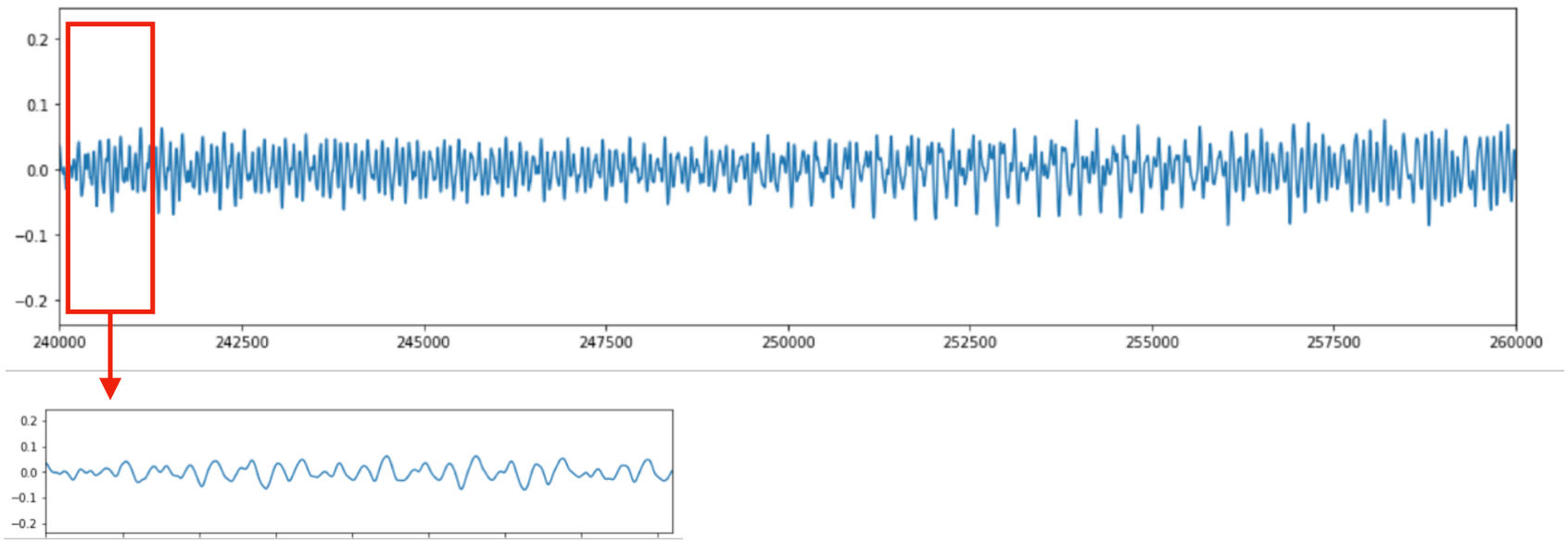
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것



Spectrogram?

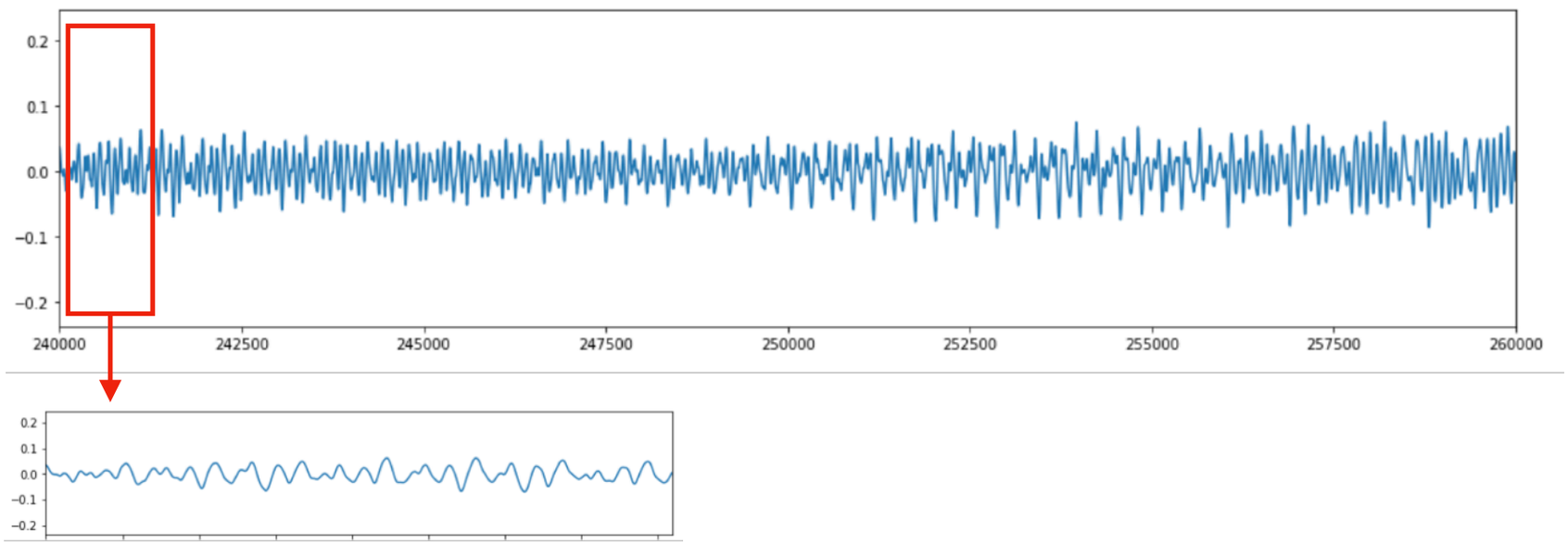
웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것



Frame

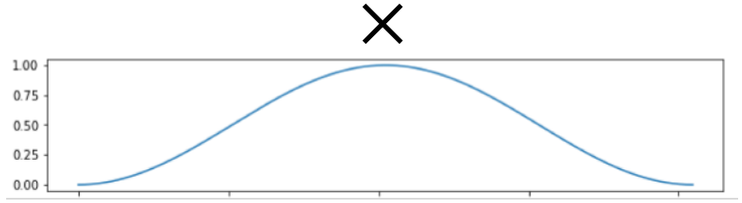
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것



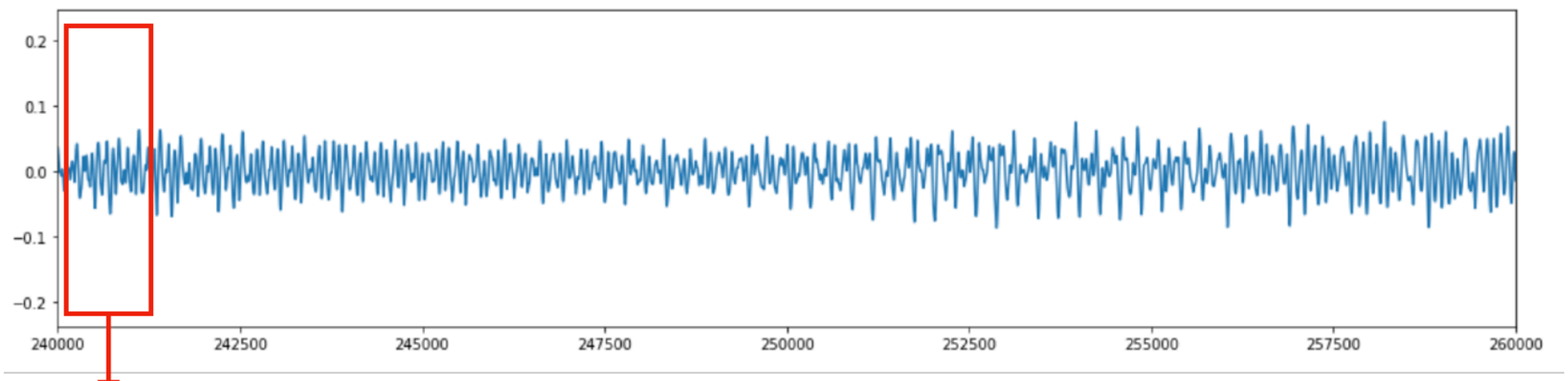
Frame

Window

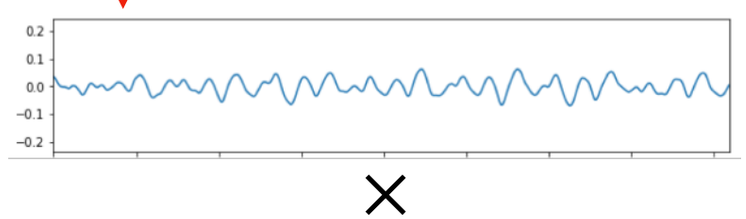


Spectrogram?

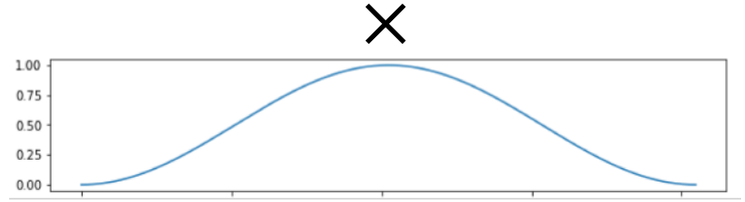
웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것

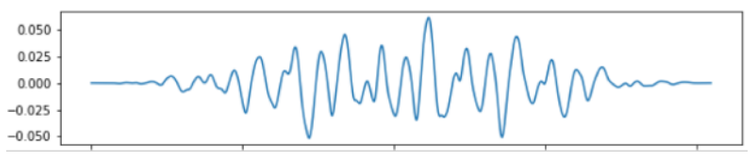


Frame



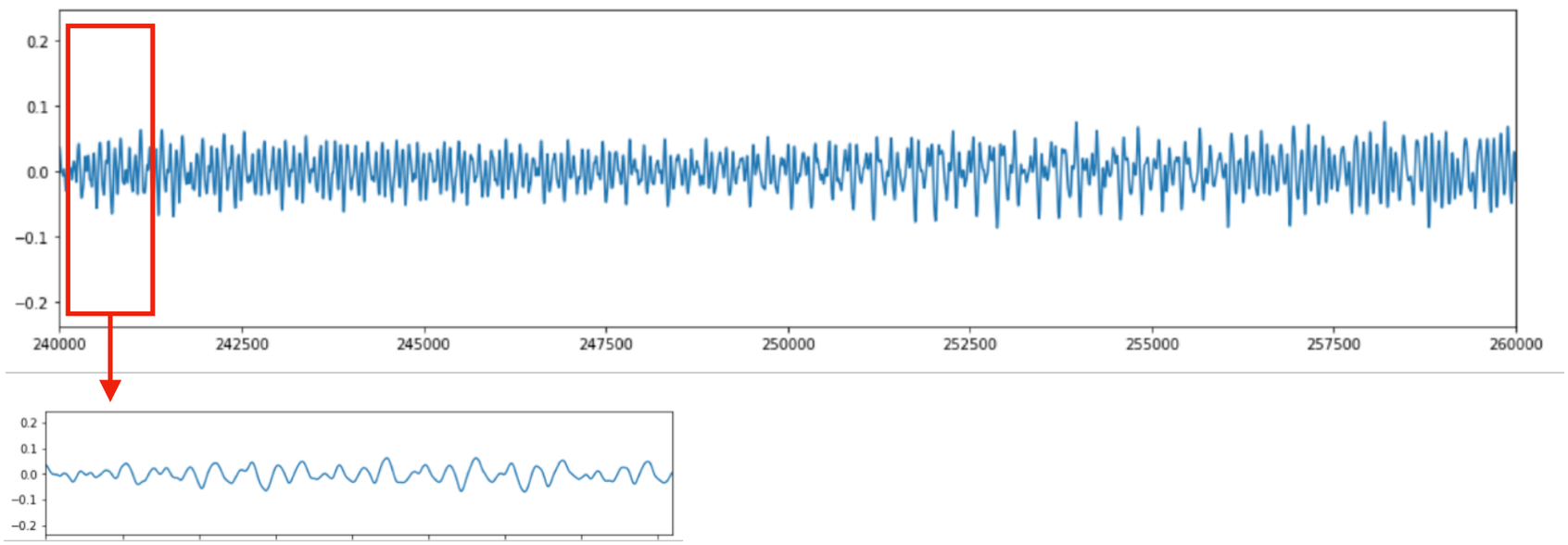
Window





Spectrogram?

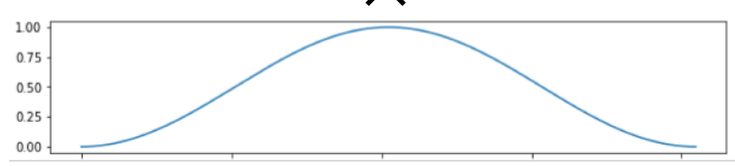
웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것



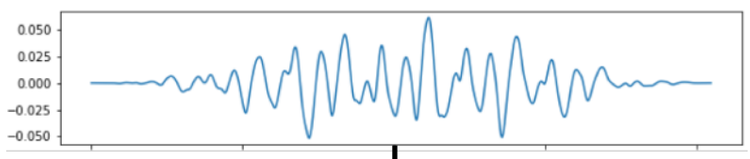
Frame

×

Window

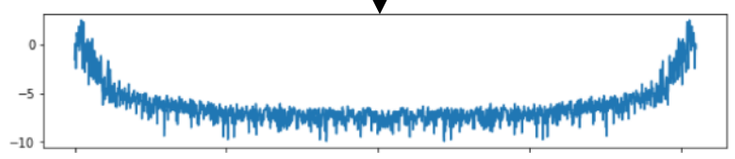


=



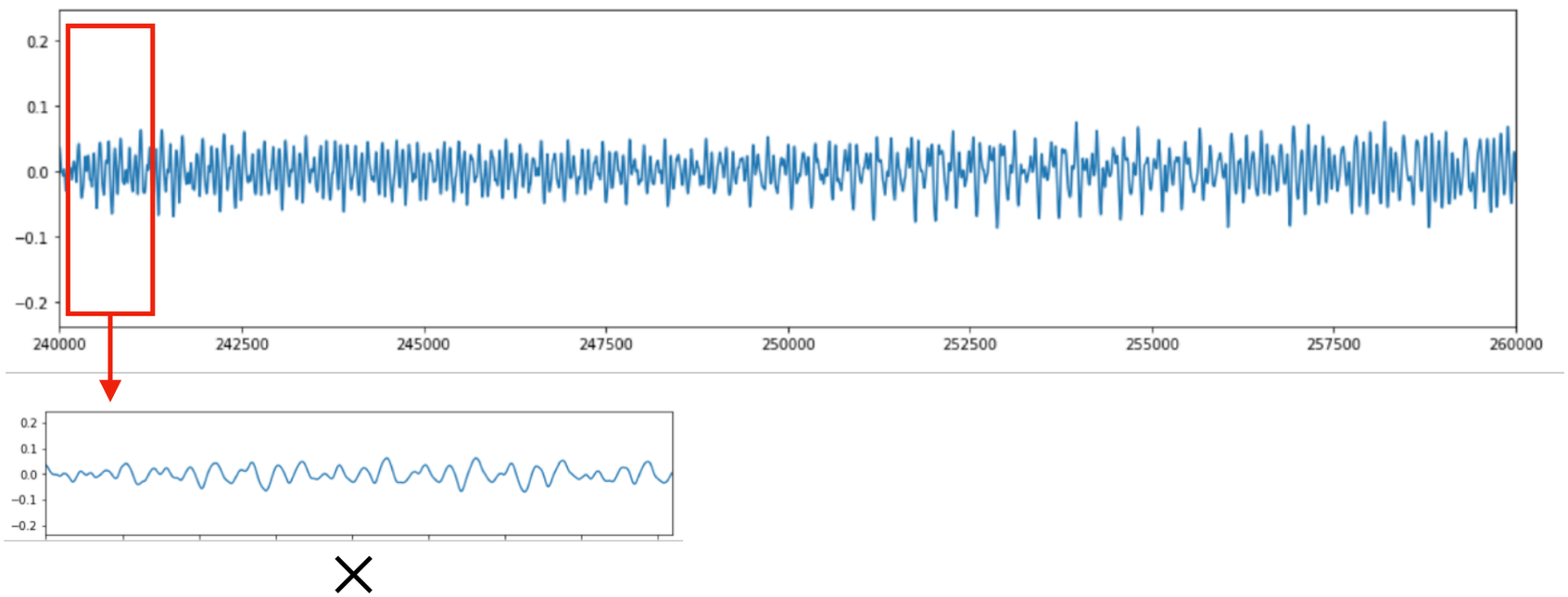
Discrete Fourier Transform

Spectrum



Spectrogram?

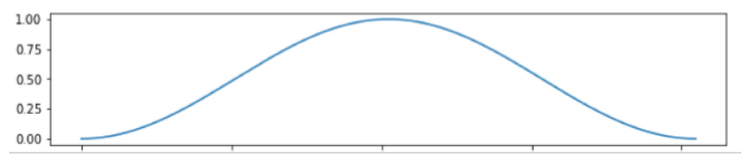
웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것



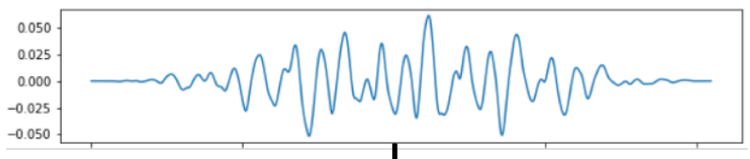
Frame

Window

×

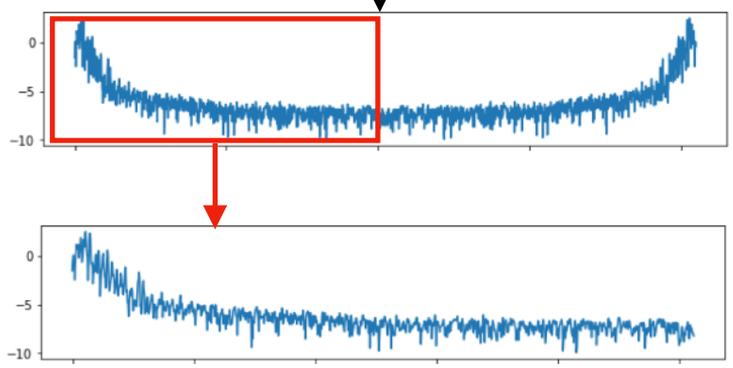


=



Discrete Fourier Transform

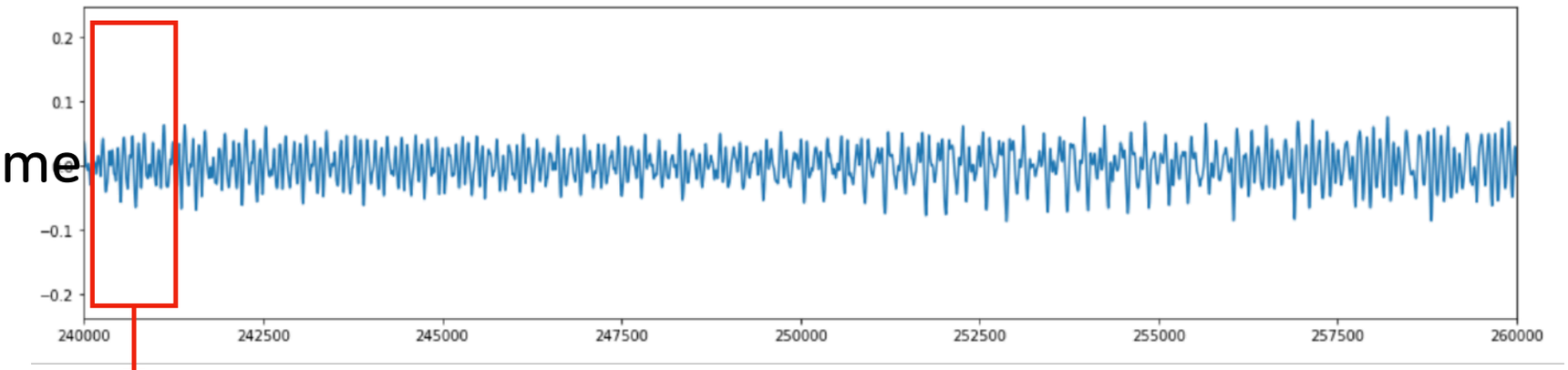
Spectrum



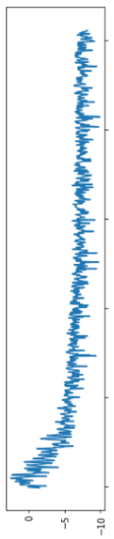
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것

Frame



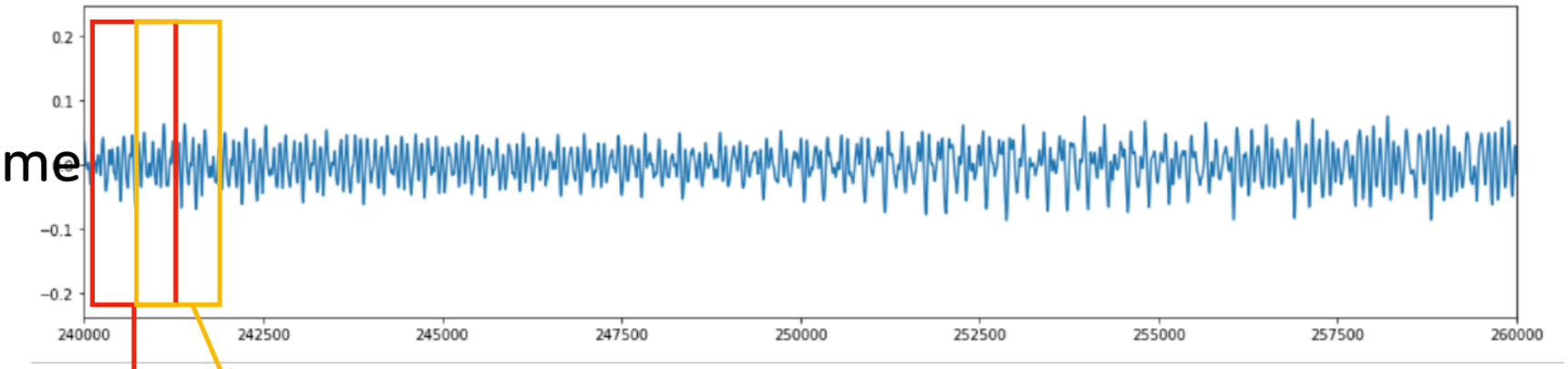
Spectrum



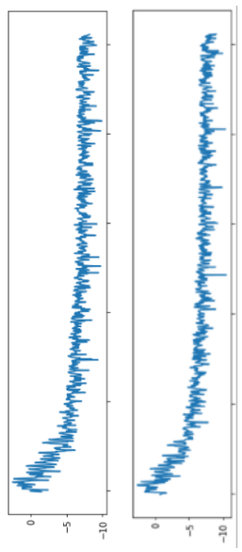
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것

Frame



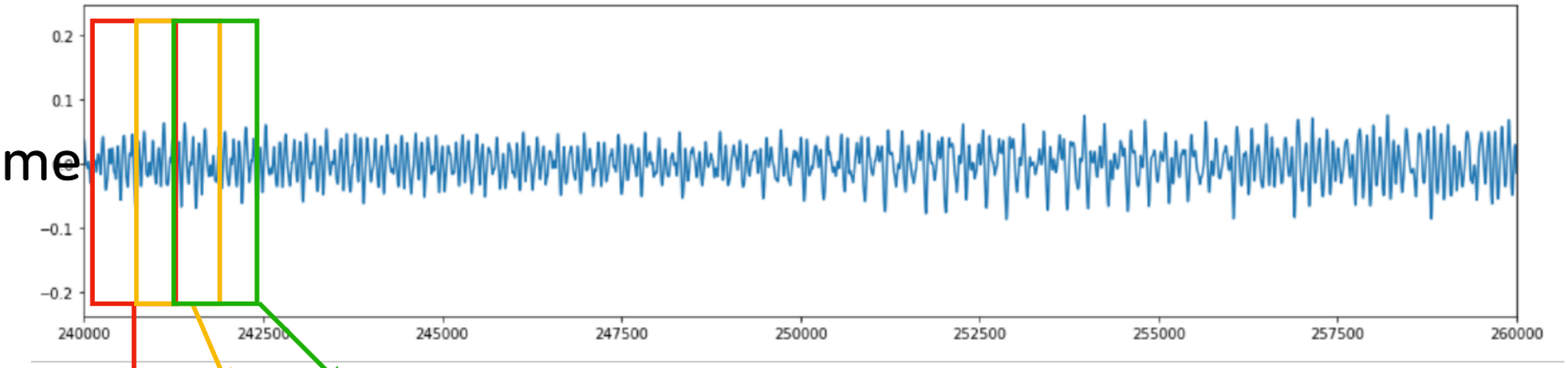
Spectrum



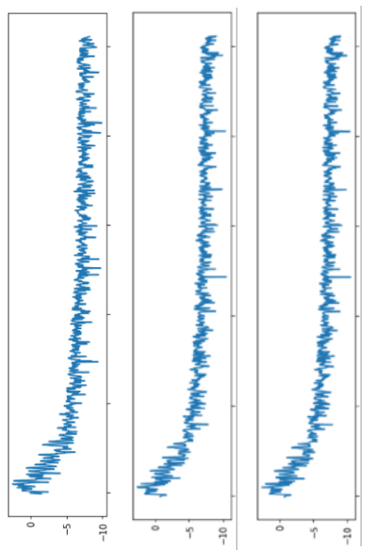
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것

Frame



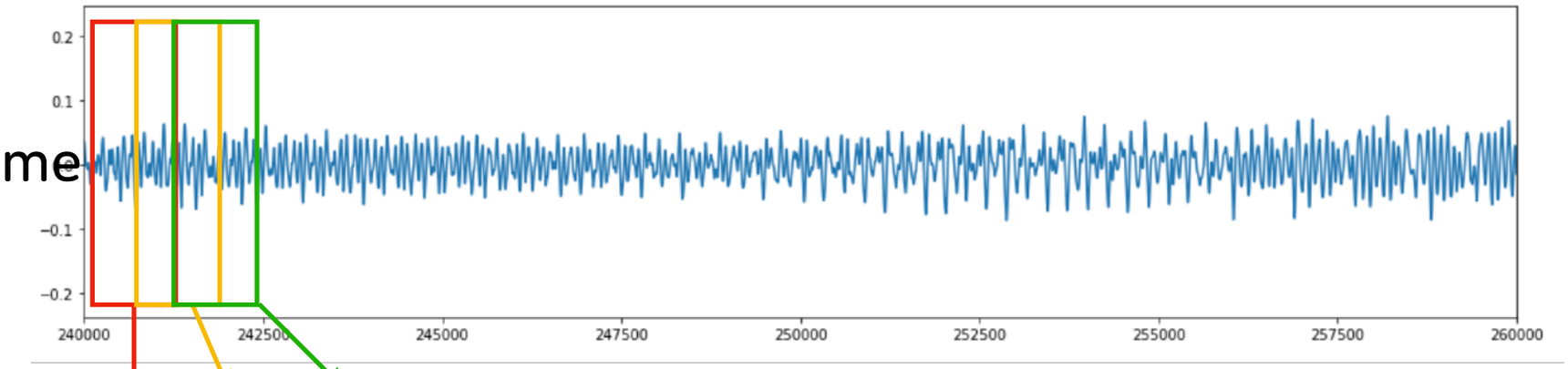
Spectrum



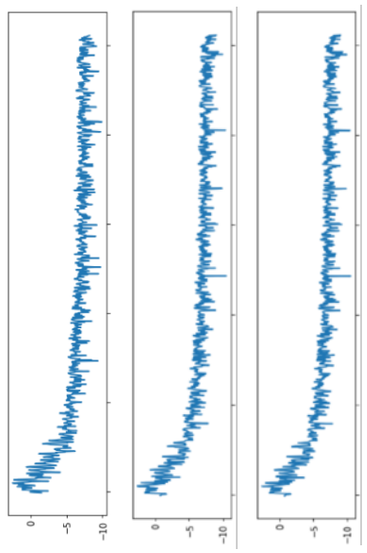
Spectrogram?

웨이브를 푸리에 변환을 통해 주파수 영역에서 표현한 것

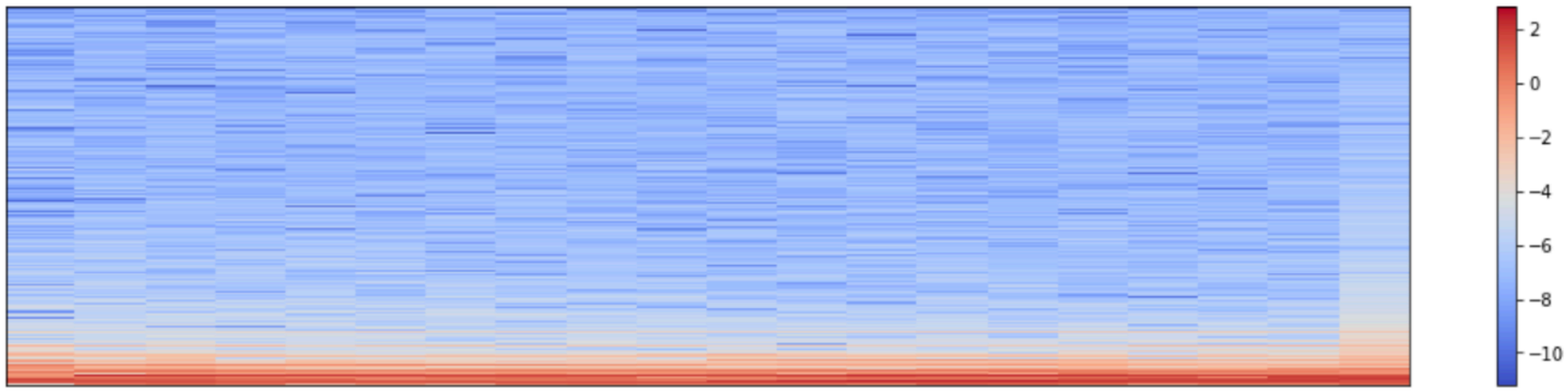
Frame



Spectrum



Spectrogram



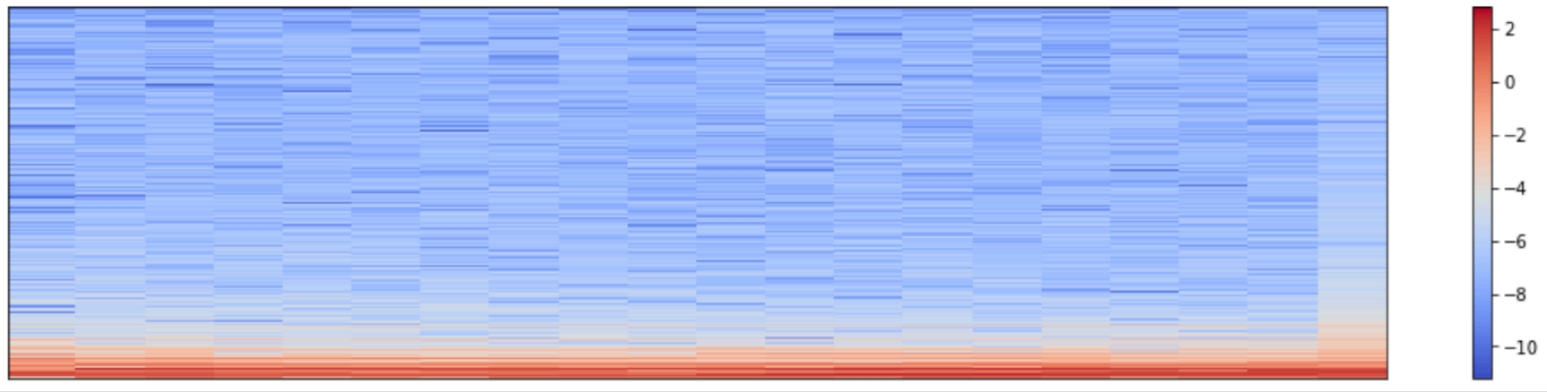
Frequency

Time

Mel-Spectrogram?

유용한 정보가 낮은 주파수대에 분포해 있는 특성을 반영해 Spectrogram을 선형 변환한 것

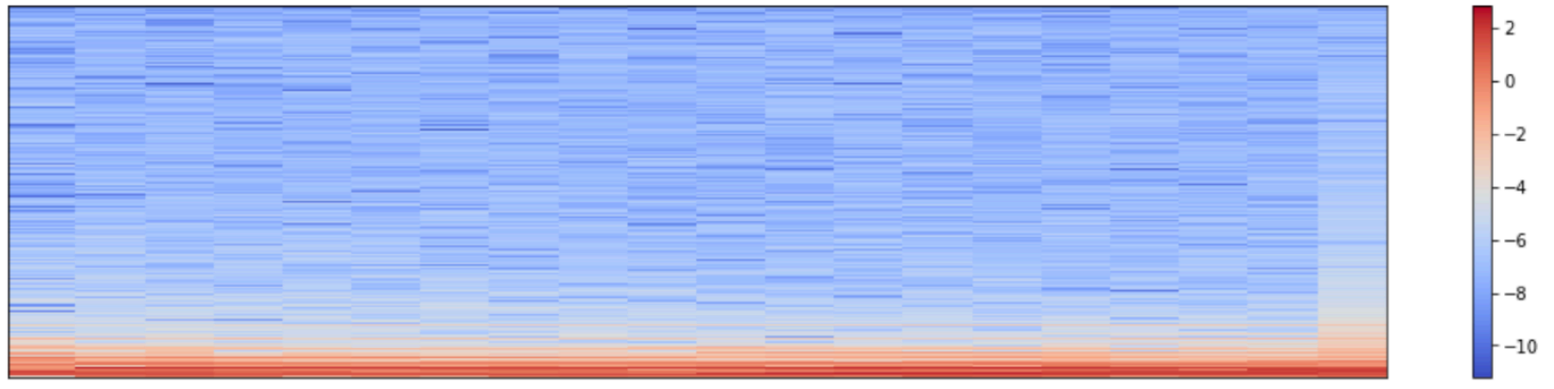
Spectrogram



Mel-Spectrogram?

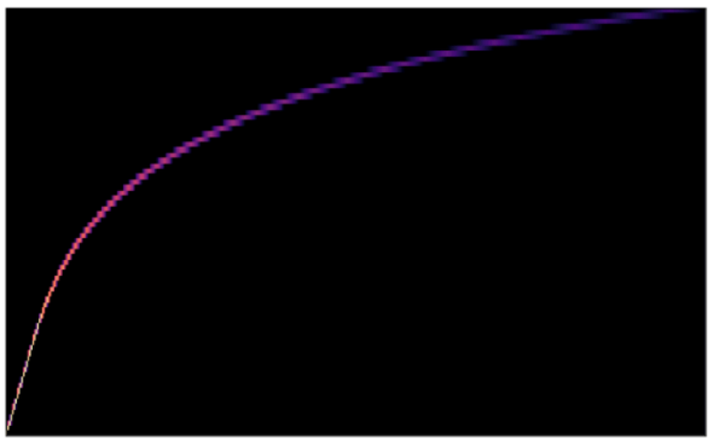
유용한 정보가 낮은 주파수대에 분포해 있는 특성을 반영해 Spectrogram을 선형 변환한 것

Spectrogram



×

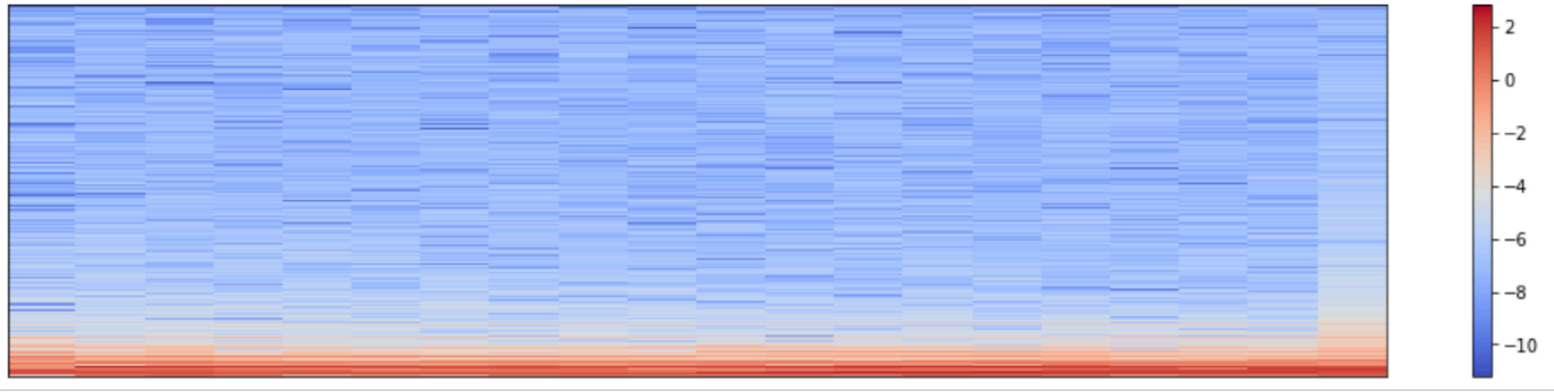
Mel-Filter Matrix



Mel-Spectrogram?

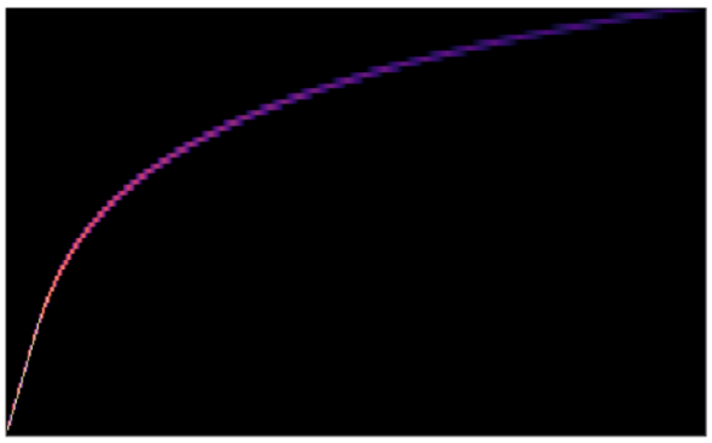
유용한 정보가 낮은 주파수대에 분포해 있는 특성을 반영해 Spectrogram을 선형 변환한 것

Spectrogram



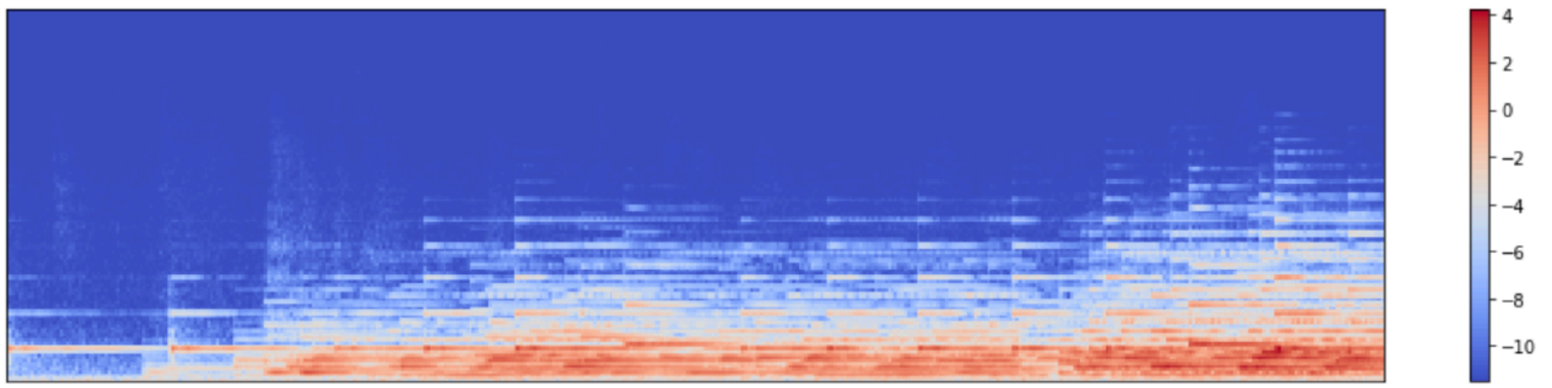
×

Mel-Filter Matrix



=

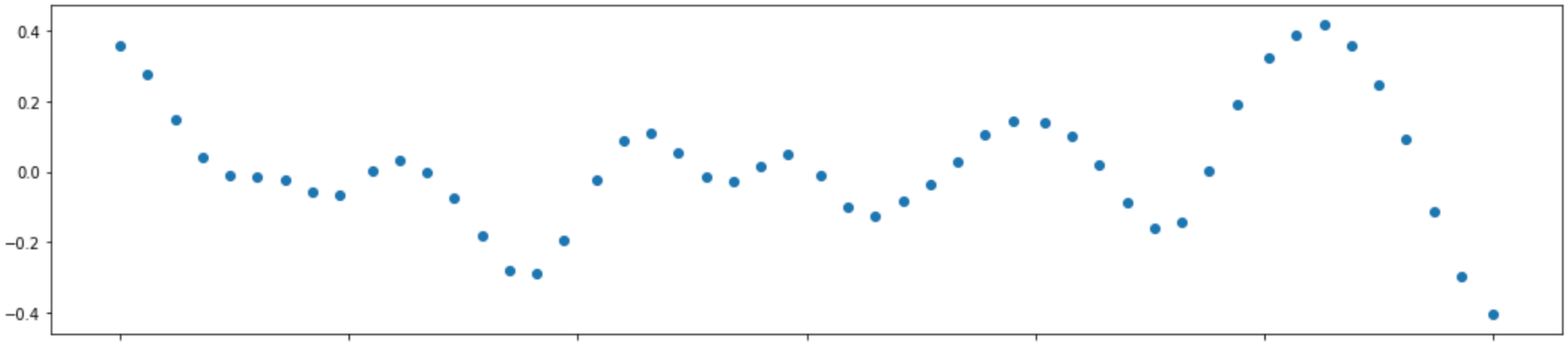
Mel-Spetrogram



Autoregressive Models

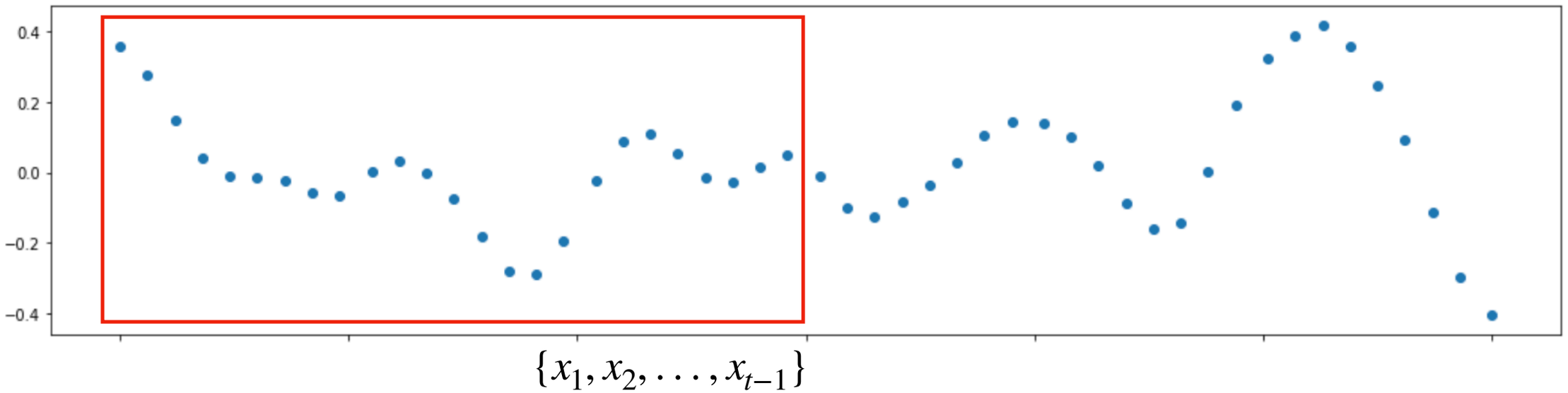
2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



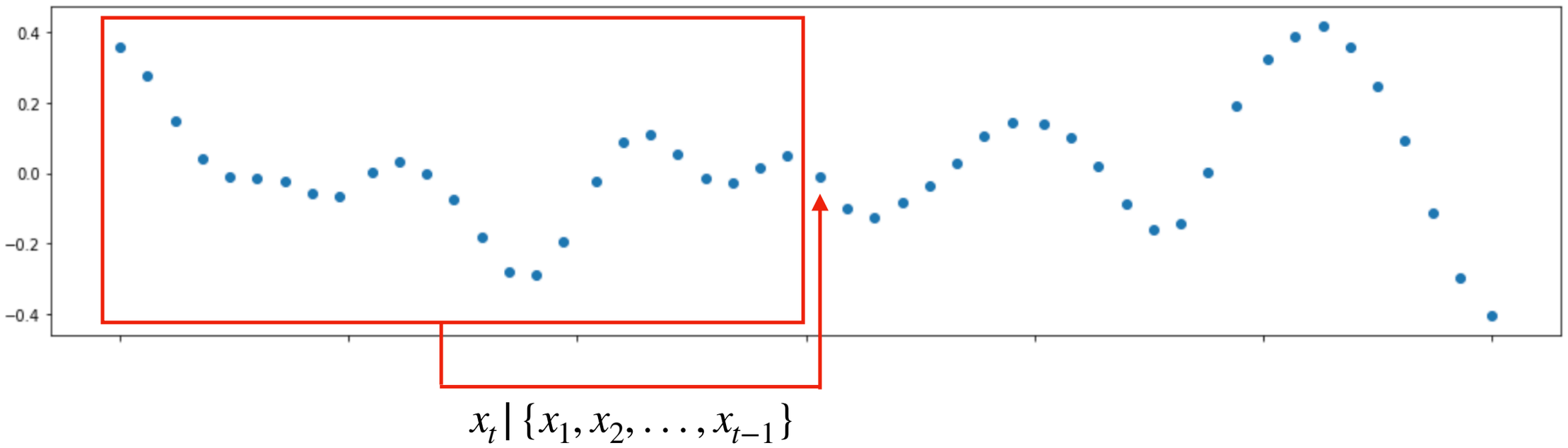
2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



2. Autoregressive Models

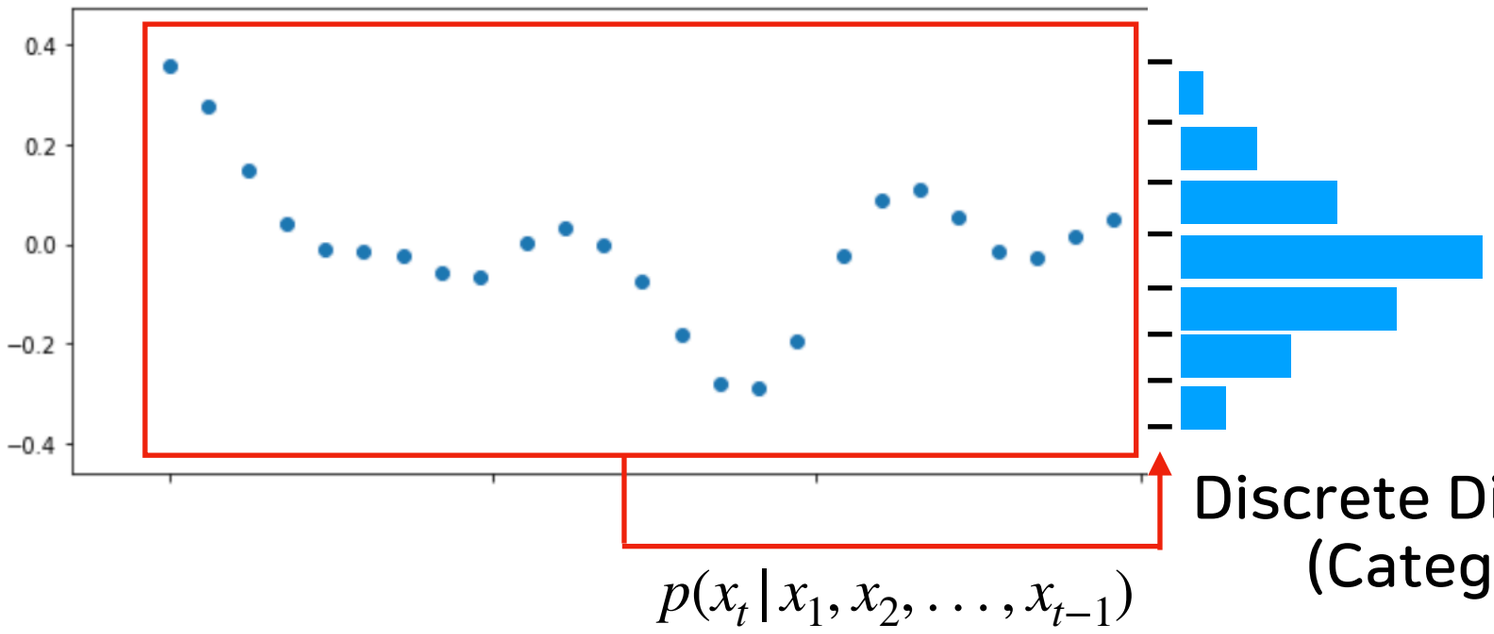
한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



$$x_t \mid \{x_1, x_2, \ldots, x_{t-1}\}$$

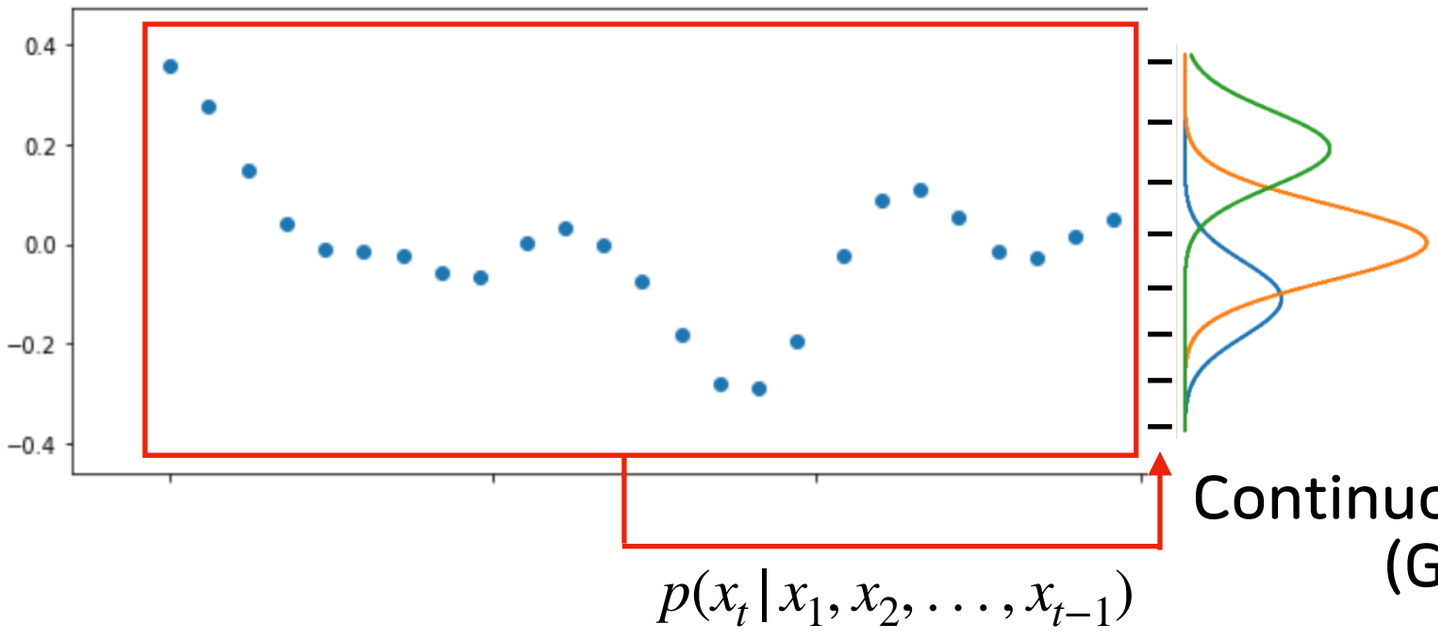
2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측

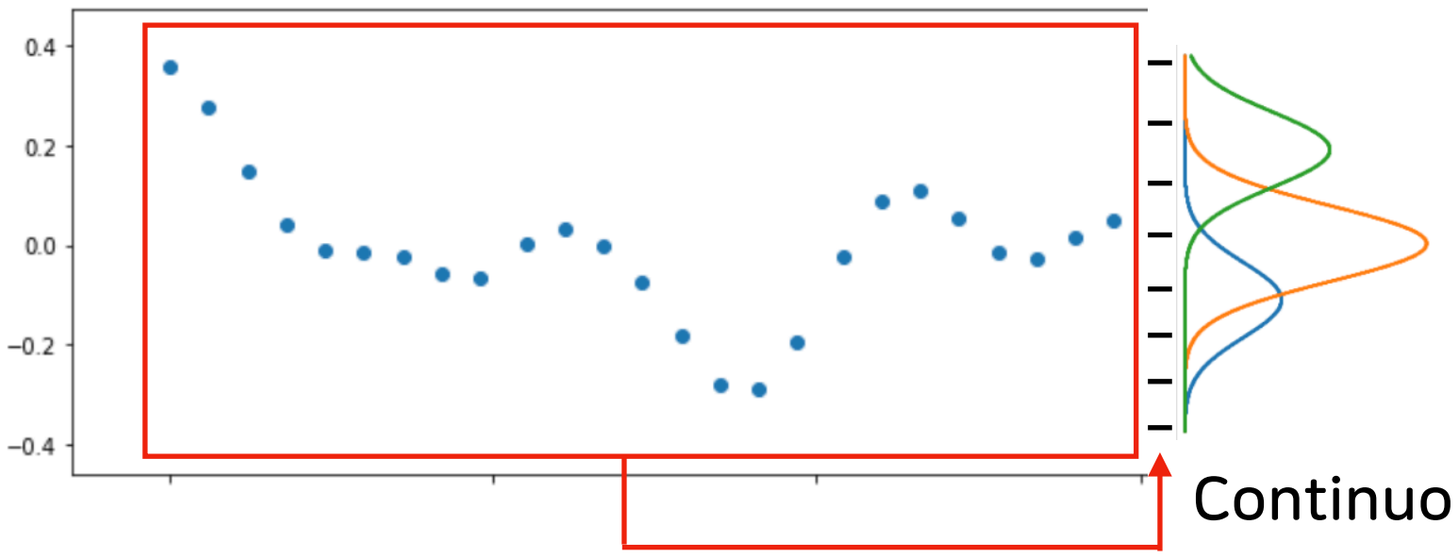


$p(x_t \mid x_1, x_2, \ldots, x_{t-1})$

Continuous Distribution
(Gaussian)

2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



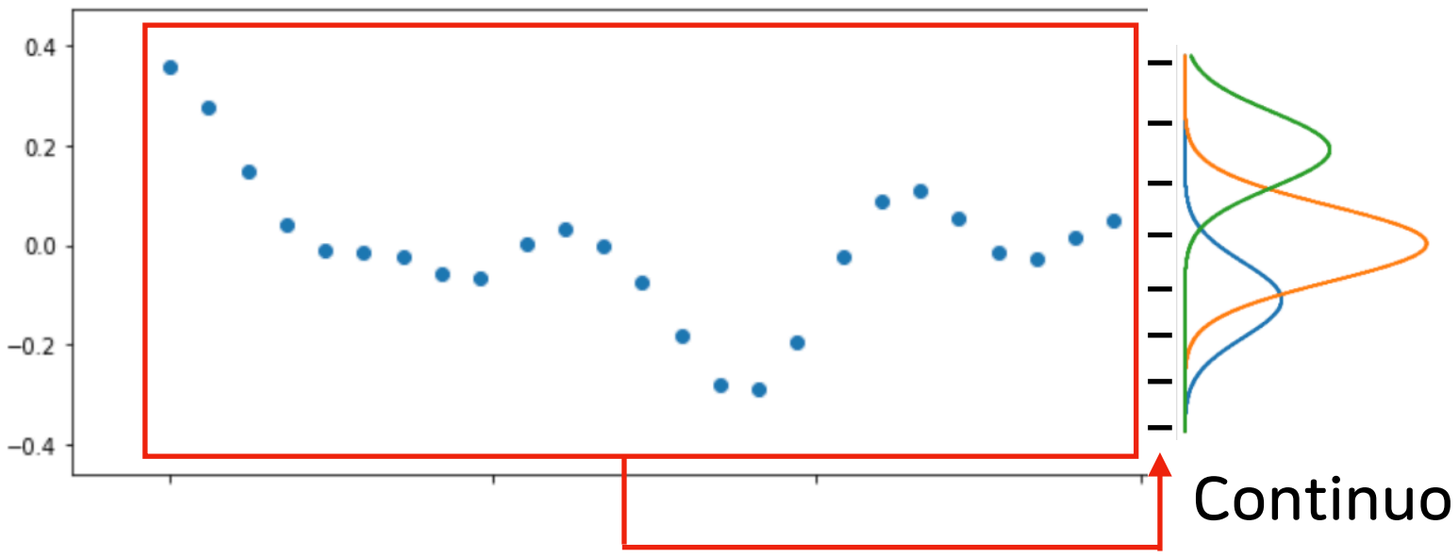
Continuous Distribution

$$p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

$$p(X) = p(x_1, x_2, \ldots, x_T)$$

2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



Continuous Distribution

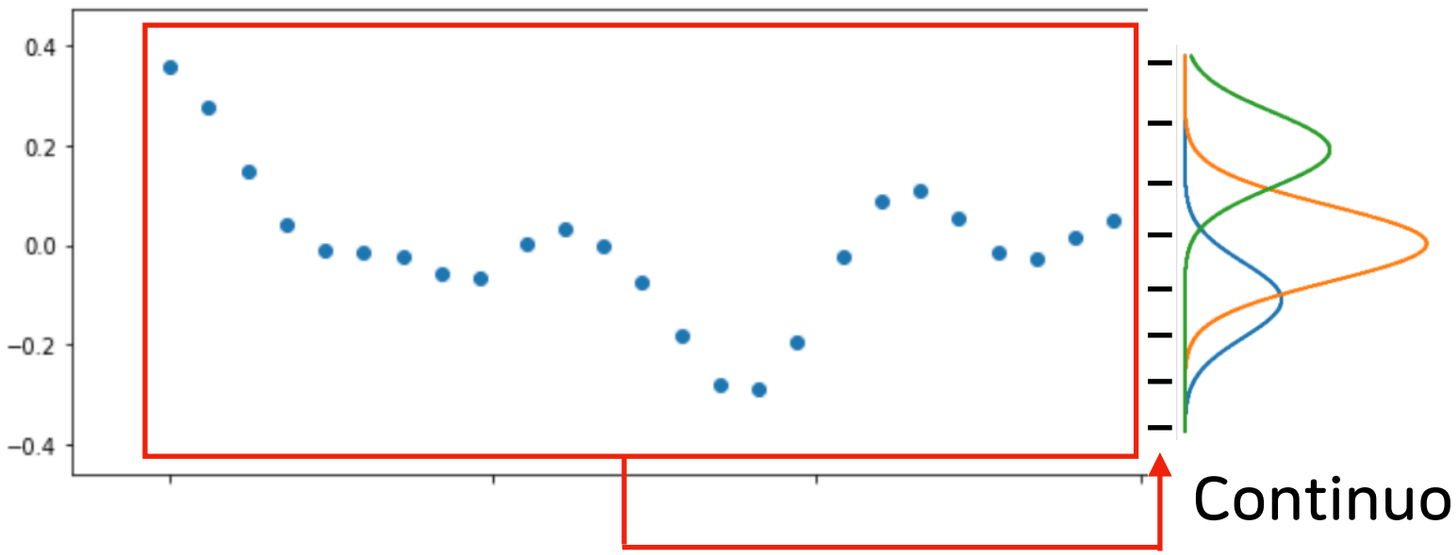
$$p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

Chain Rule

$$p(X) = p(x_1, x_2, \ldots, x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \ldots p(x_T \mid x_1, \ldots, x_{T-1})$$

2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



Continuous Distribution

$$p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

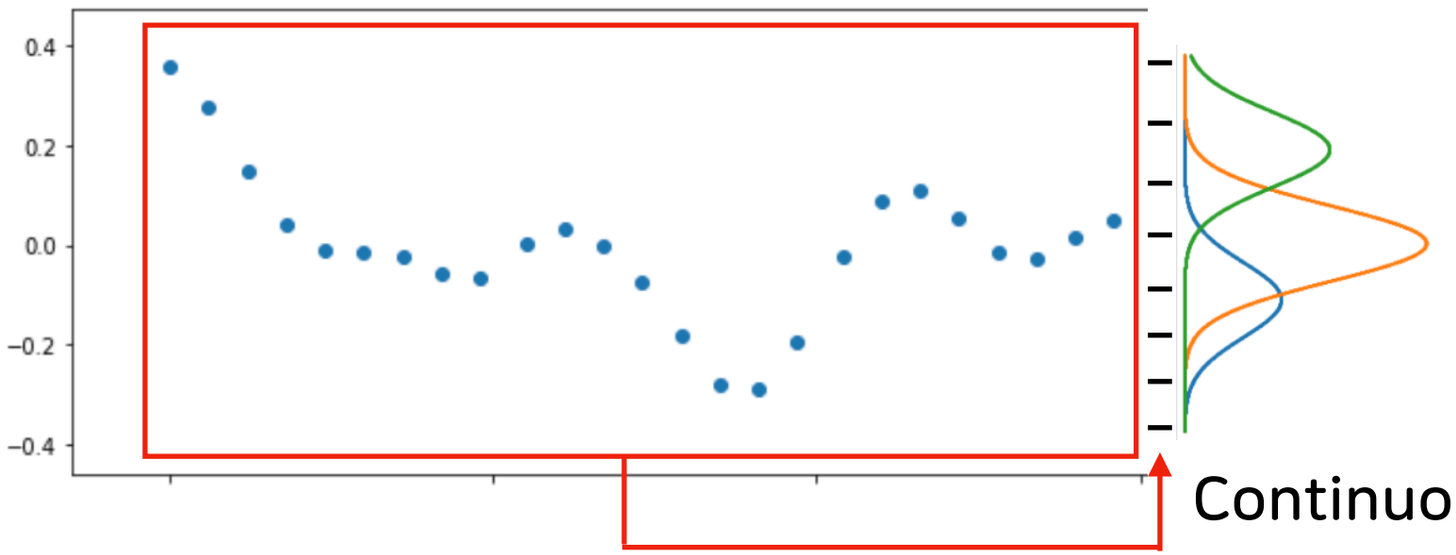
Chain Rule

$$p(X) = p(x_1, x_2, \ldots, x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \ldots p(x_T \mid x_1, \ldots, x_{T-1})$$

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

2. Autoregressive Models

한단계 앞까지의 과거 데이터를 토대로 현재 시점의 데이터의 확률 분포를 예측



Continuous Distribution

$$p(x_t \mid x_1, x_2, \ldots, x_{t-1})$$

Chain Rule

$$p(X) = p(x_1, x_2, \ldots, x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \ldots p(x_T \mid x_1, \ldots, x_{T-1})$$

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

Wavenet

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

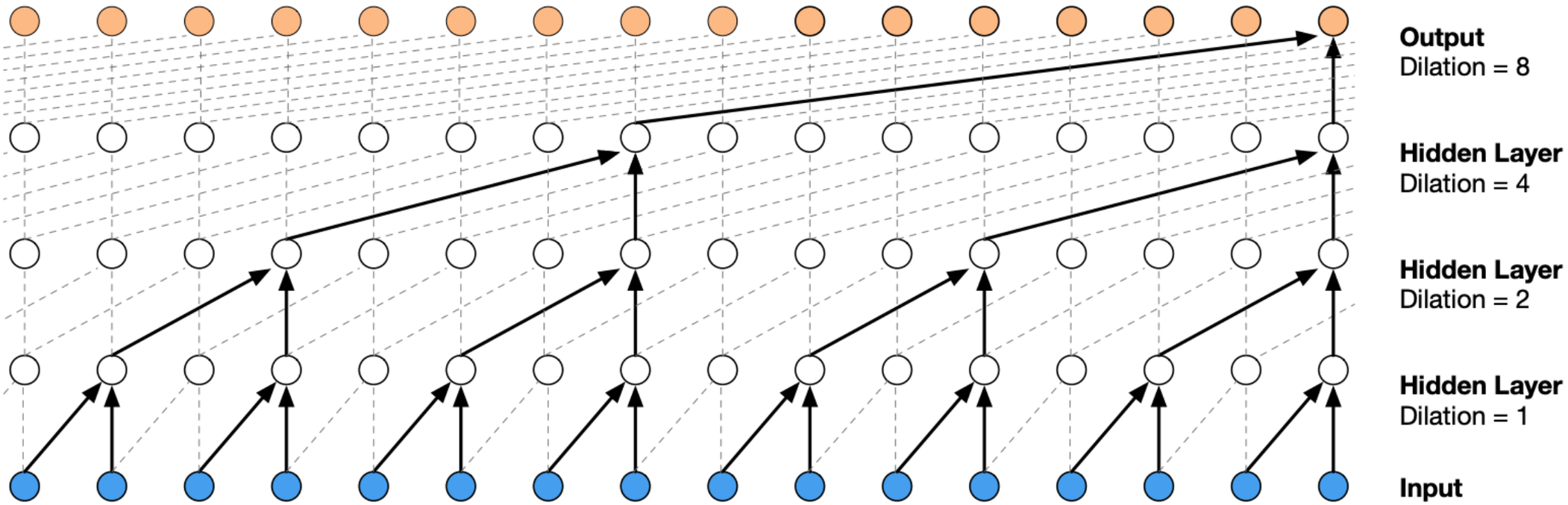
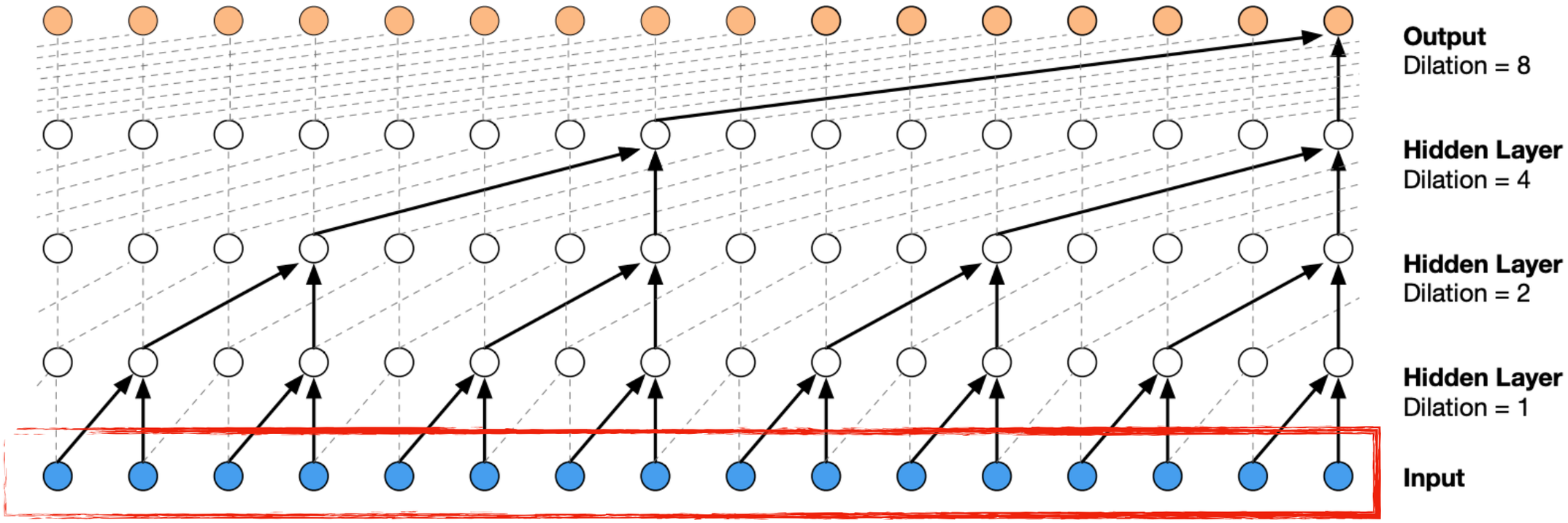


Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio



Wave Samples

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

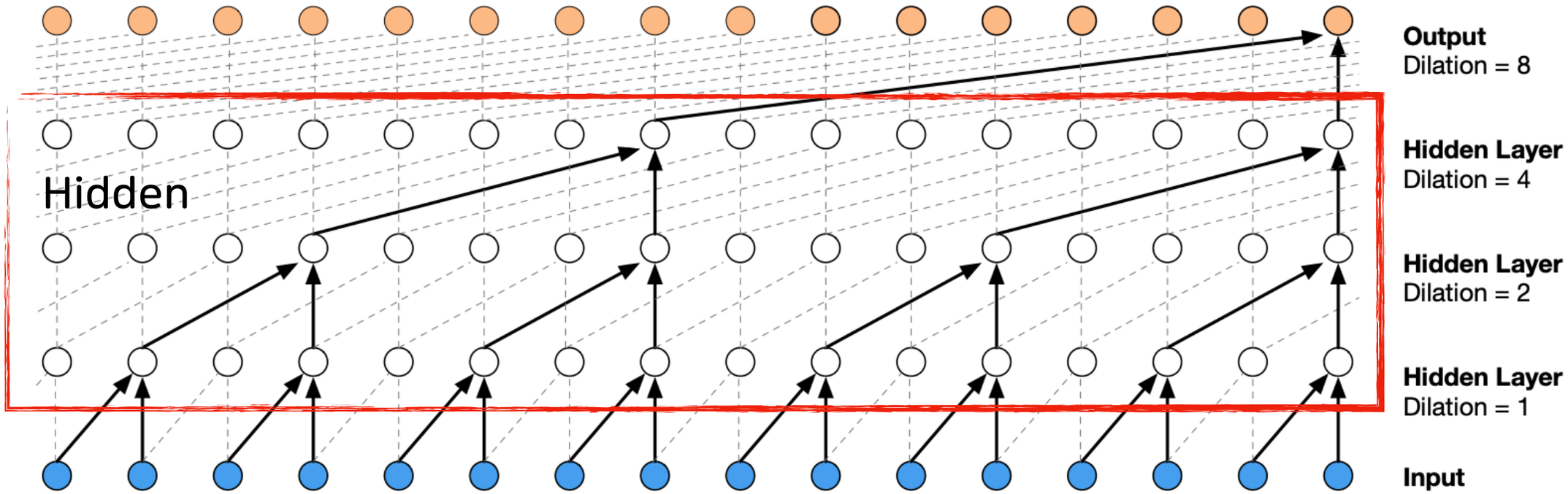


Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

Output = Parameters of Predictive Distribution

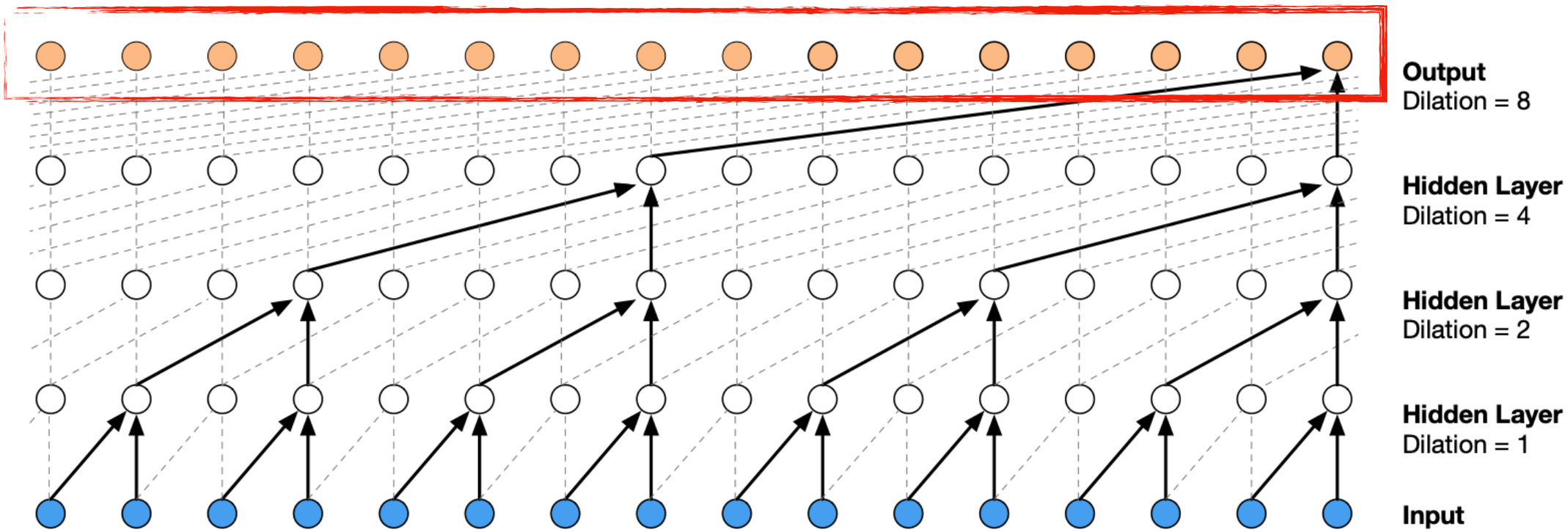
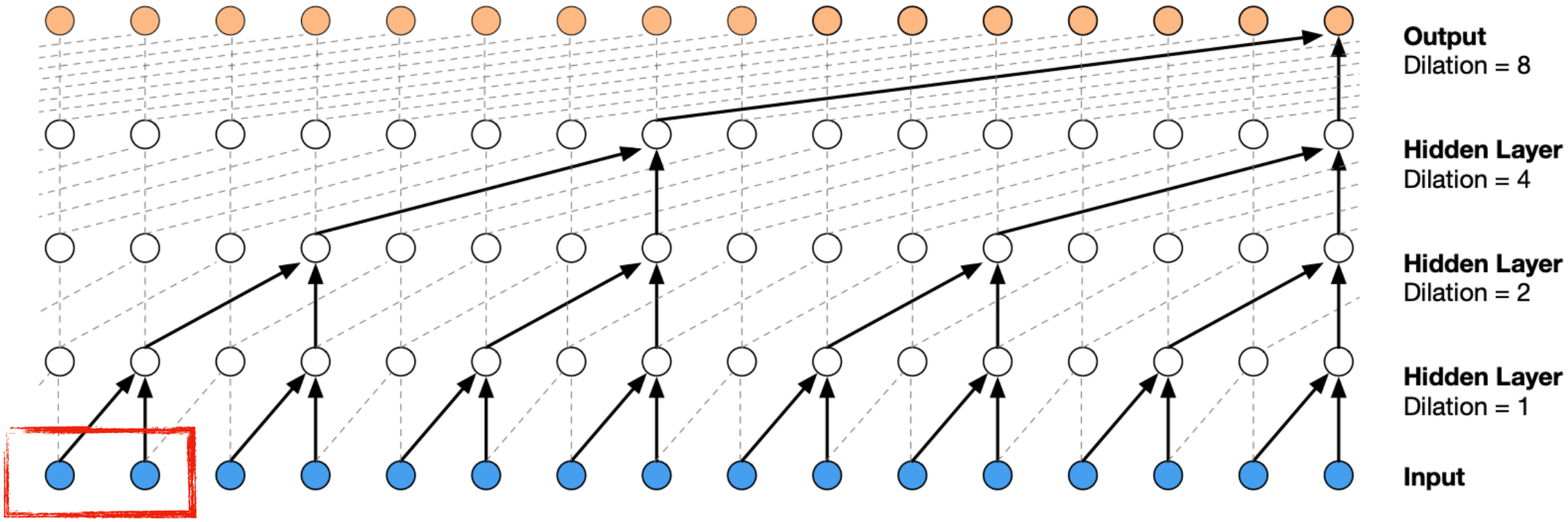


Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

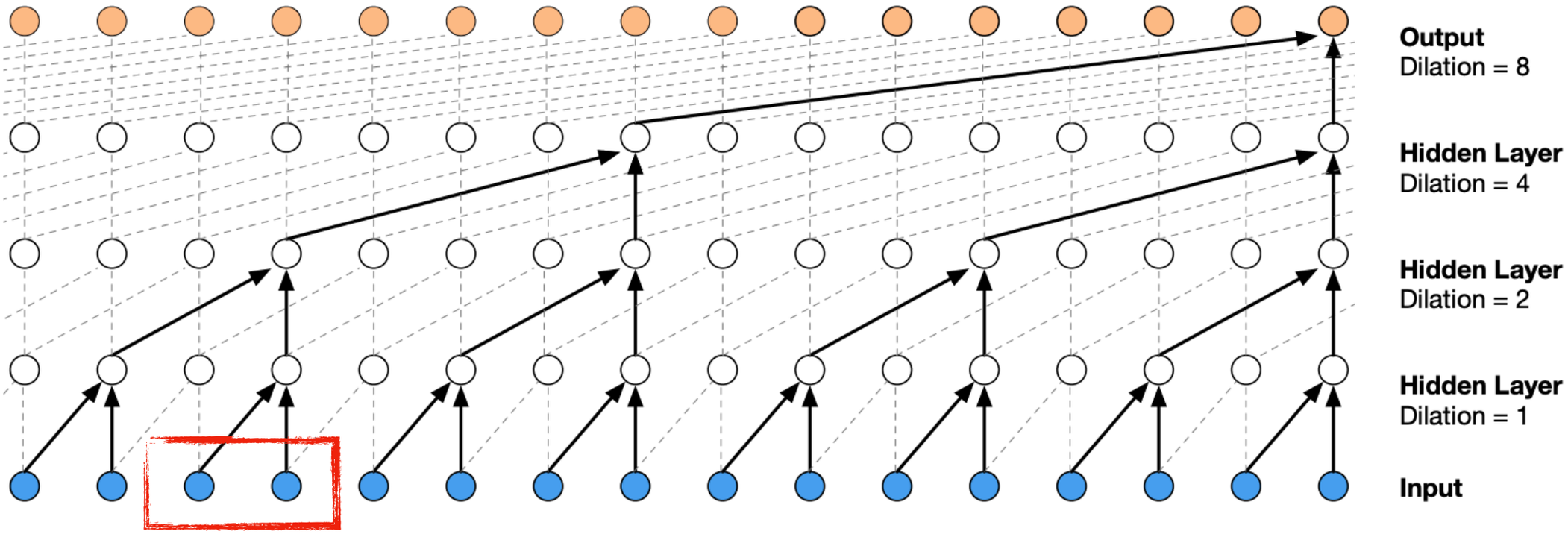


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

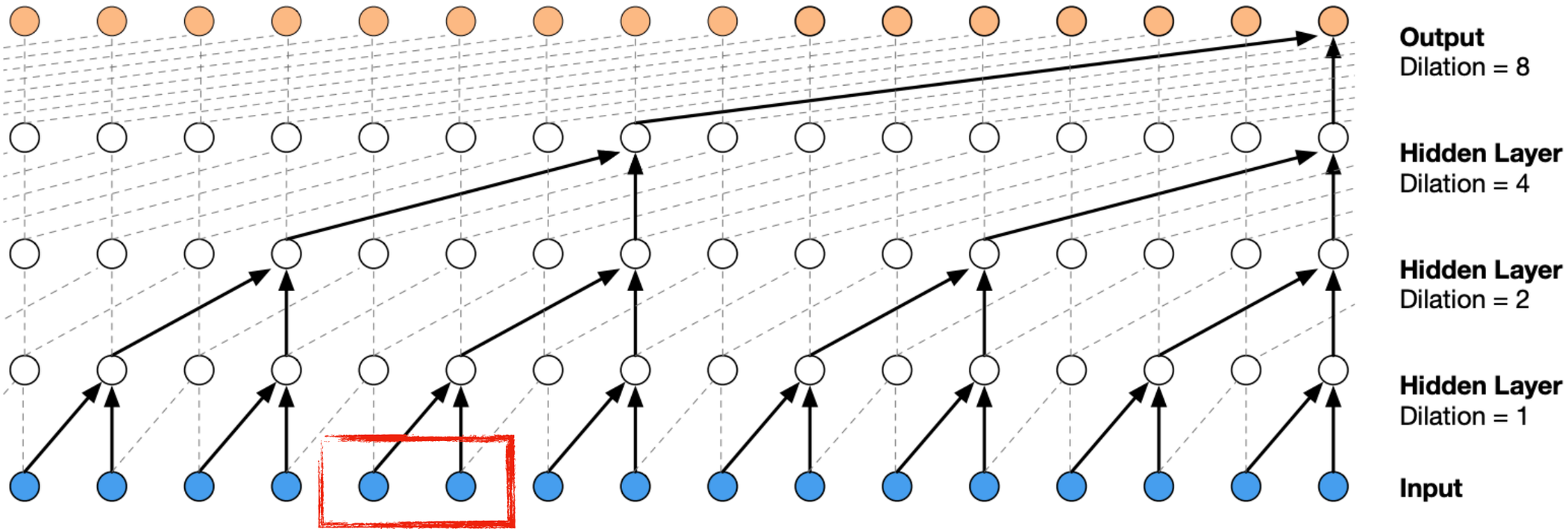


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

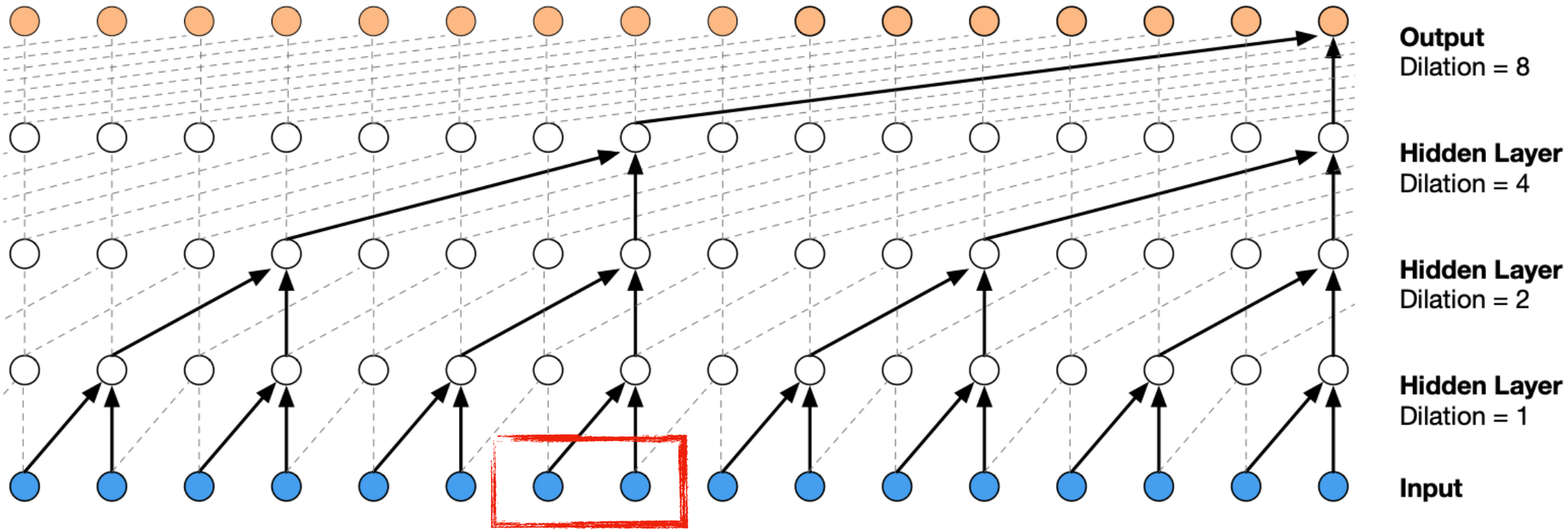


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

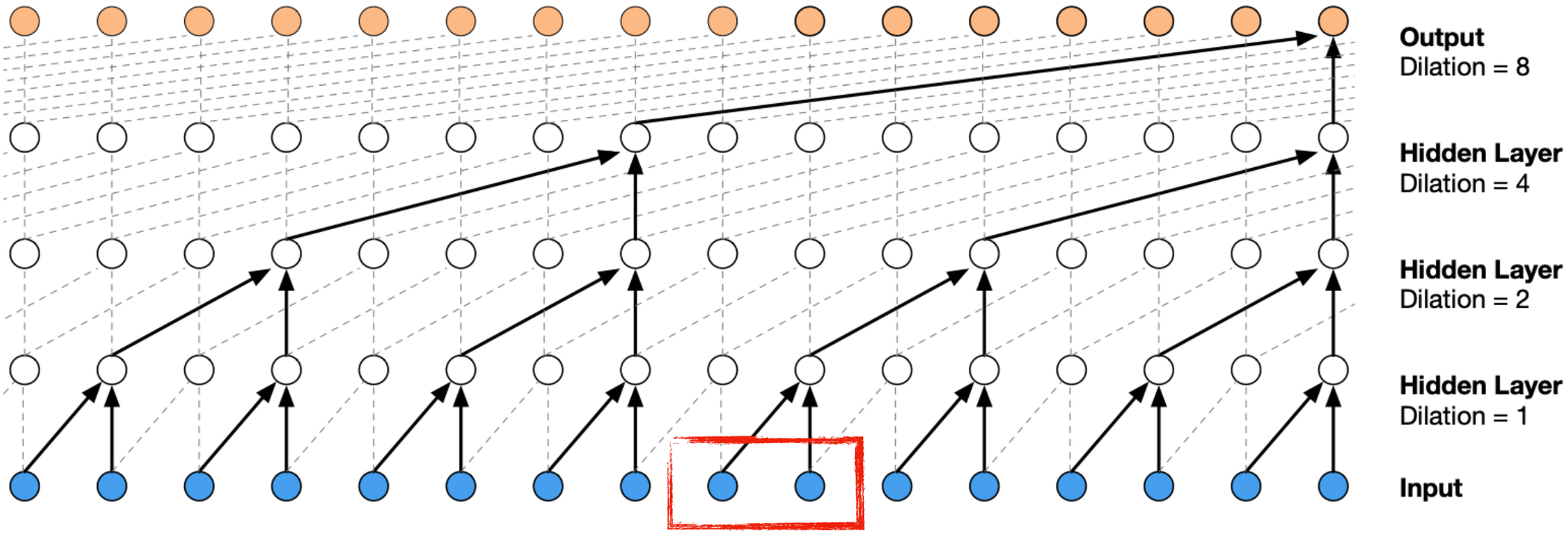


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

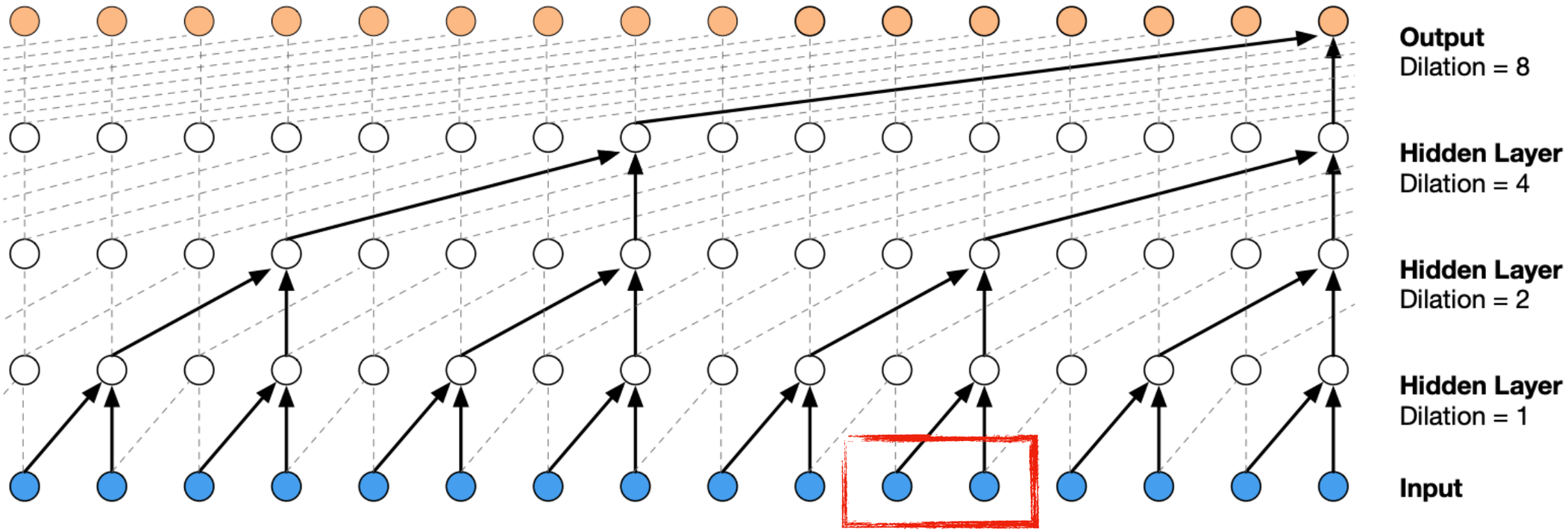


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

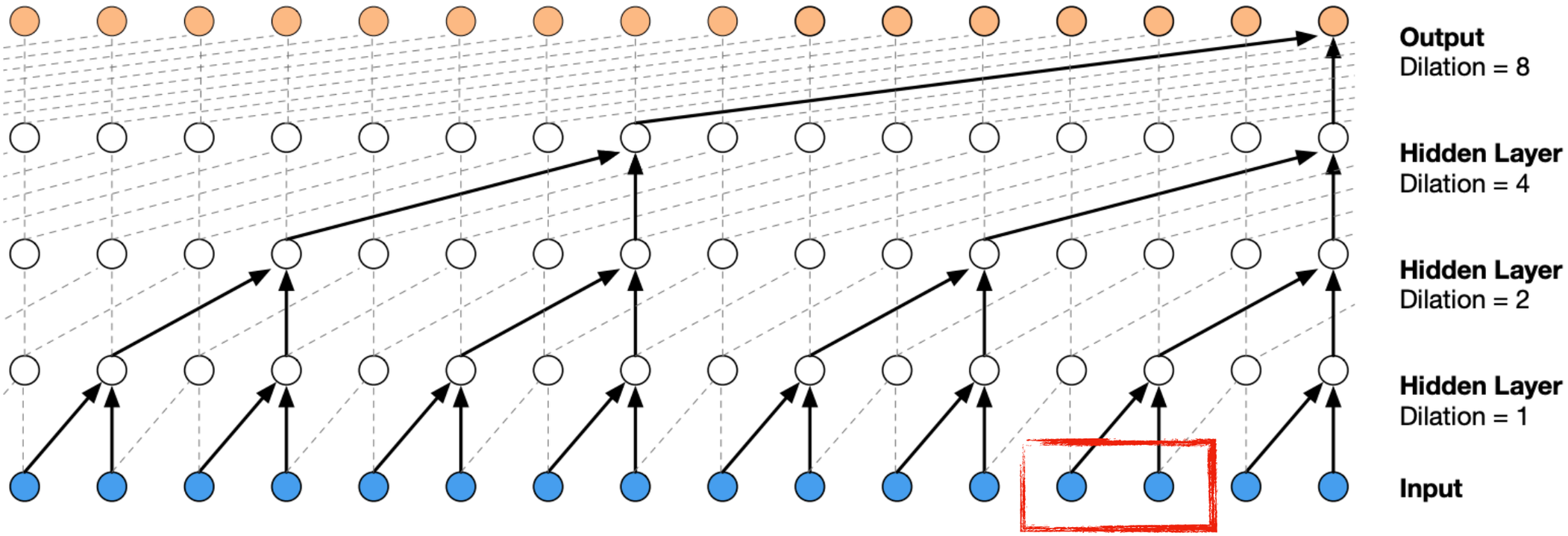


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

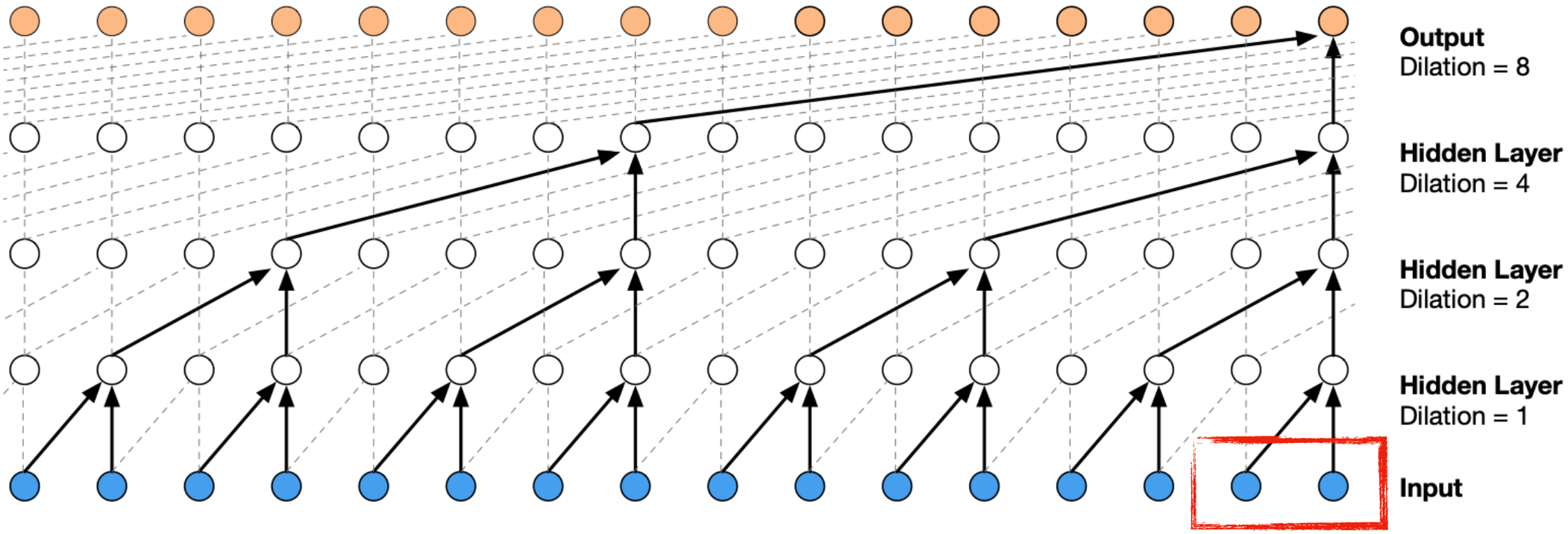


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

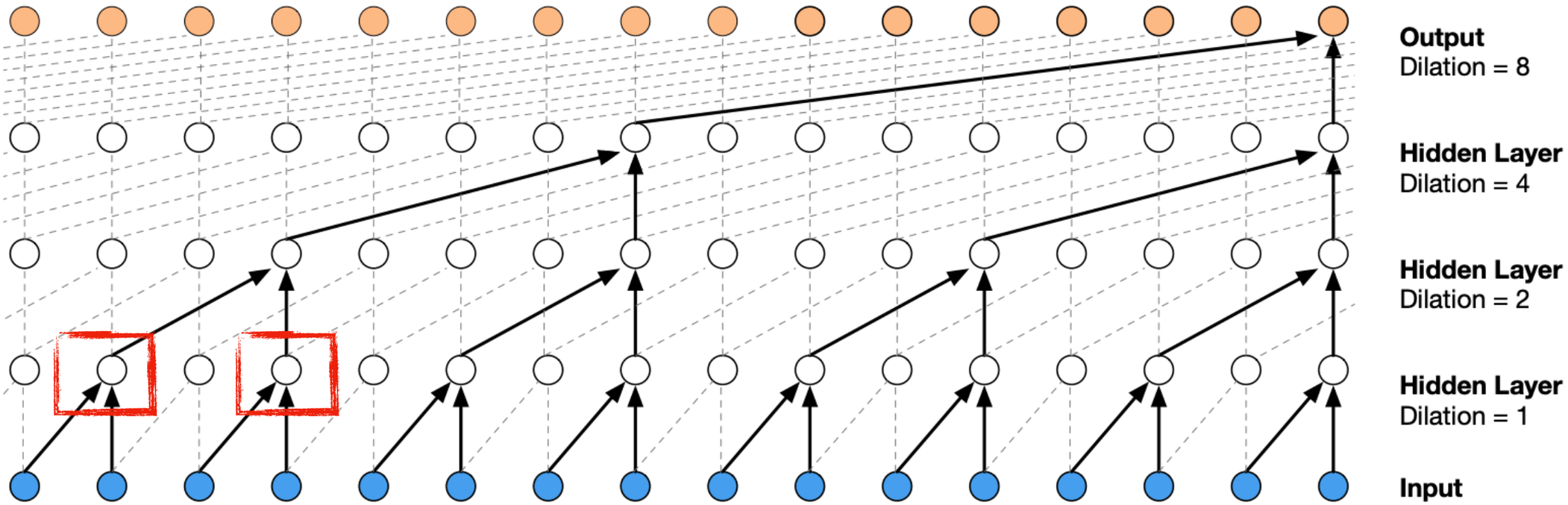


1D-Conv. Filter Length=2, Dilation=1

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

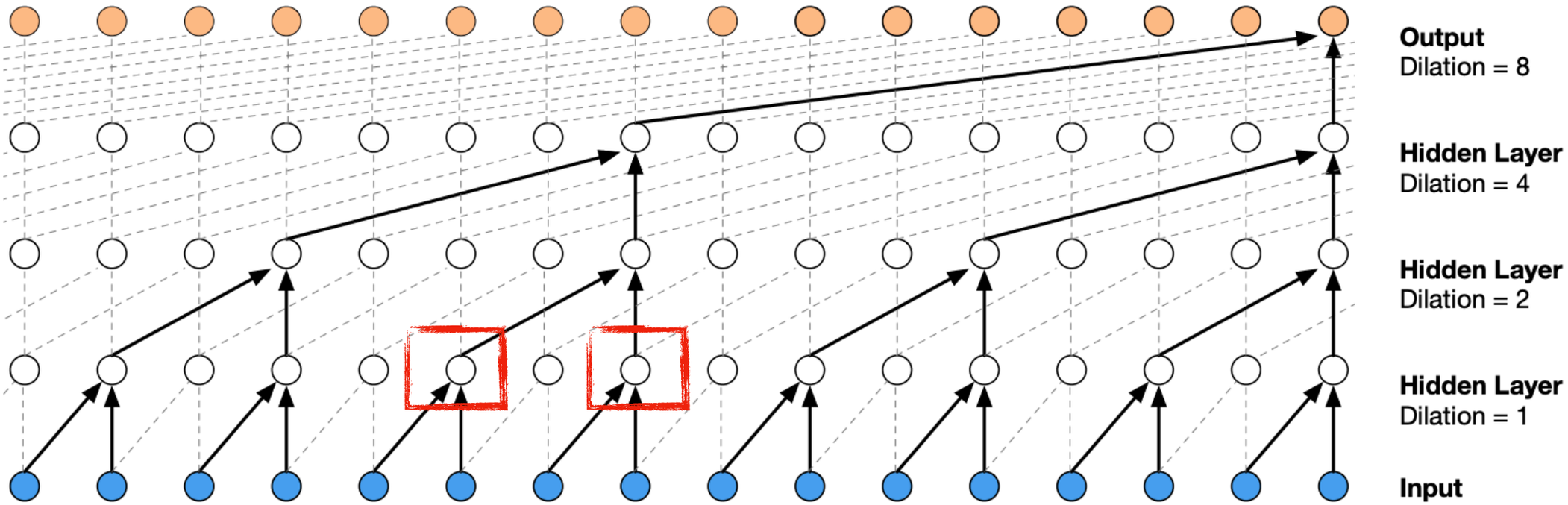


1D-Conv. Filter Length=2, Dilation=2

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

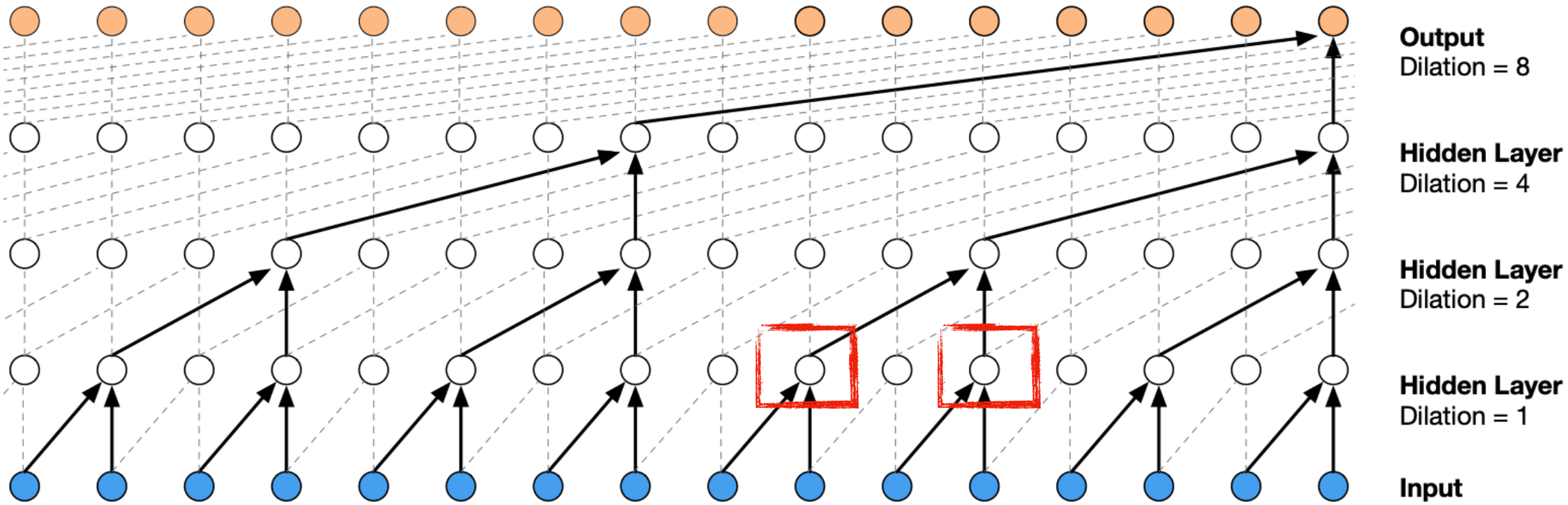


1D-Conv. Filter Length=2, Dilation=2

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

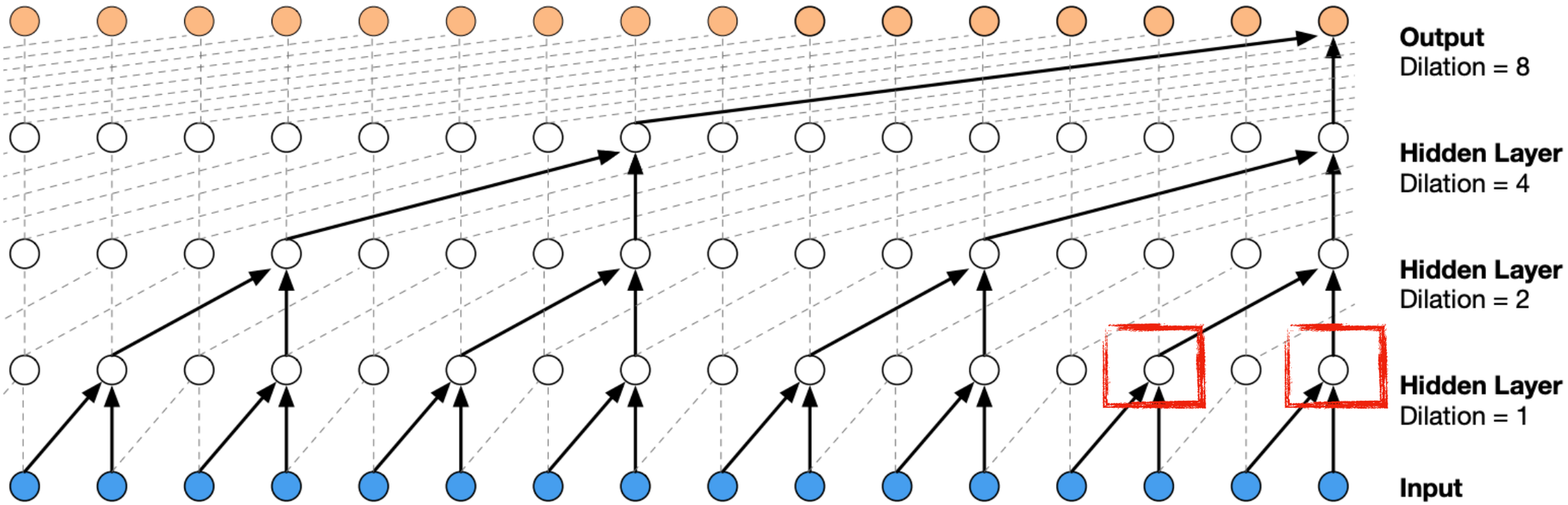


1D-Conv. Filter Length=2, Dilation=2

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

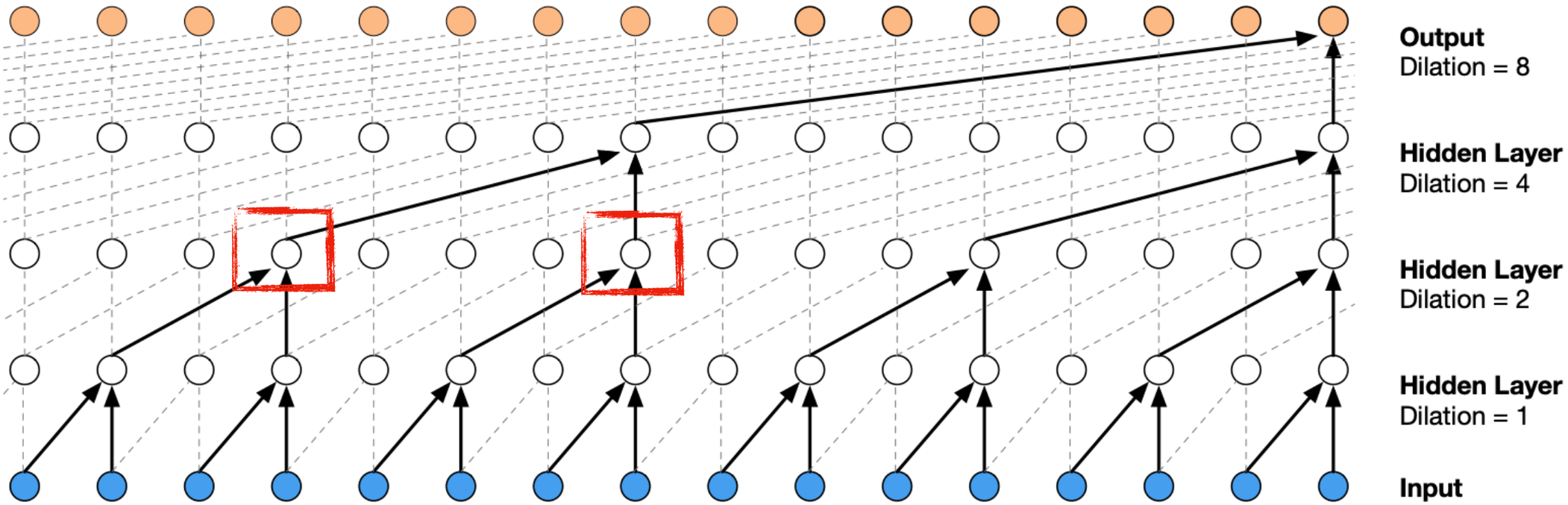


1D-Conv. Filter Length=2, Dilation=2

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

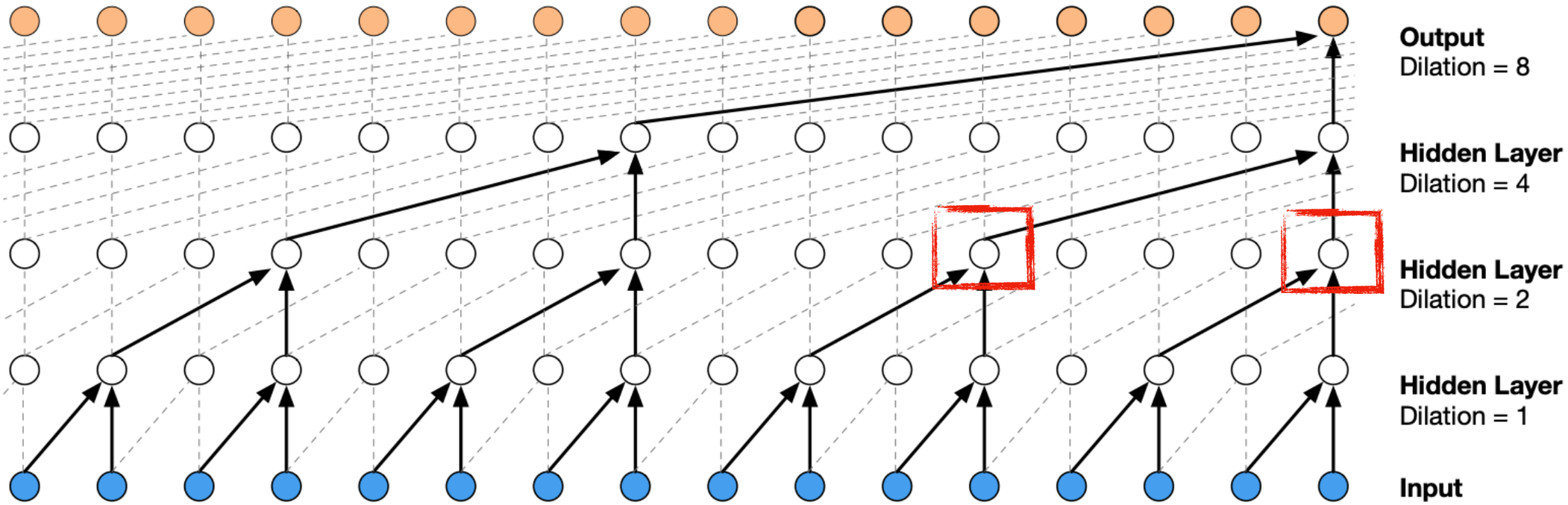


1D-Conv. Filter Length=2, Dilation=4

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

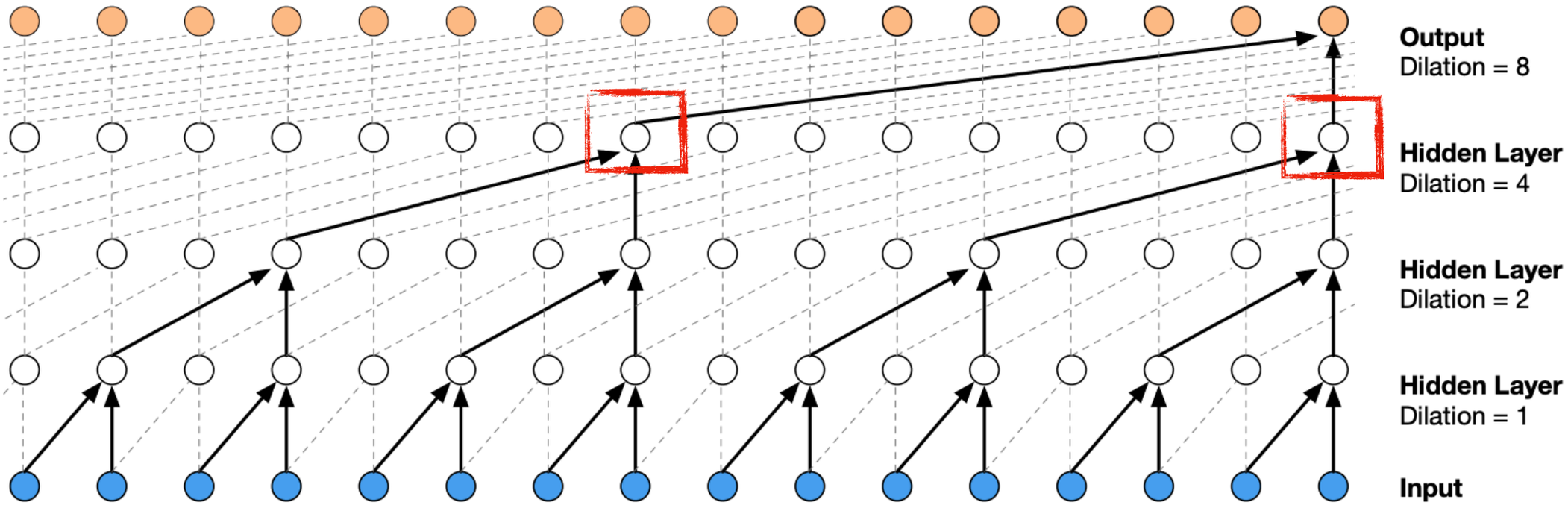


1D-Conv. Filter Length=2, Dilation=4

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

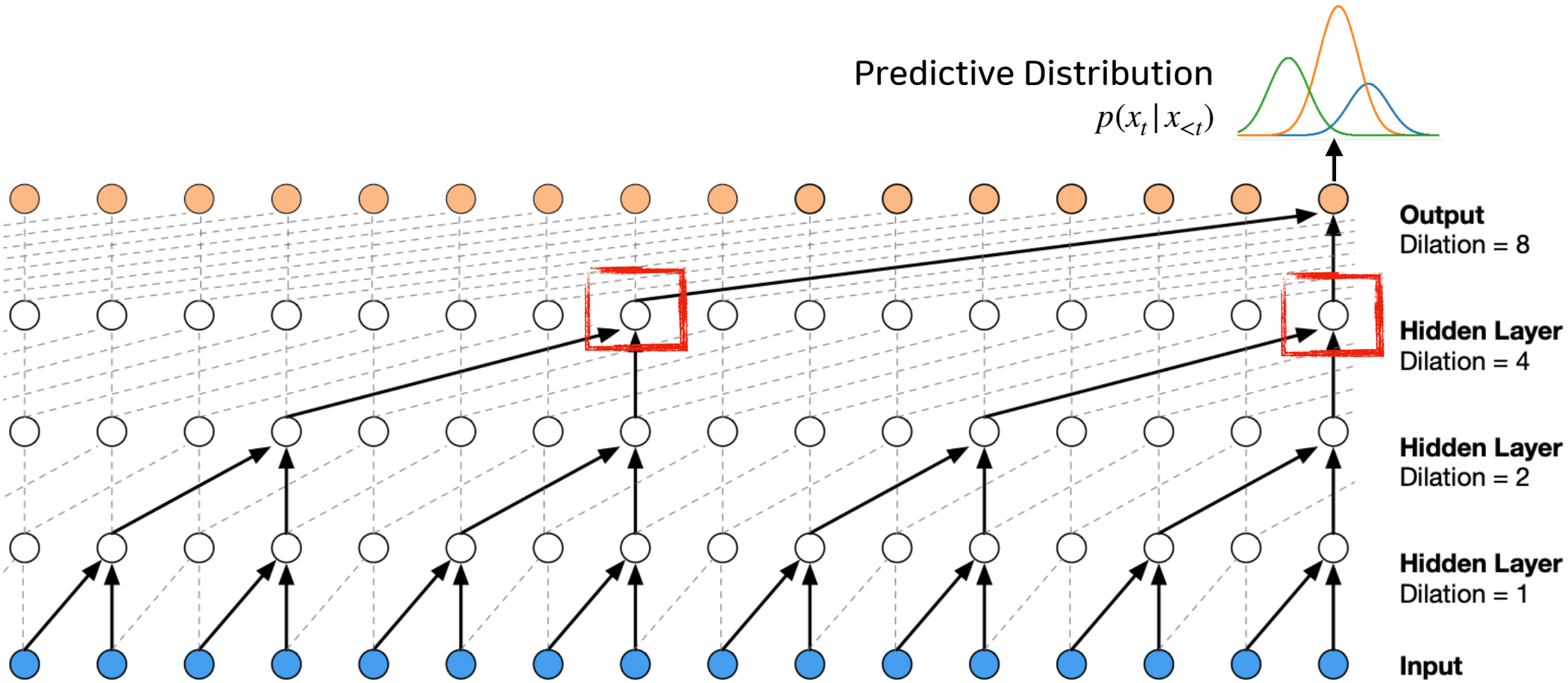


1D-Conv. Filter Length=2, Dilation=8

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

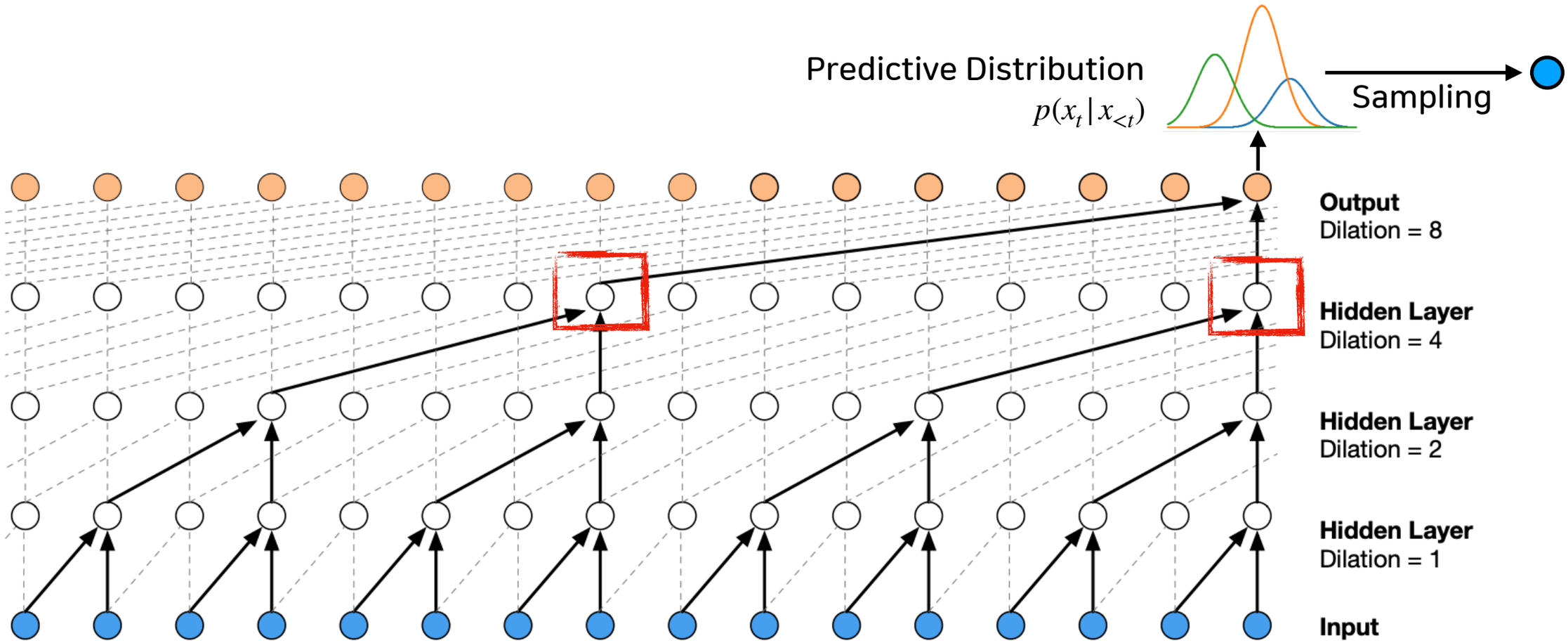


1D-Conv. Filter Length=2, Dilation=8

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio



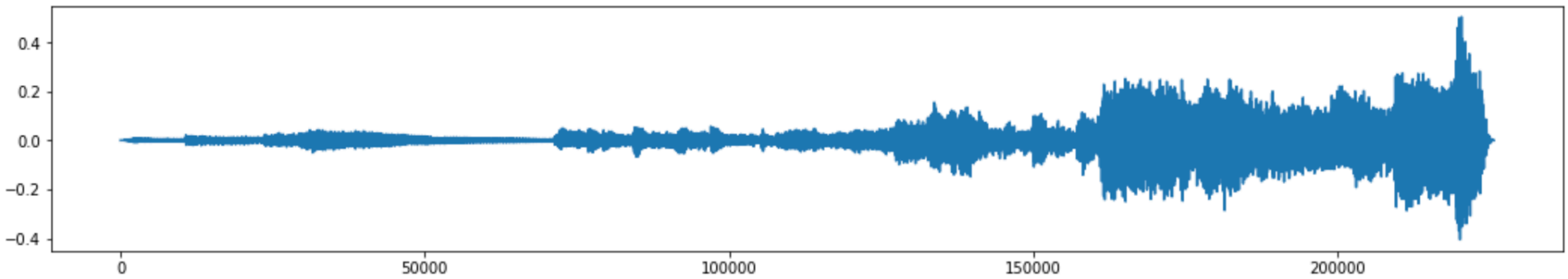
1D-Conv. Filter Length=2, Dilation=8

Figure credit: A Oord, WaveNet: A Generative Model for Raw Audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

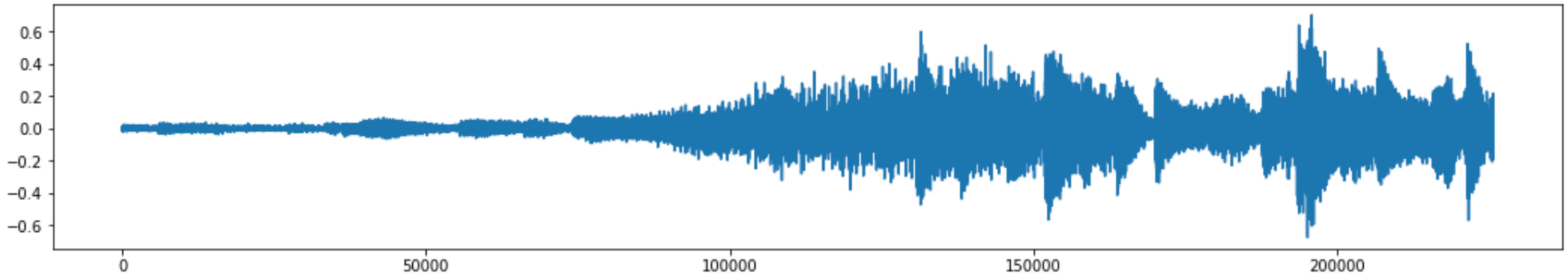
Generated Music. 1



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

2. Autoregressive Models
2.1 Wavenet : A Generative Model for Raw Audio

Generated Music. 2

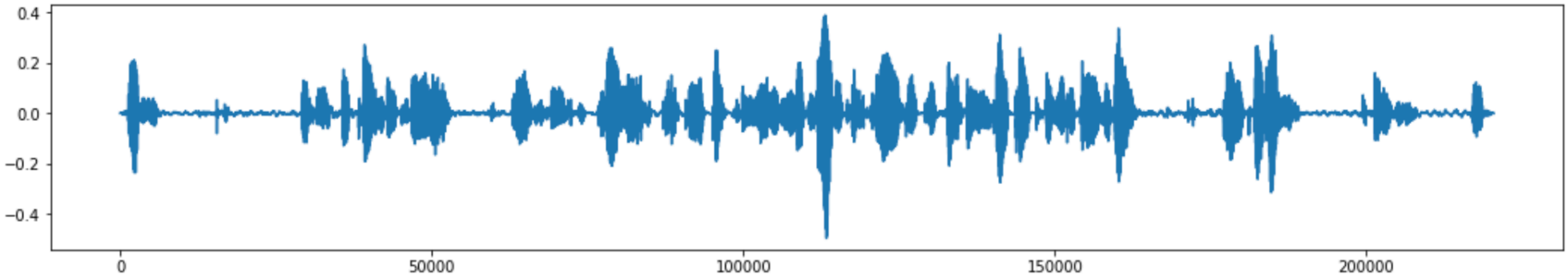


Audio Reference : https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

Generated Speech. 1

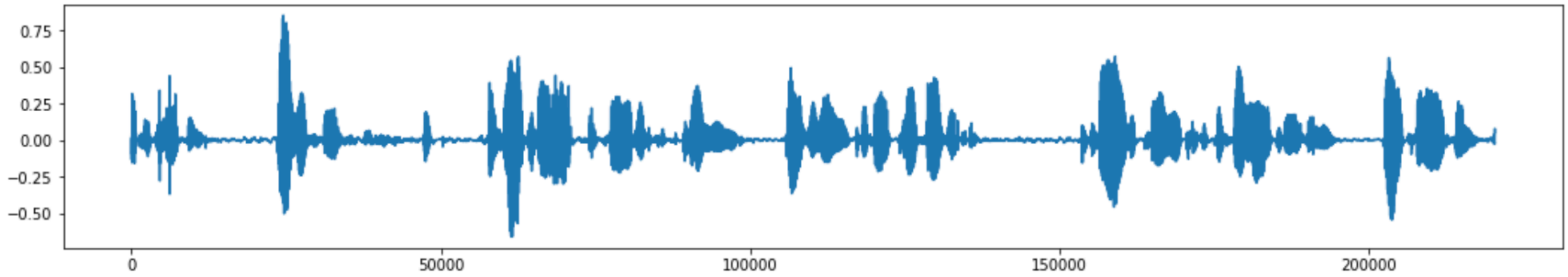


Audio Reference : https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

2. Autoregressive Models

2.1 Wavenet : A Generative Model for Raw Audio

Generated Speech. 2



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

2. Autoregressive Models
2.1 Wavenet : A Generative Model for Raw Audio

Long-Term Dependency

오디오 샘플을 하나하나씩 생성하므로 거시적인 의존관계를 반영하지 못한다.

2. Autoregressive Models
2.1 Wavenet : A Generative Model for Raw Audio

Long-Term Dependency

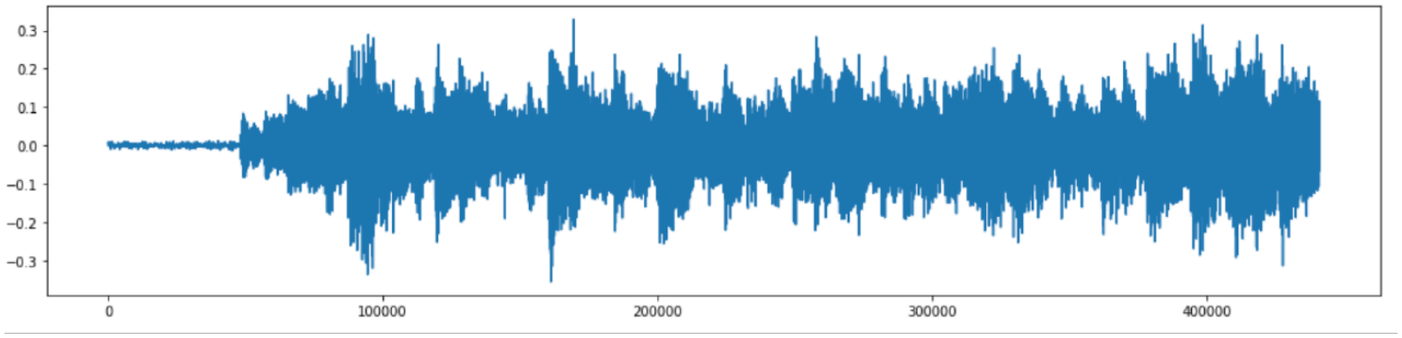
오디오 샘플을 하나하나씩 생성하므로 거시적인 의존관계를 반영하지 못한다.

샘플 단위가 아니라 프레임 단위의 생성 모델을 만든다.

VQ-VAE

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

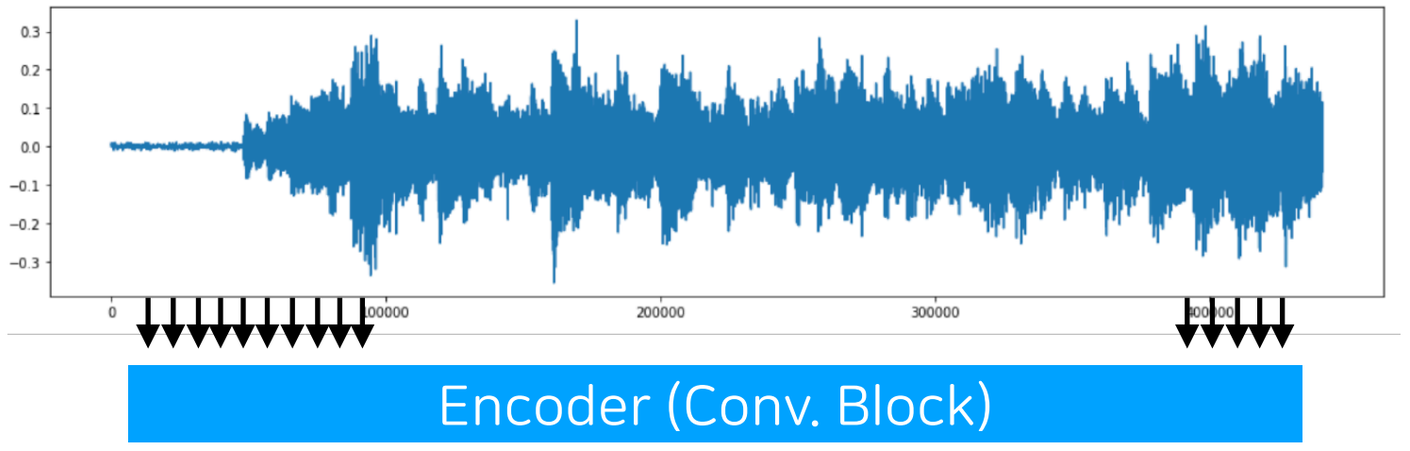


Shape

[Time, 1]

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

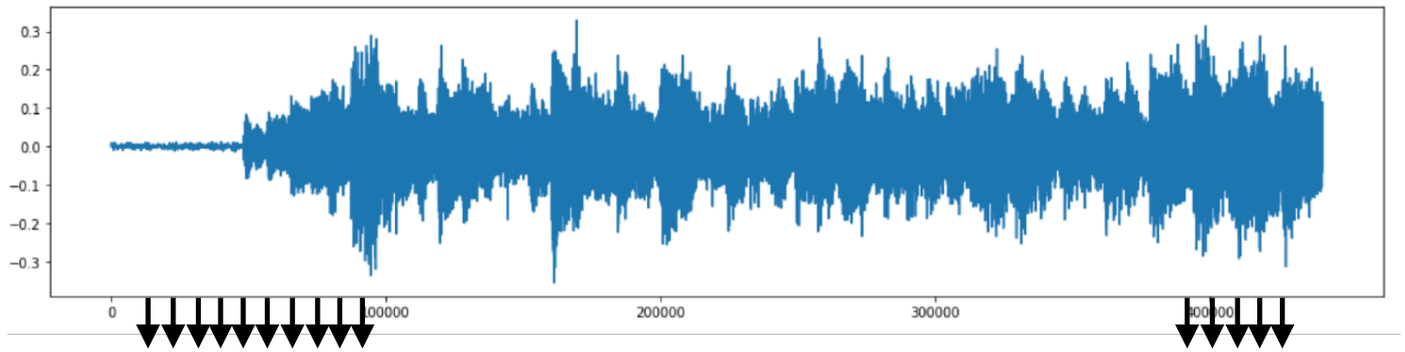


Shape

[Time, 1]

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

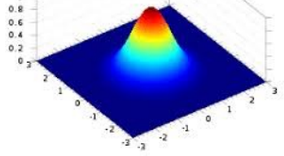


Shape

[Time, 1]

VAE 적용!

z_1 z_2



$$KL(q_\phi(z|x)\|p_\theta(z))$$

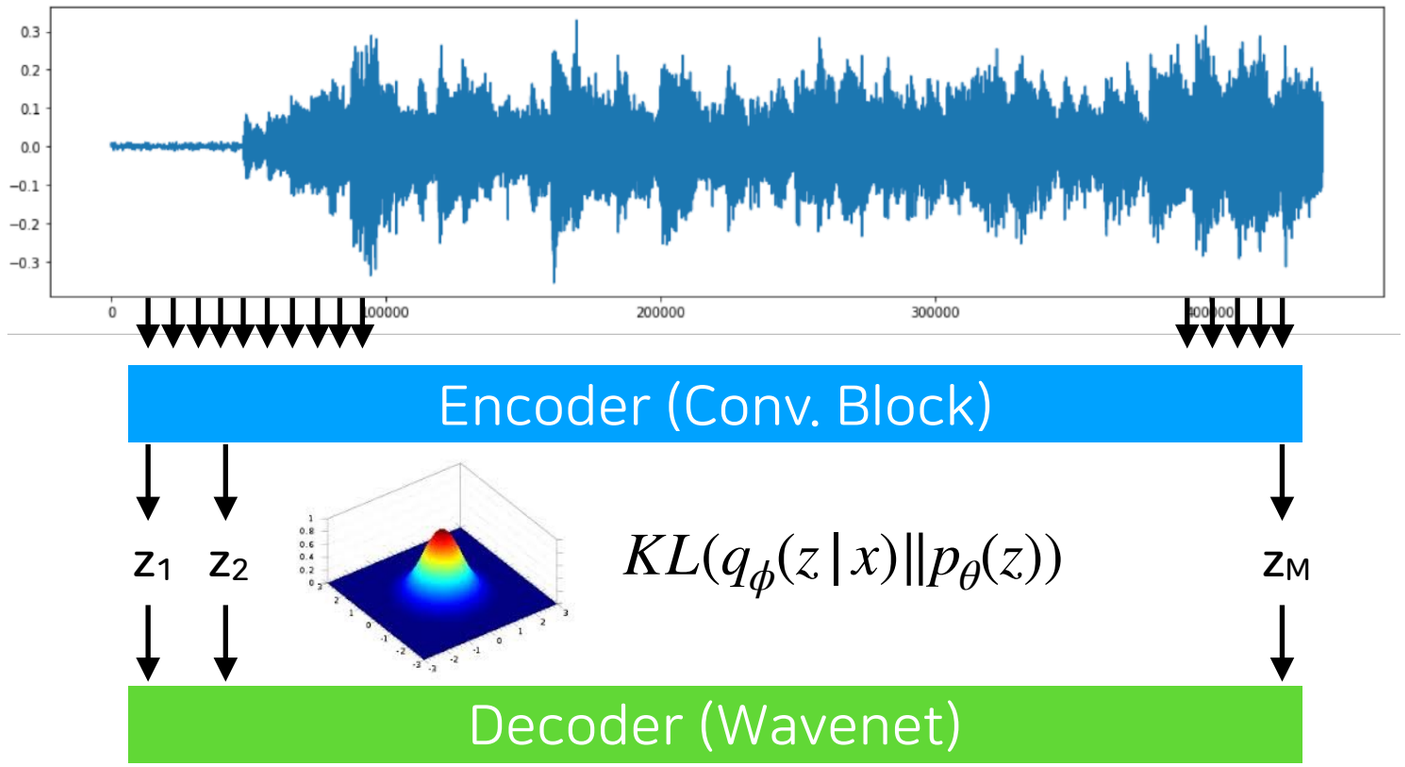
z_M

[Frame, |Z|]

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

VAE 적용!



Shape

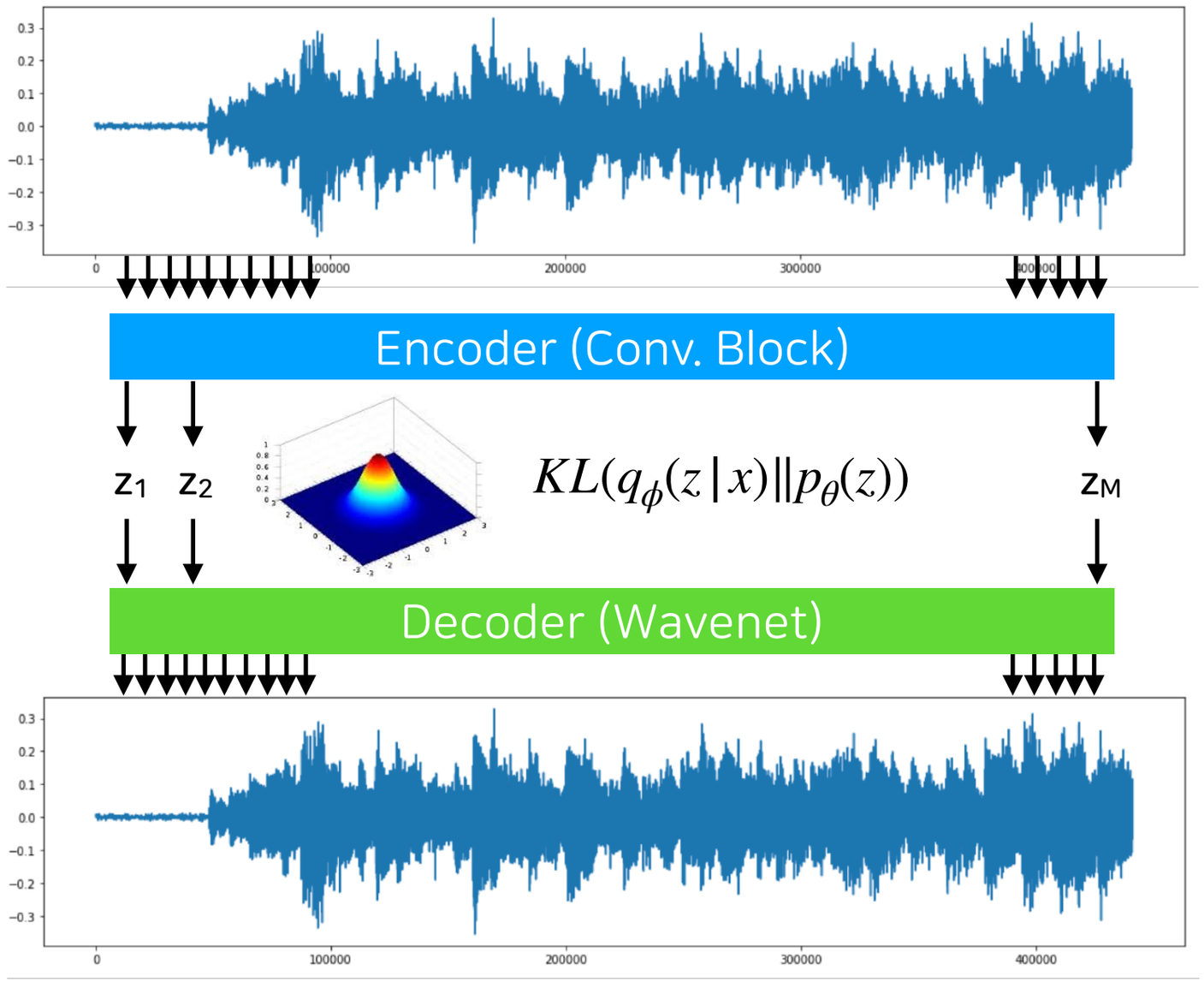
[Time, 1]

[Frame, |Z|]

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

VAE 적용!



Shape

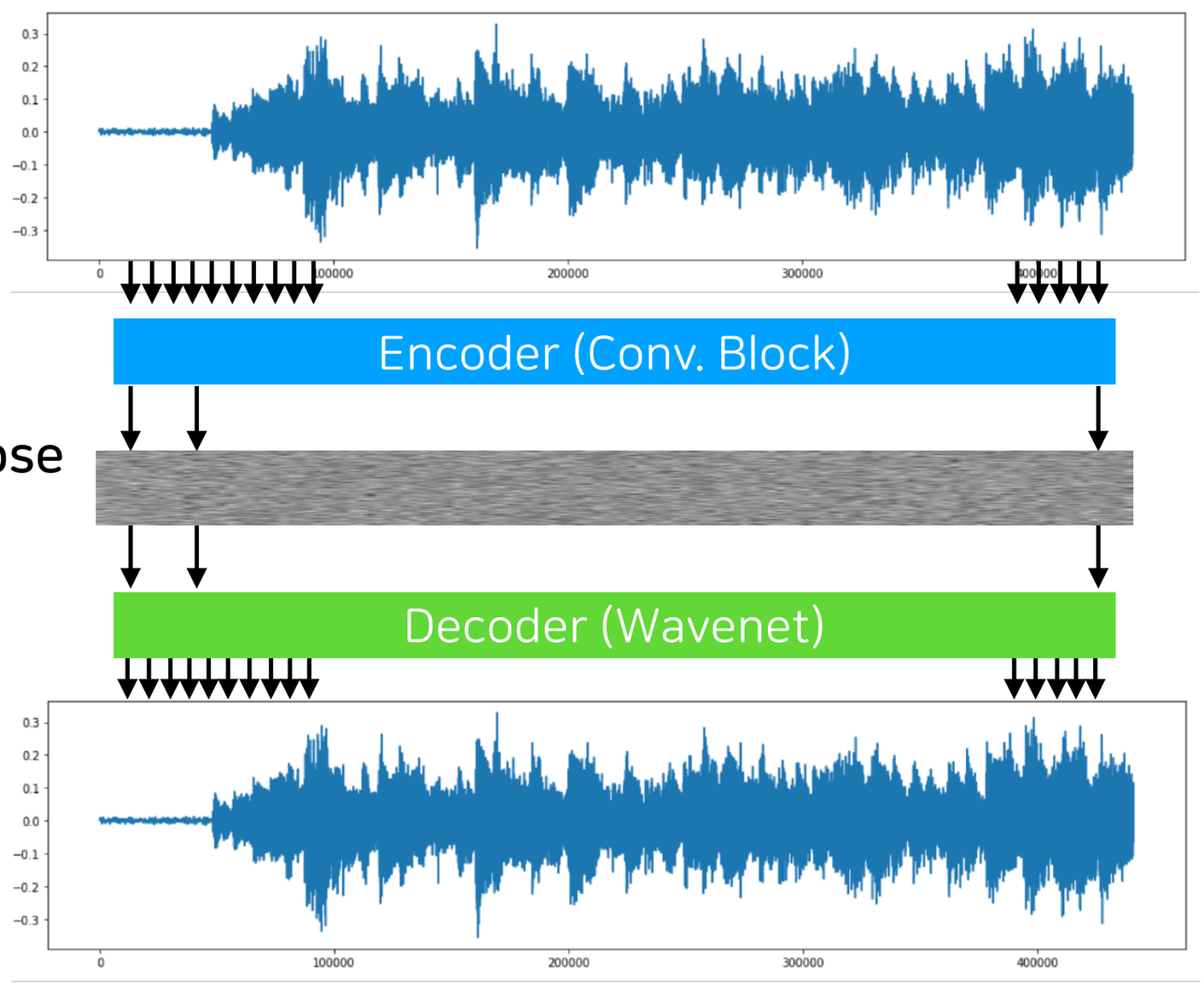
[Time, 1]

[Frame, |Z|]

[Time, 1]

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

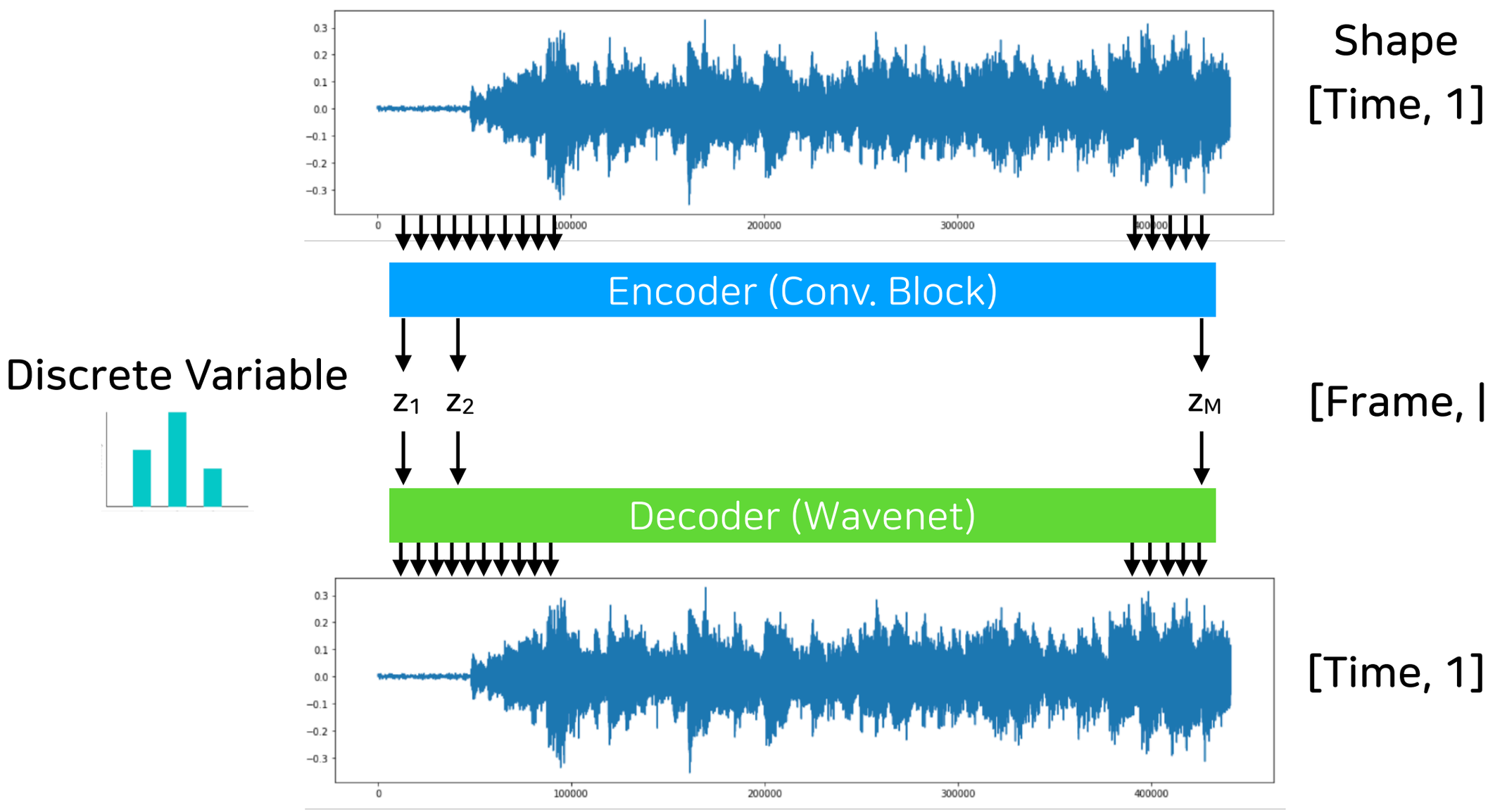


Posterior Collapse

$$q_\phi(z \mid x) \approx p_\theta(z)$$

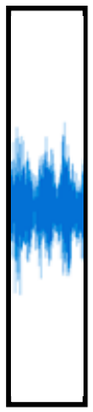
2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



2. Autoregressive Models

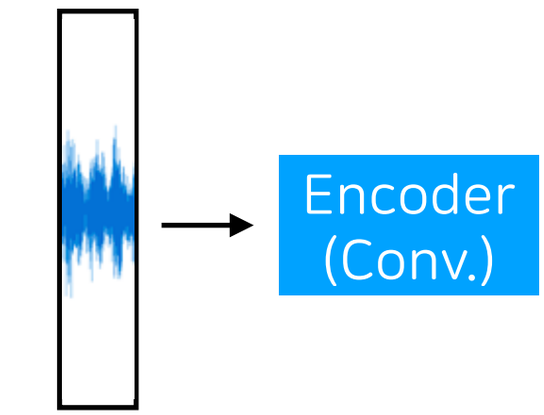
2.2 VQ-VAE : Neural Discrete Representation Learning



Frame

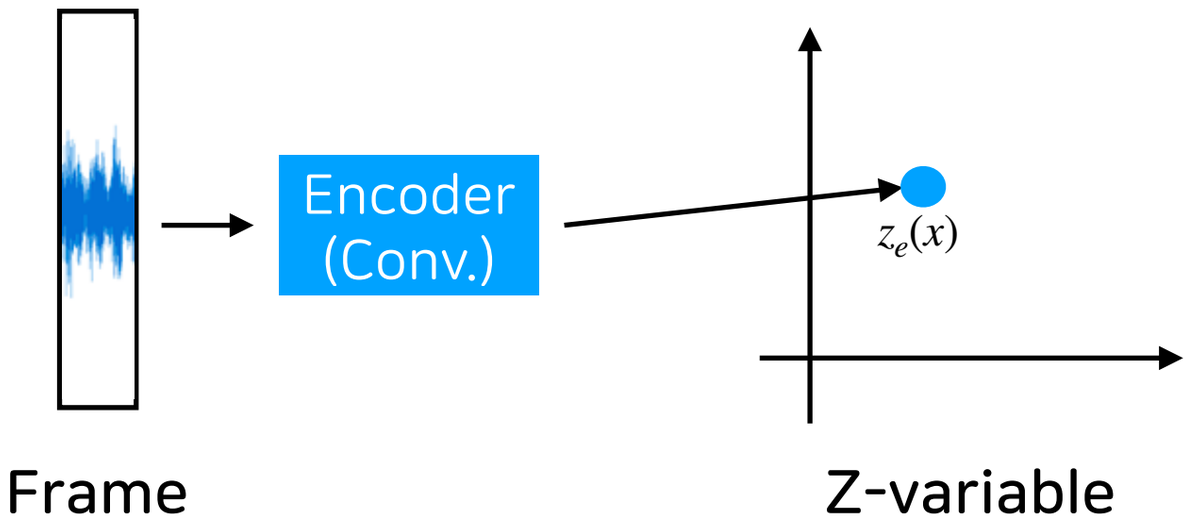
2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



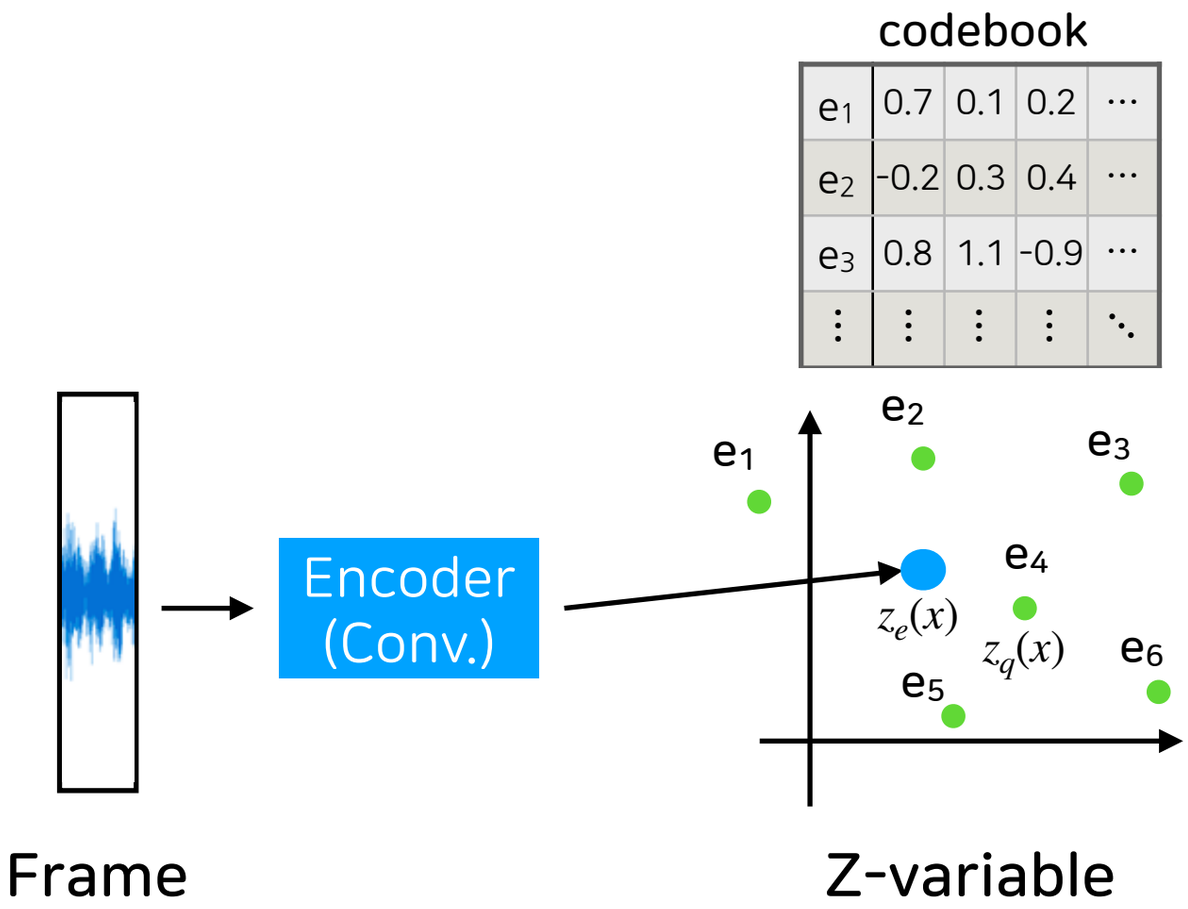
2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

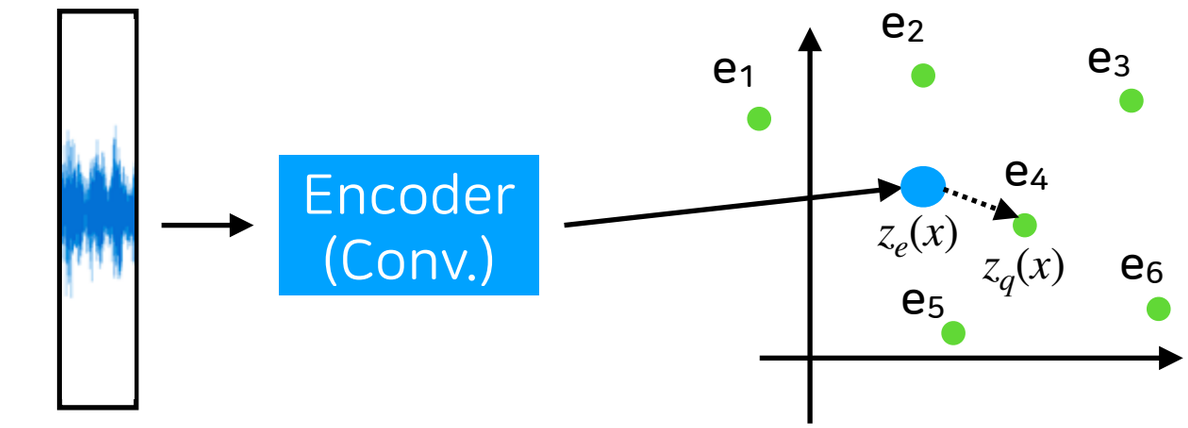


2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

codebook

e_1	0.7	0.1	0.2	⋯
e_2	-0.2	0.3	0.4	⋯
e_3	0.8	1.1	-0.9	⋯
⋮	⋮	⋮	⋮	⋱



Frame

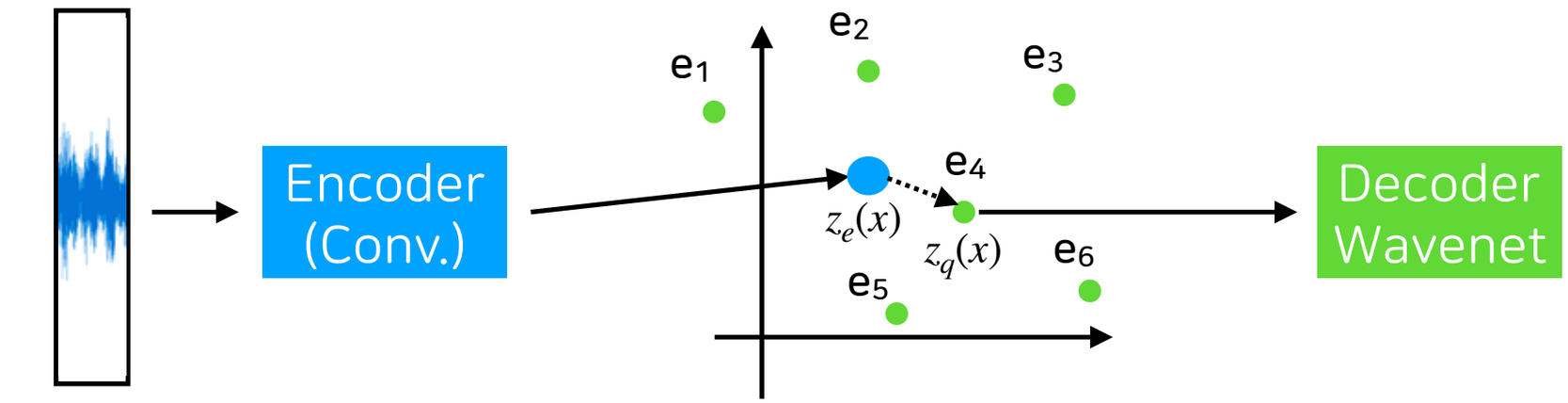
Z-variable

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

codebook

e_1	0.7	0.1	0.2	⋯
e_2	-0.2	0.3	0.4	⋯
e_3	0.8	1.1	-0.9	⋯
⋮	⋮	⋮	⋮	⋱

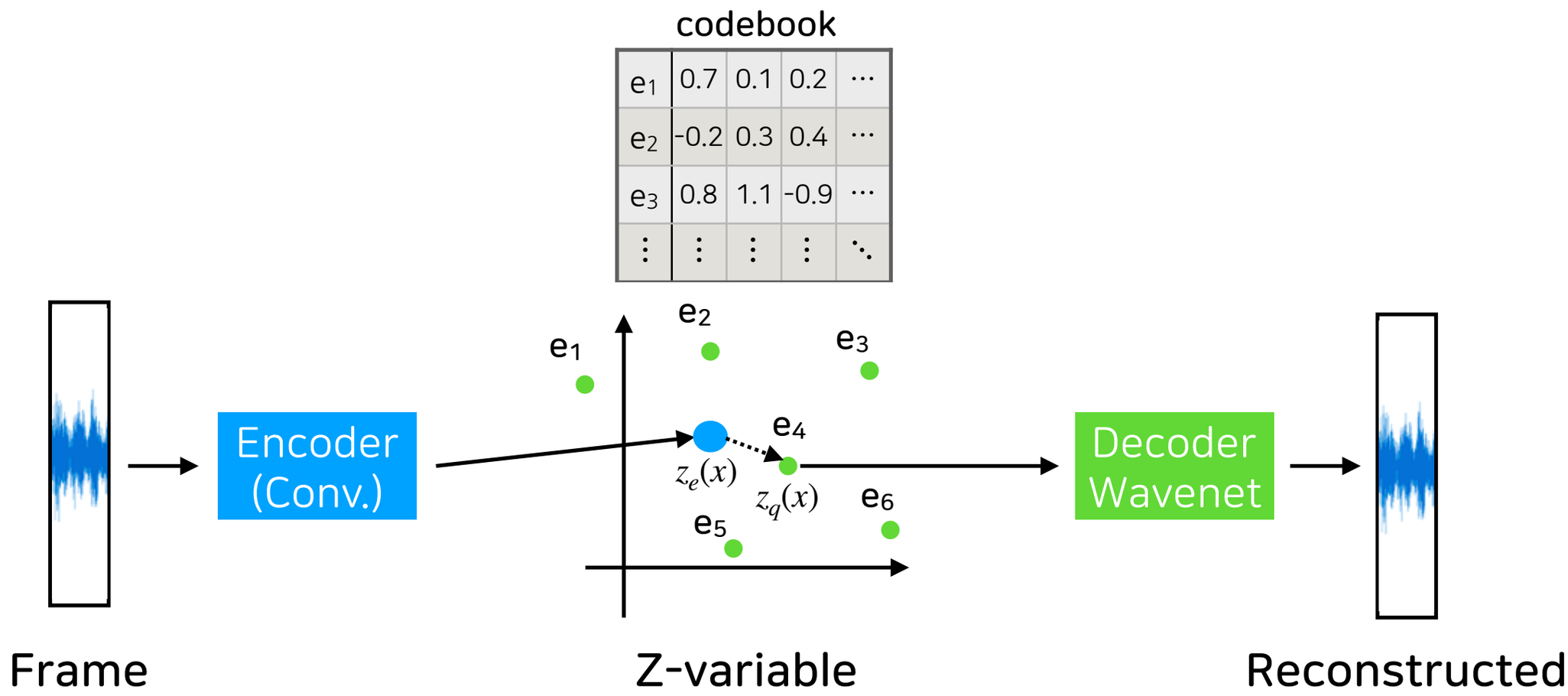


Frame

Z-variable

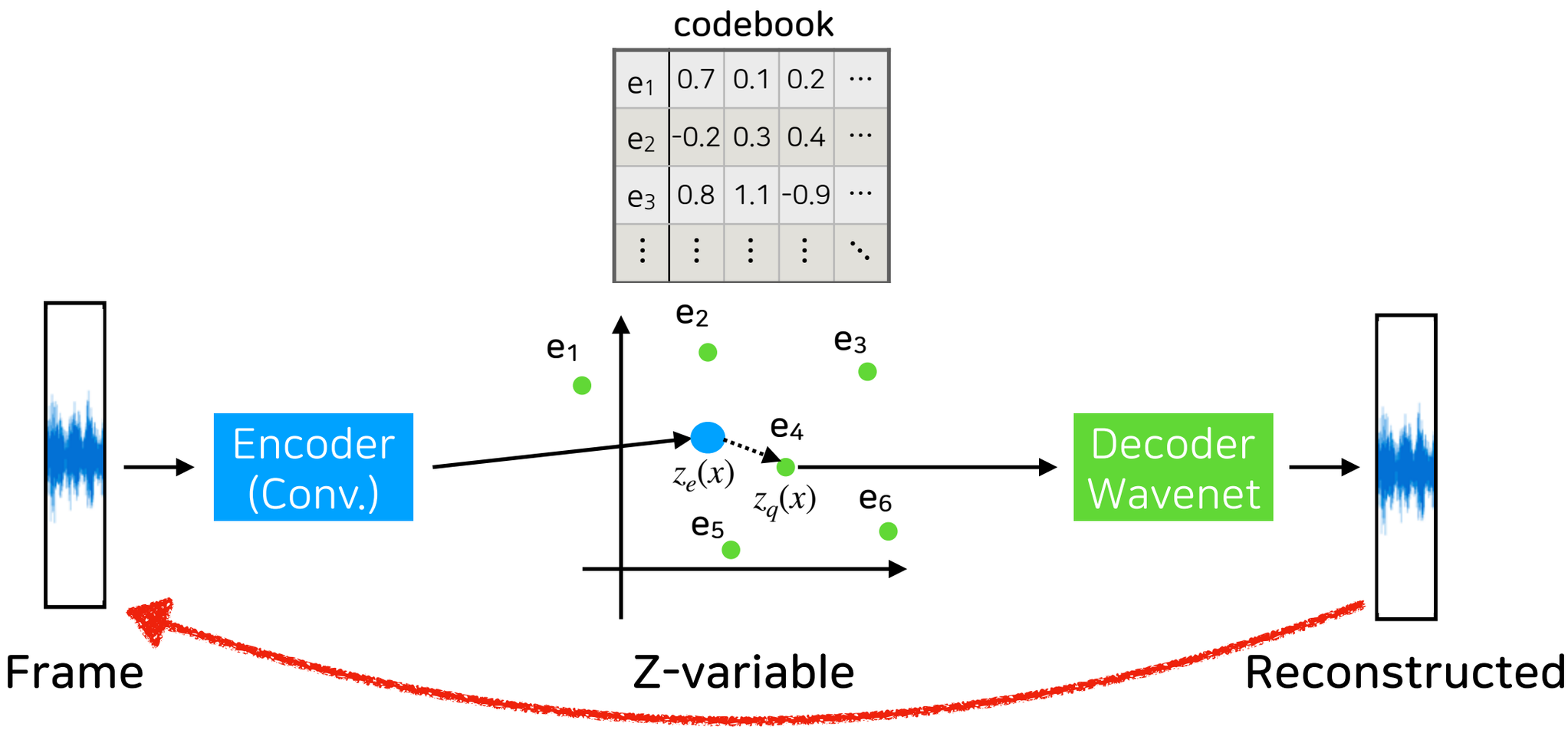
2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



$$Loss = -\,logp(x \mid z_q(x))$$

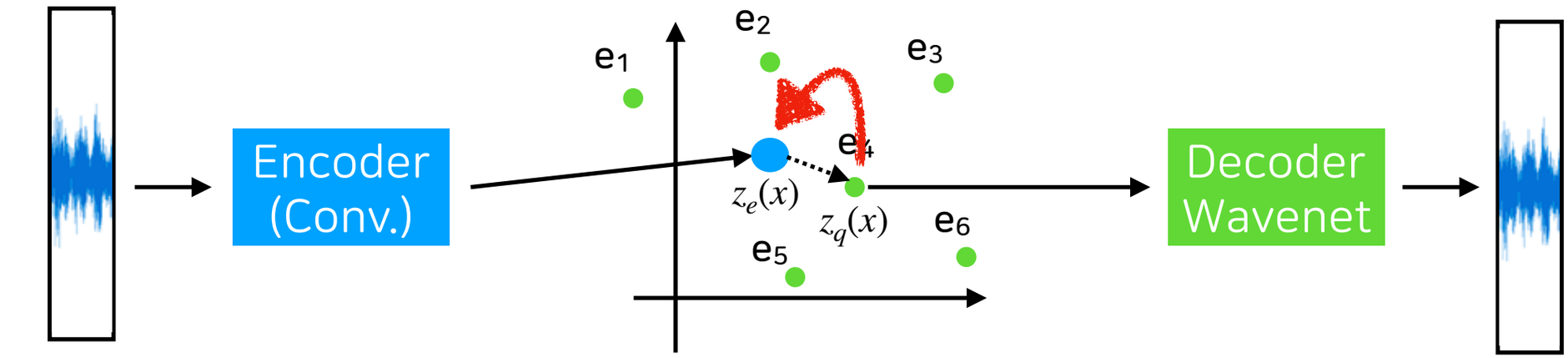
reconstruction

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

codebook

e_1	0.7	0.1	0.2	⋯
e_2	-0.2	0.3	0.4	⋯
e_3	0.8	1.1	-0.9	⋯
⋮	⋮	⋮	⋮	⋱



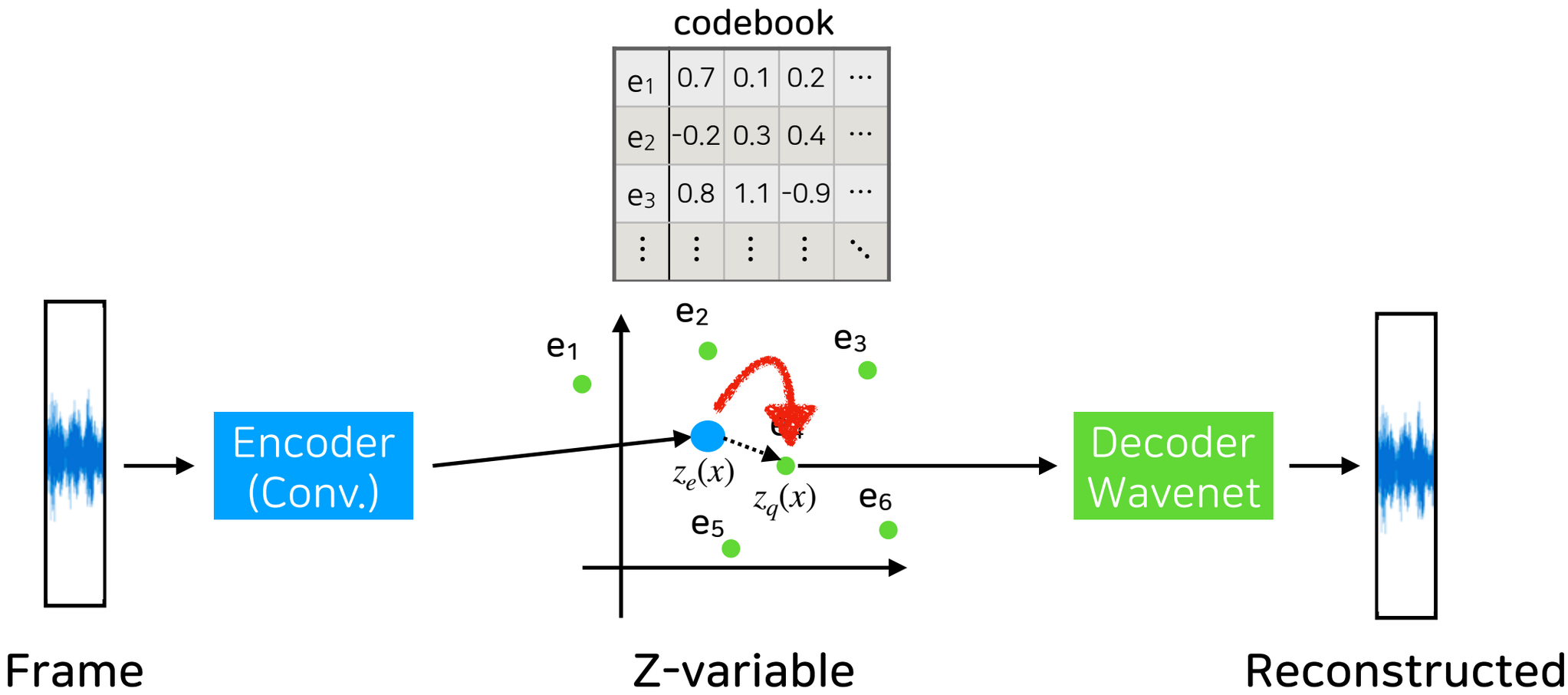
Frame Z-variable Reconstructed

$$Loss = -logp(x|z_q(x)) + \|sg[z_e(x)] - e\|_2^2$$

reconstruction codebook

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning



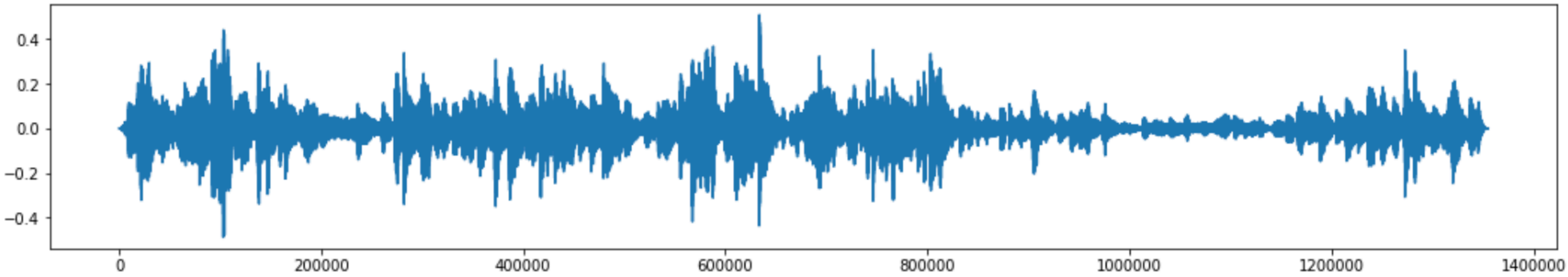
$$Loss = -logp(x|z_q(x)) + \|sg[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - sg[e]\|_2^2$$

reconstruction　　codebook　　commitment

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

Generated Music. 1

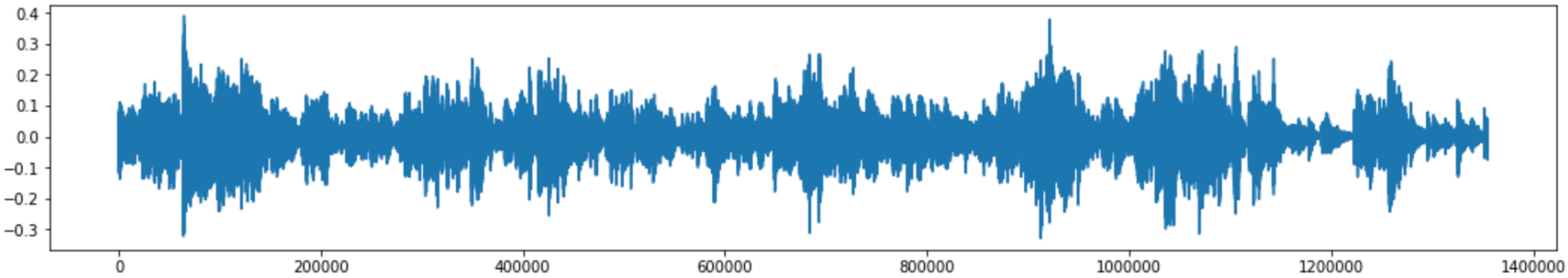


Audio Reference : https://drive.google.com/drive/folders/1fvS-DU8AcK078-5k6WGudiBn0XSeE0_D

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

Generated Music. 2

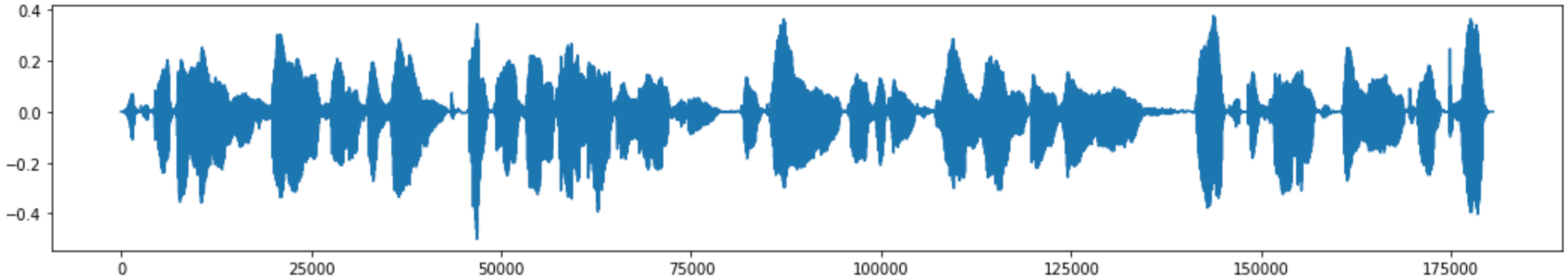


Audio Reference : https://drive.google.com/drive/folders/1fvS-DU8AcK078-5k6WGudiBn0XSeE0_D

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

Generated Speech. 1

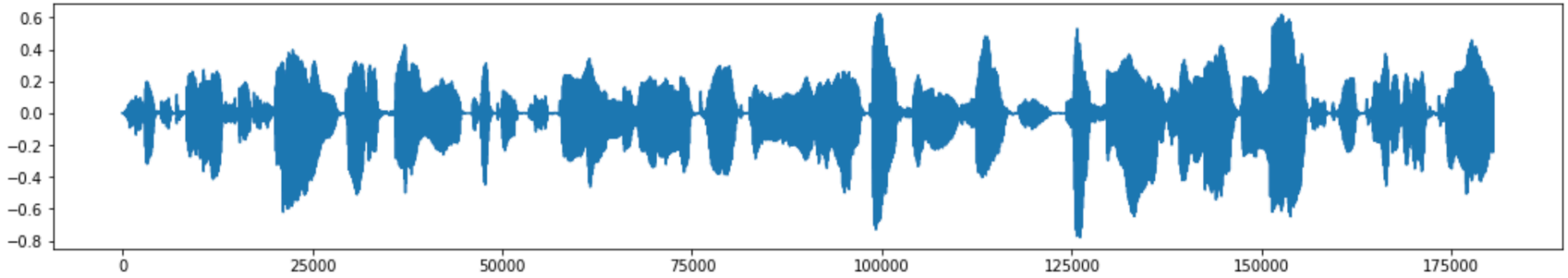


Audio Reference : https://avdnoord.github.io/homepage/vqvae/

2. Autoregressive Models

2.2 VQ-VAE : Neural Discrete Representation Learning

Generated Speech. 2

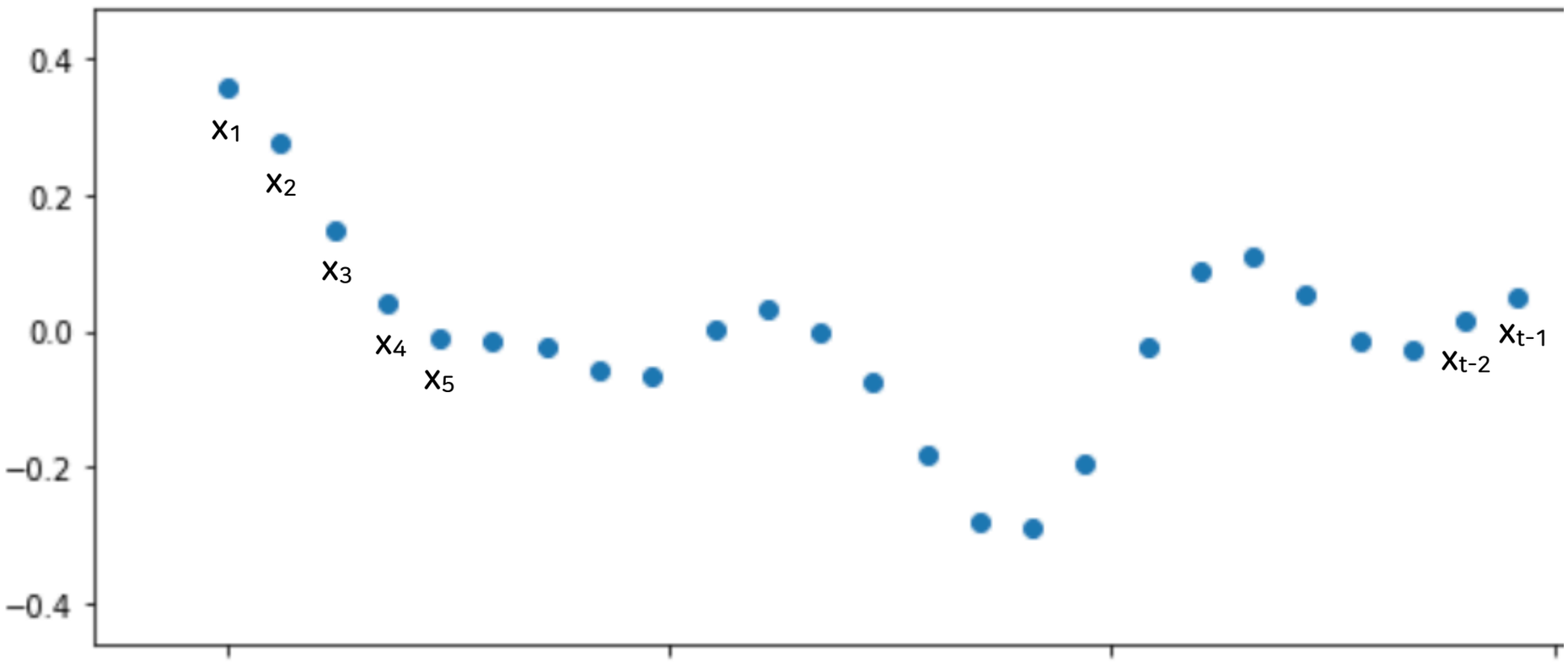


Audio Reference : https://avdnoord.github.io/homepage/vqvae/

Sparse Transformer

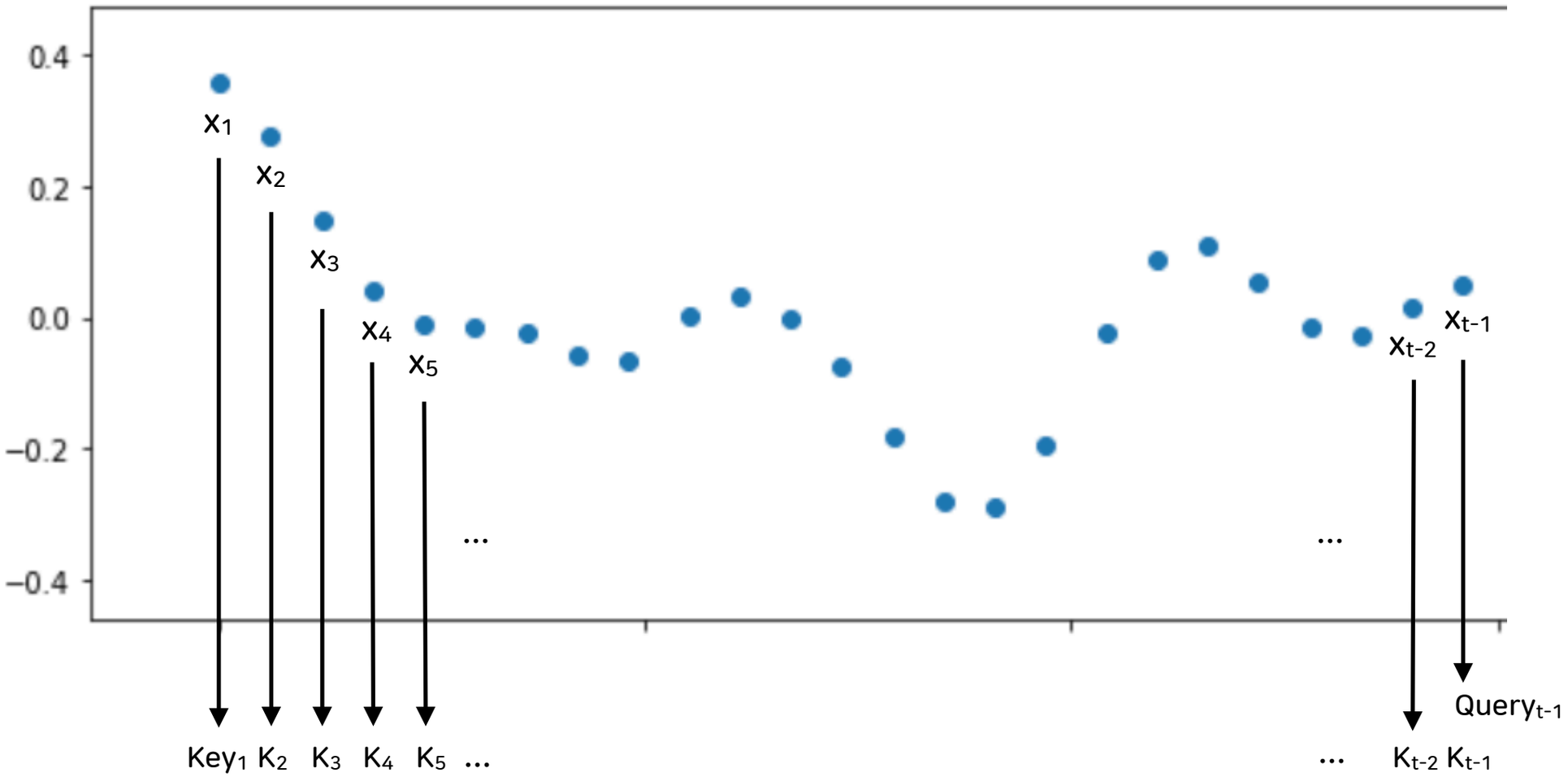
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



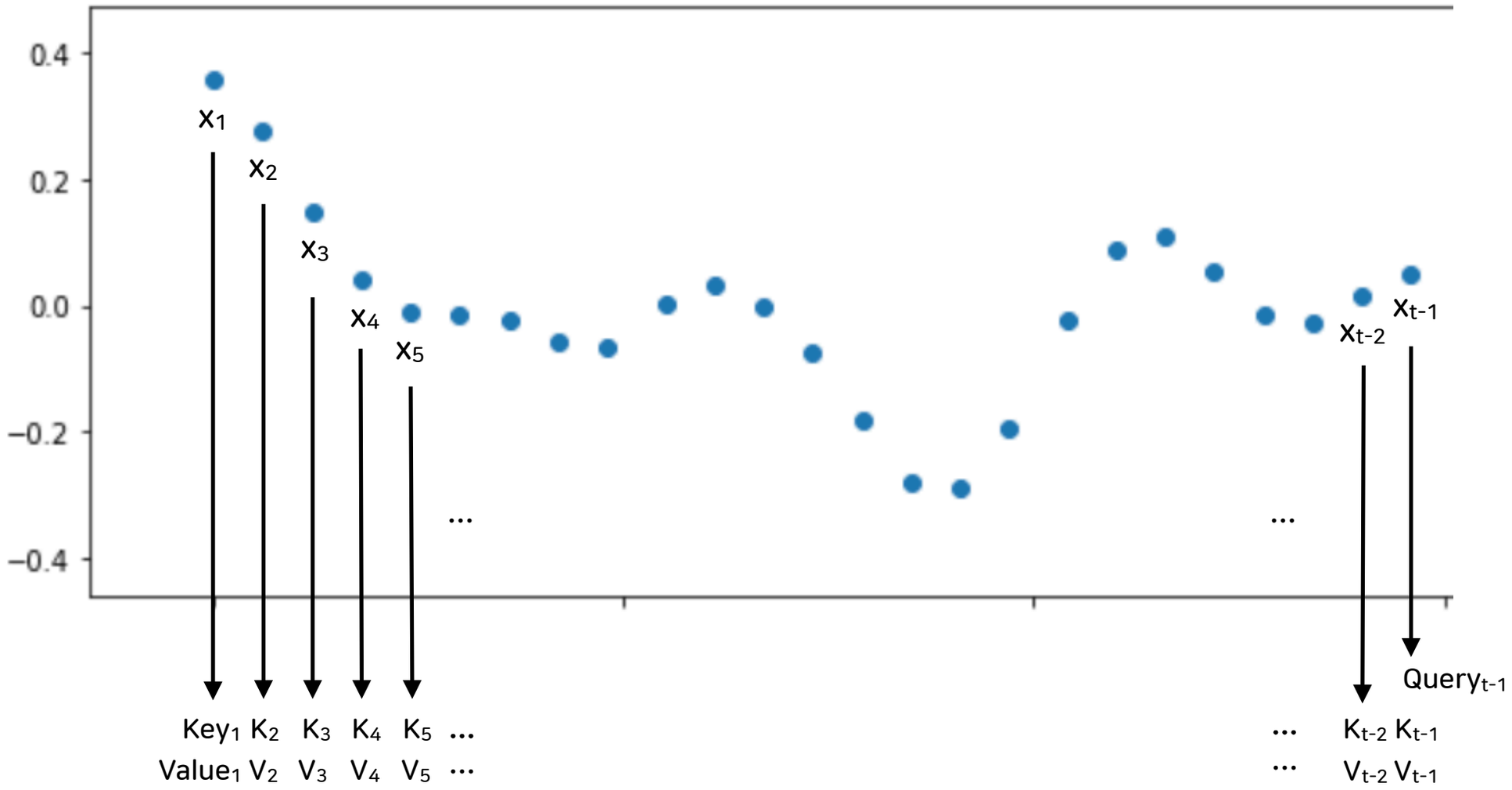
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



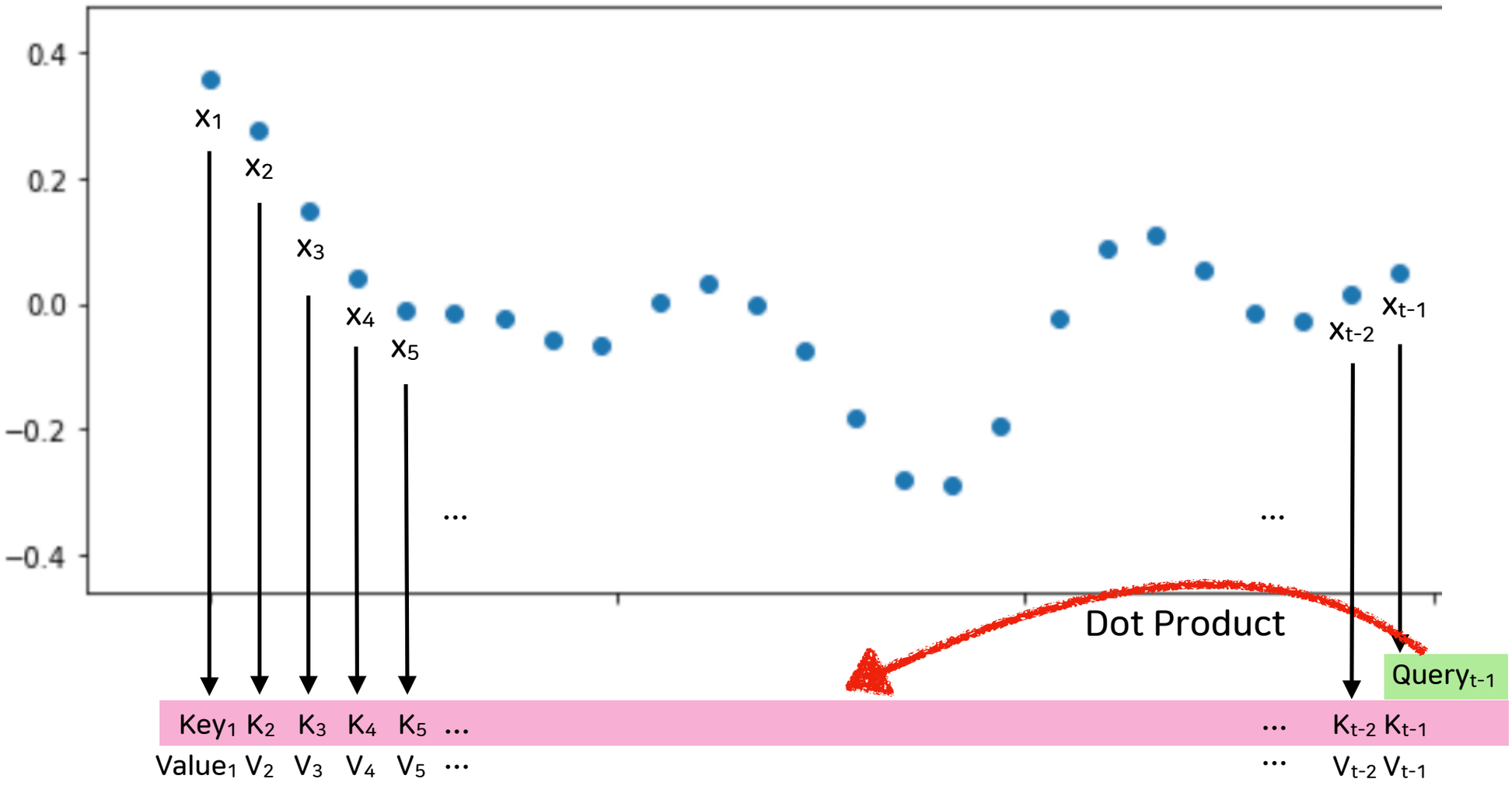
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



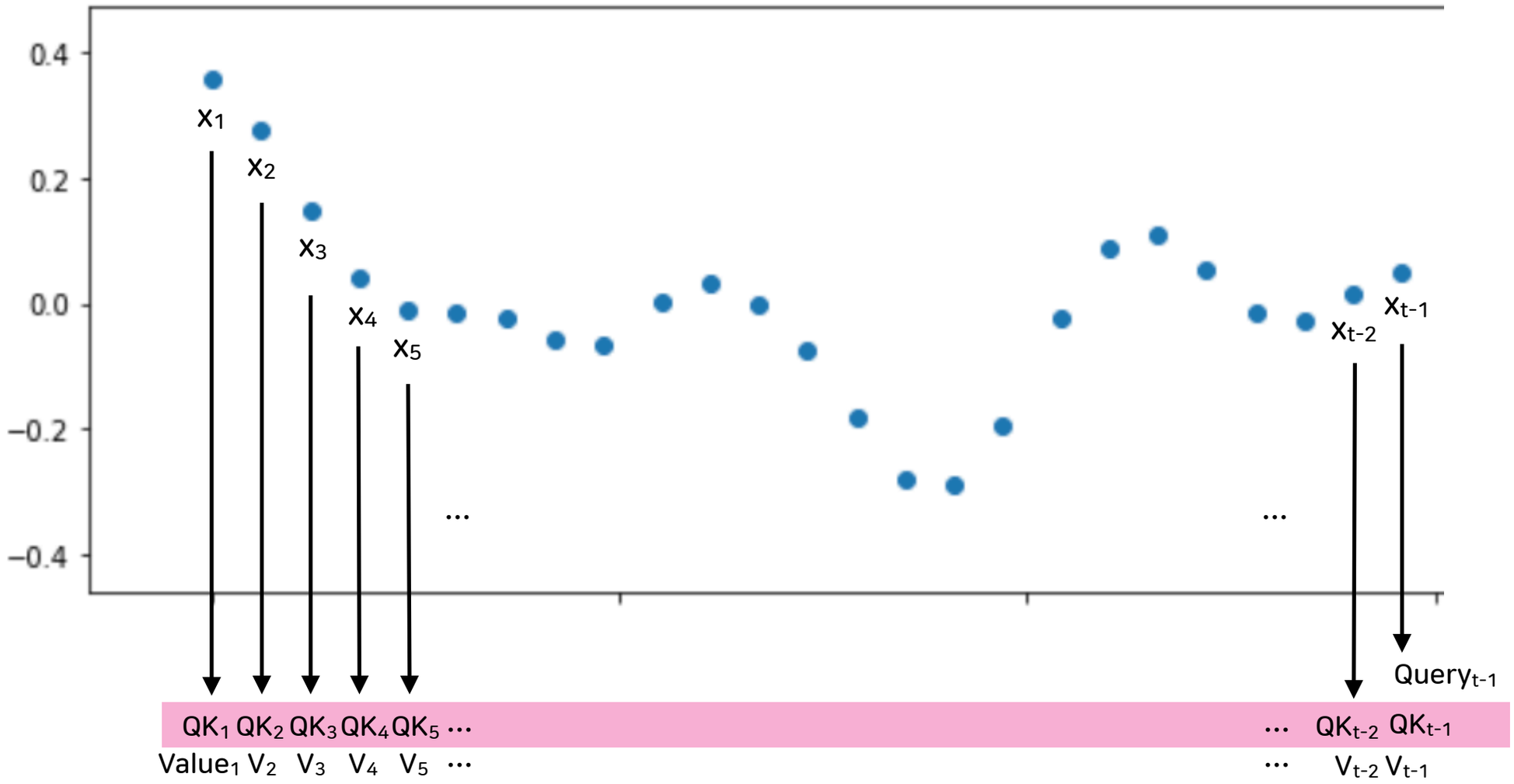
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



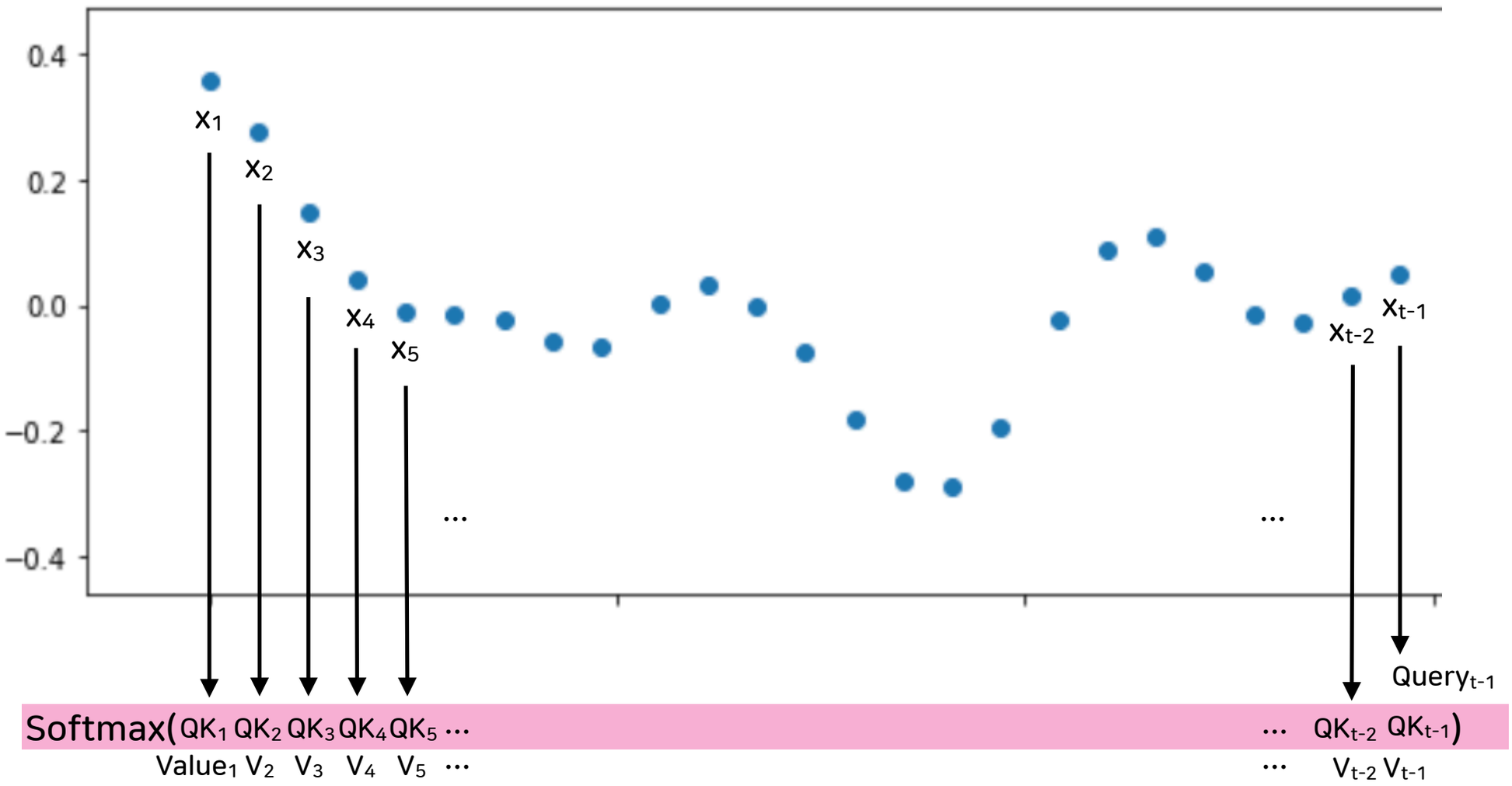
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



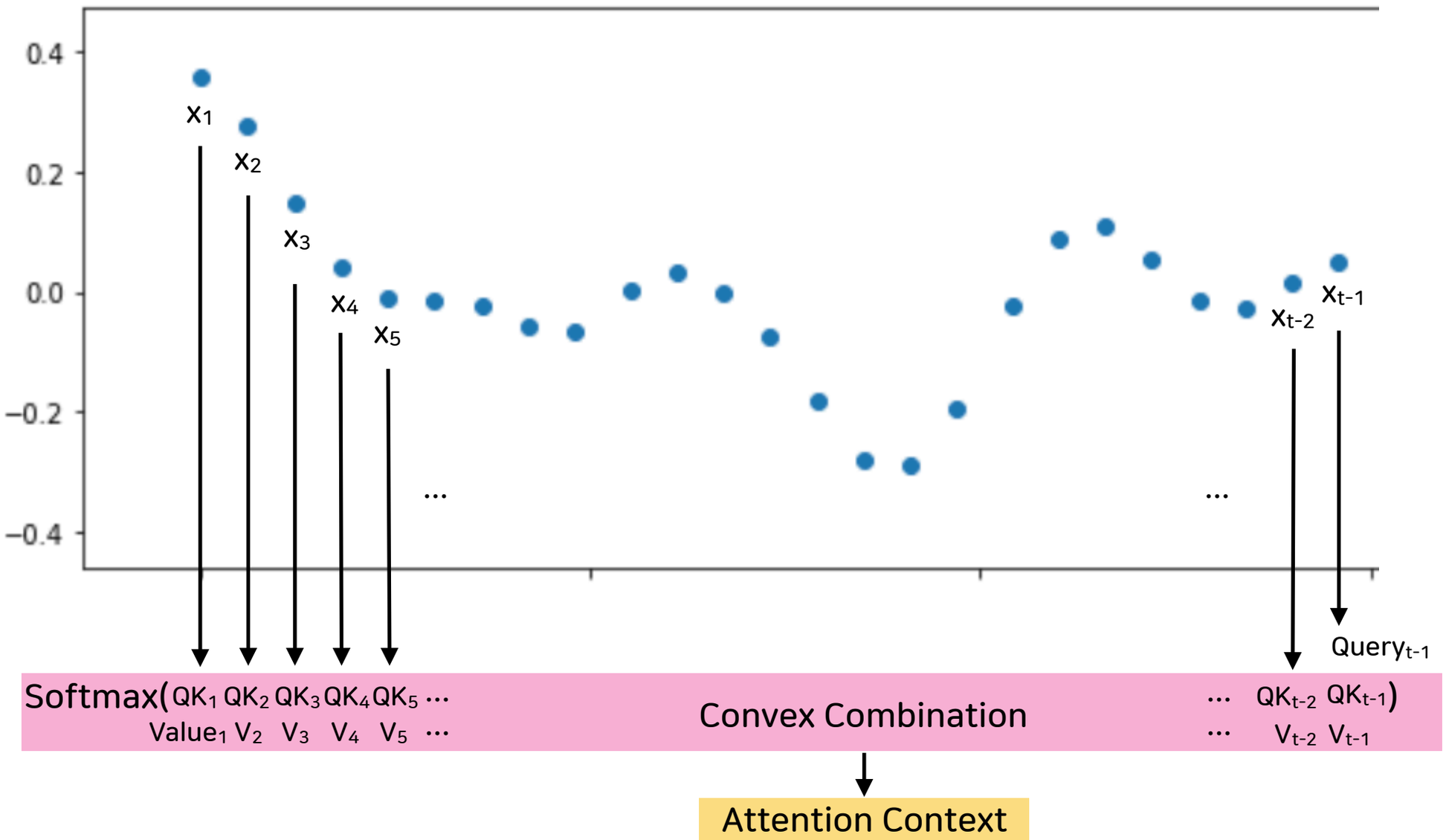
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



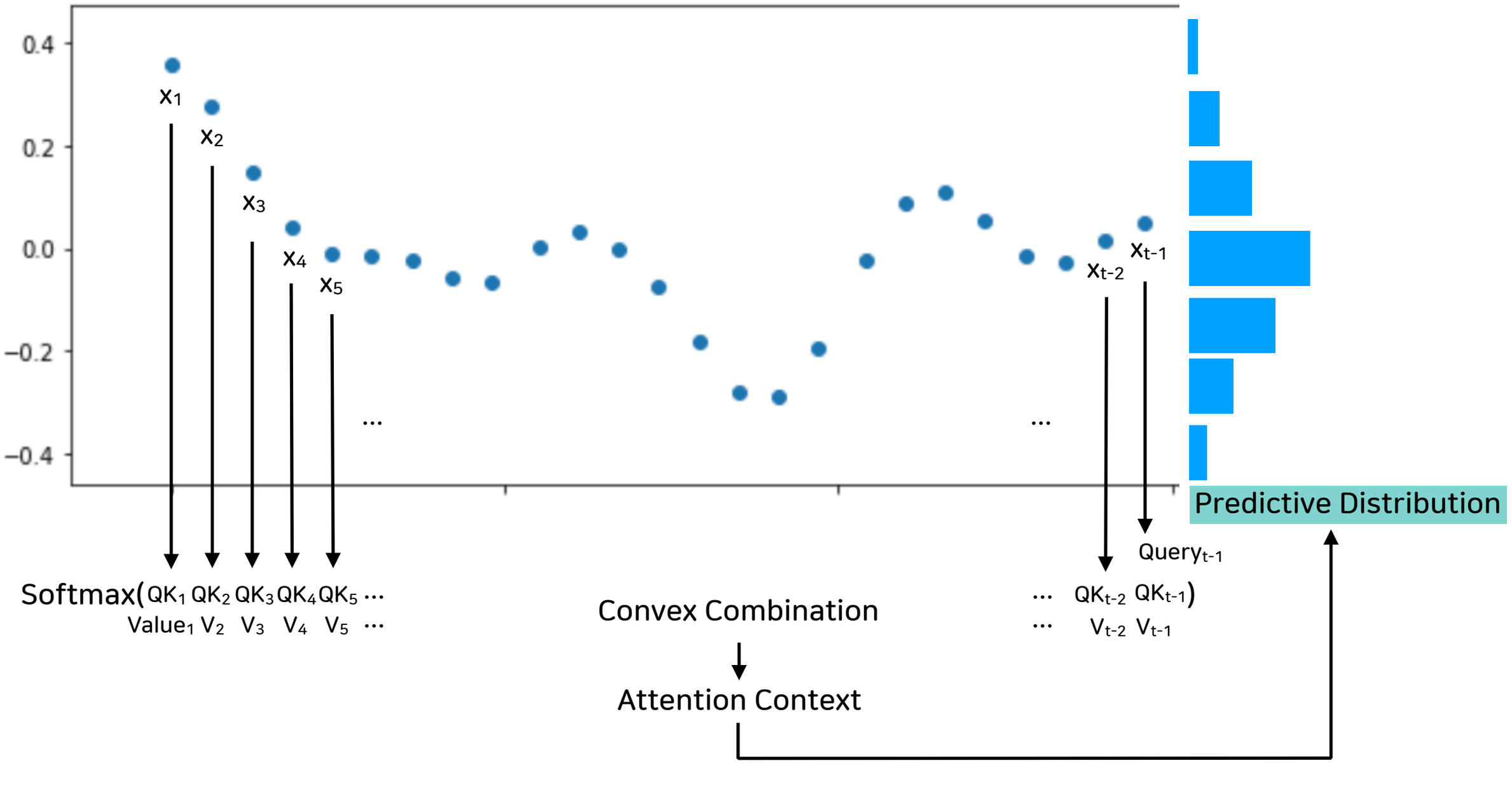
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



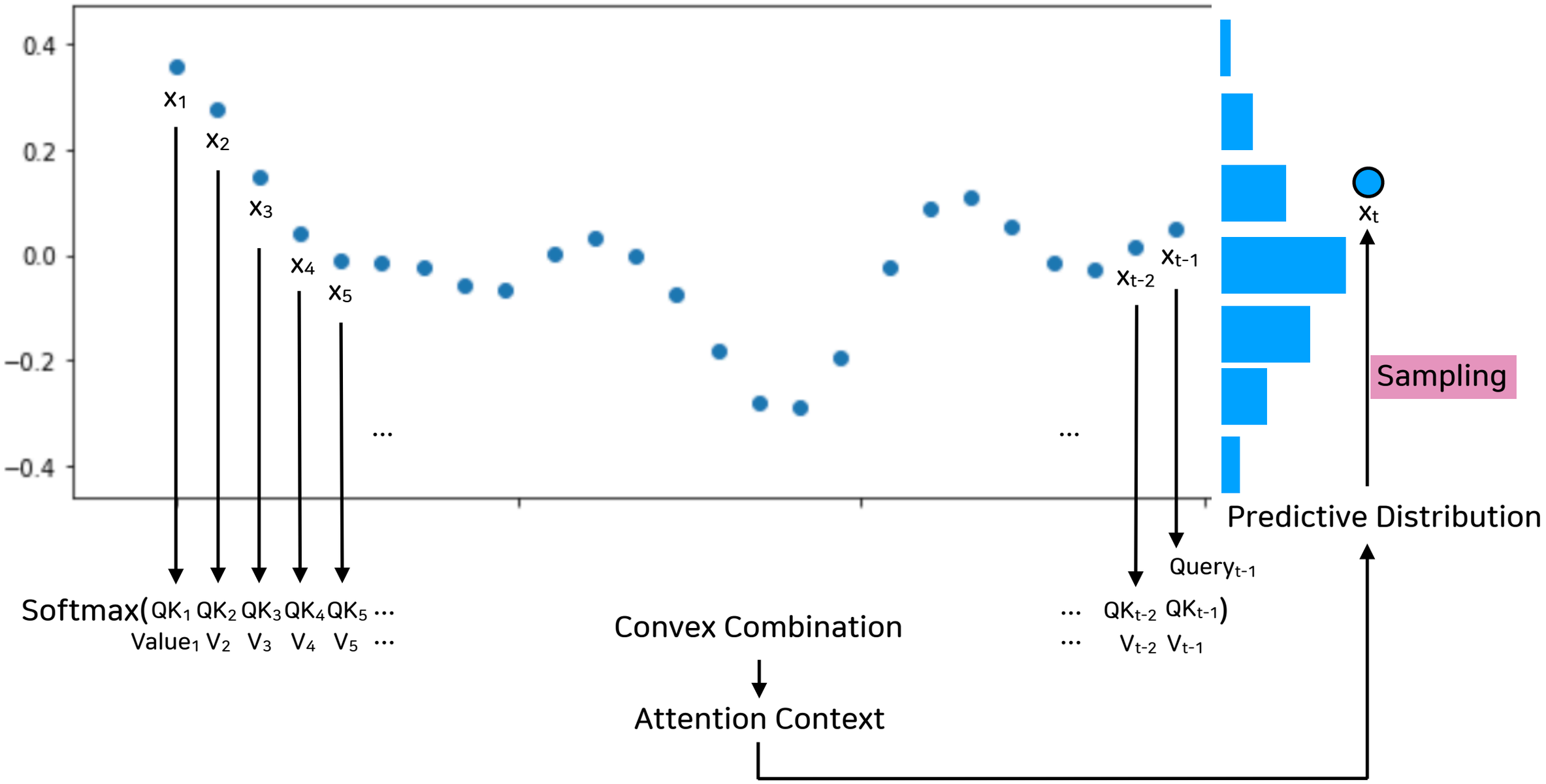
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



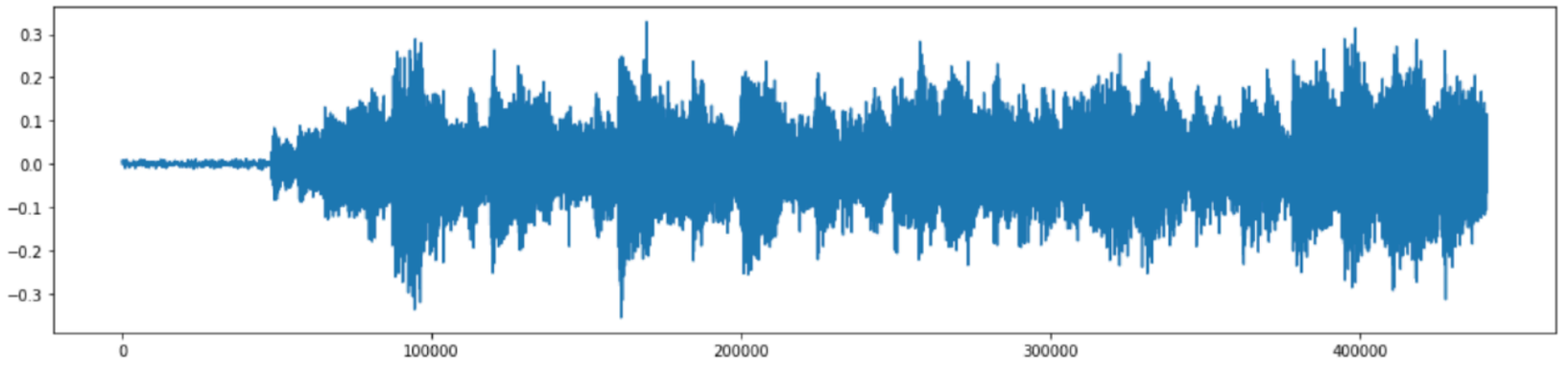
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



2. Autoregressive Models

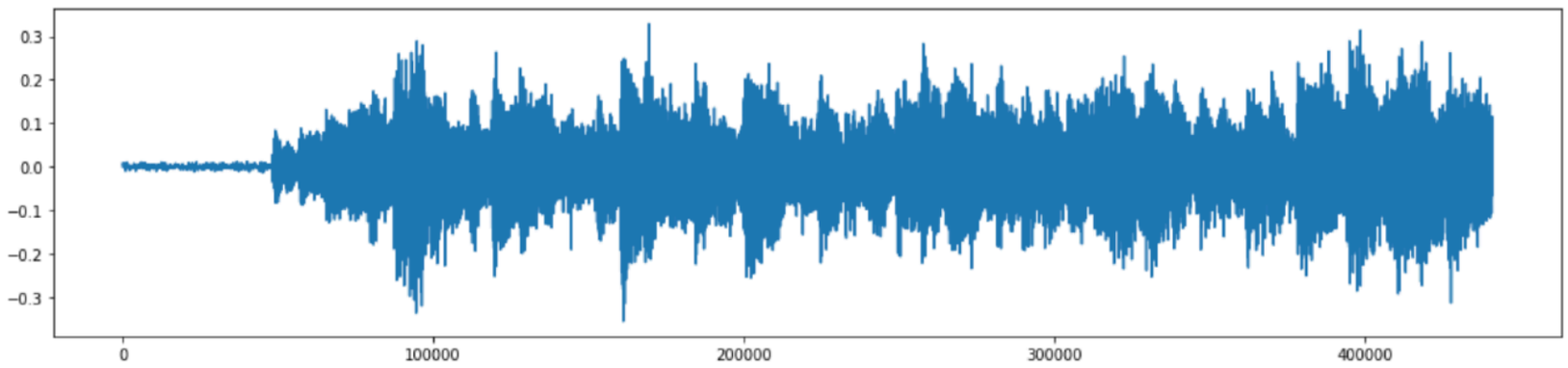
2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



1초=44100샘플

2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



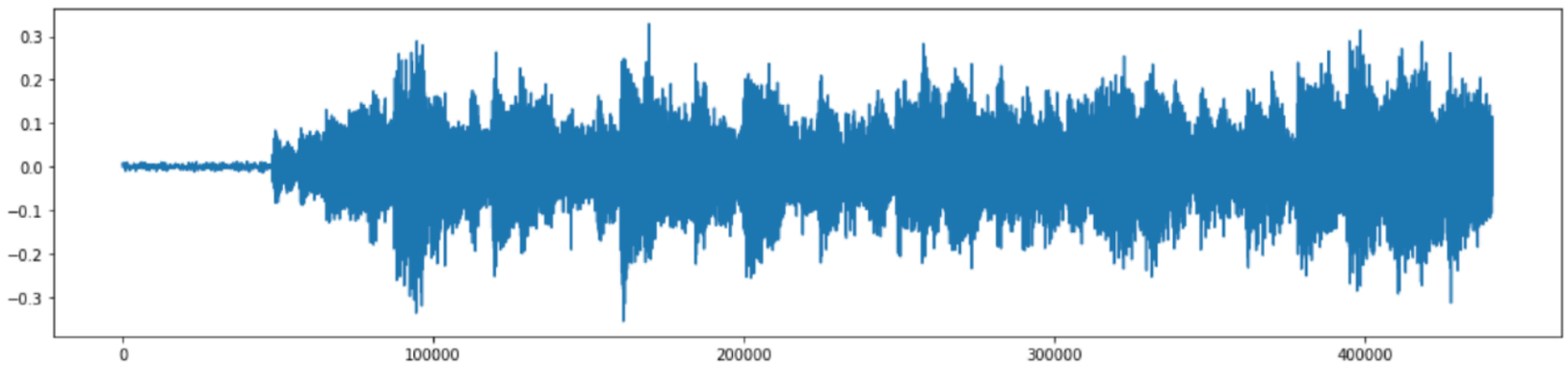
1초=44100샘플

오디오는 1초에 16000-48000 샘플링

Transformer를 적용하기에 길이가 너무 길다.

2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



1초=44100샘플

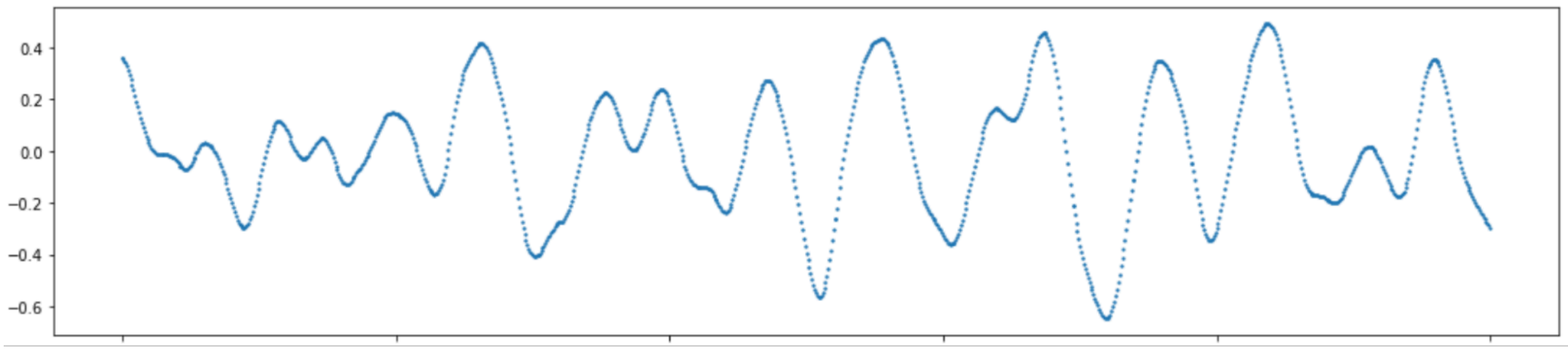
오디오는 1초에 16000-48000 샘플링

Transformer를 적용하기에 길이가 너무 길다.

어텐션을 띄엄띄엄(Sparse)해주자.

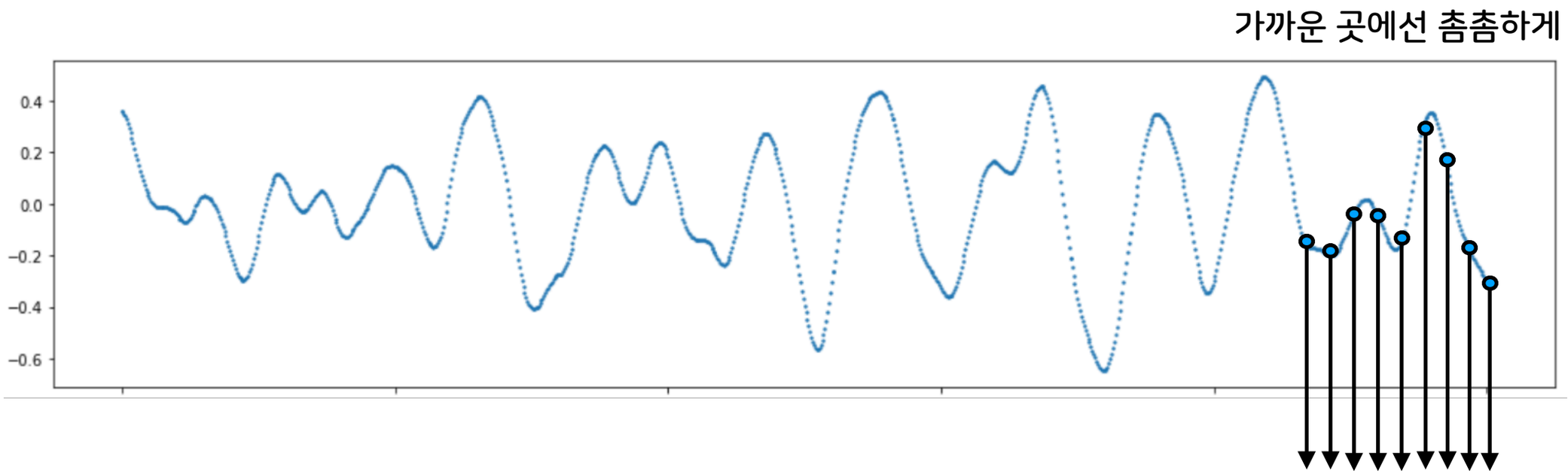
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



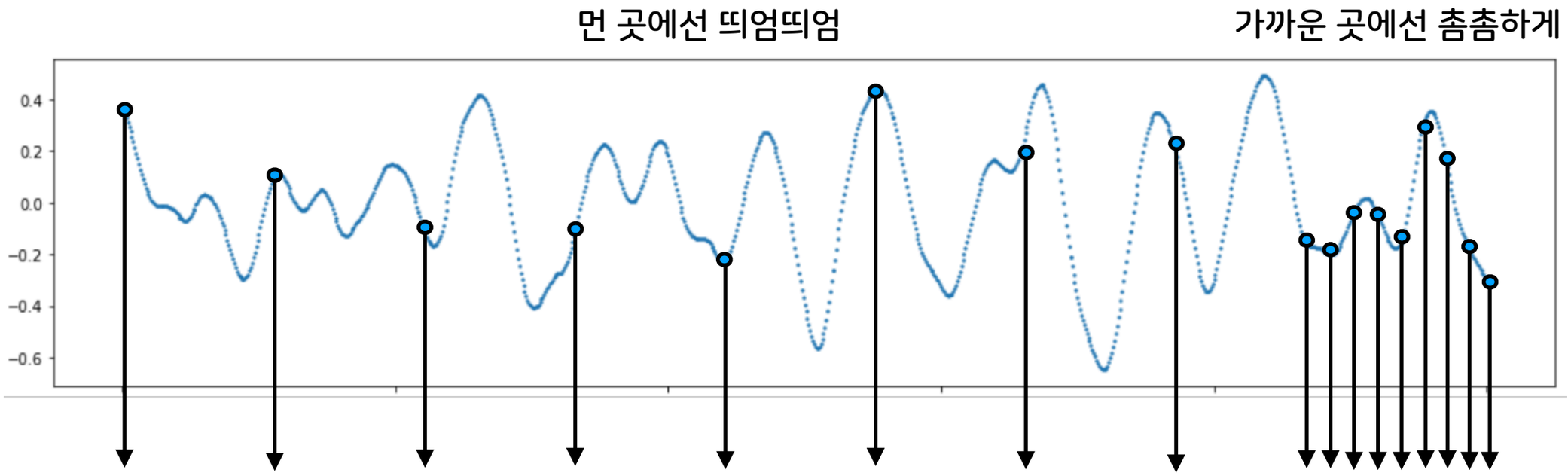
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



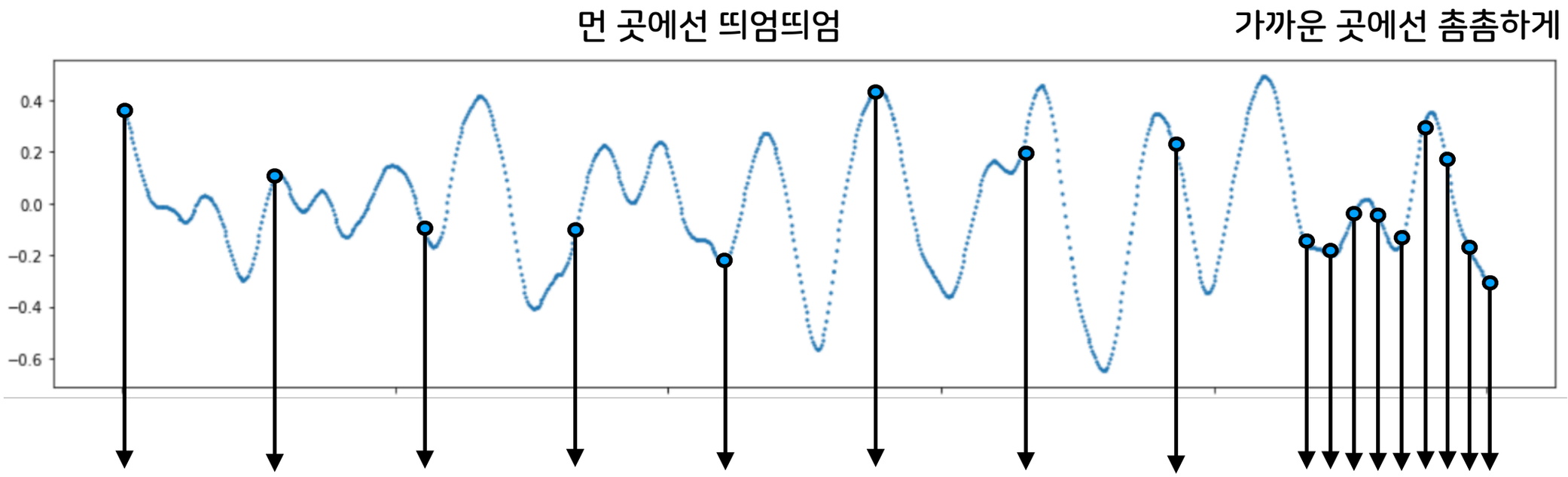
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



2. Autoregressive Models

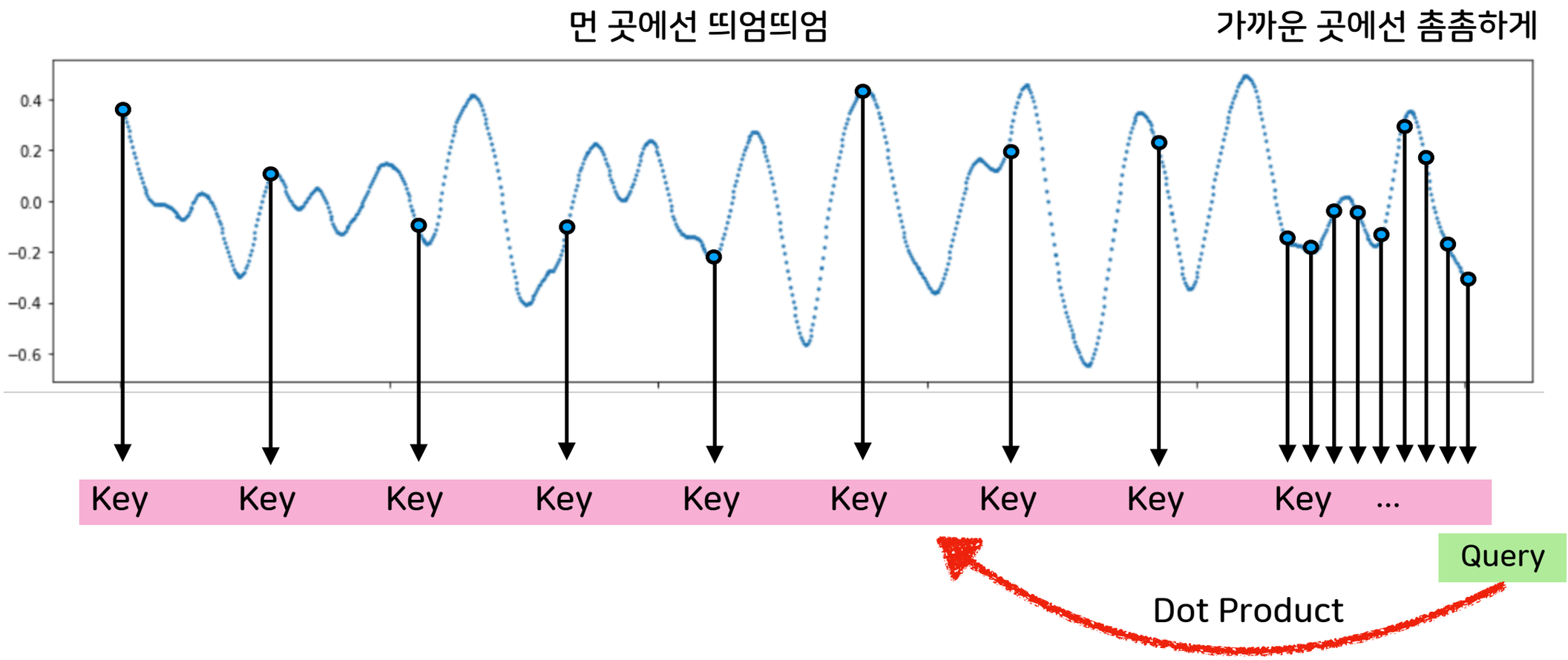
2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



Key	Key	Key	Key	Key	Key	Key	Key	Key	…

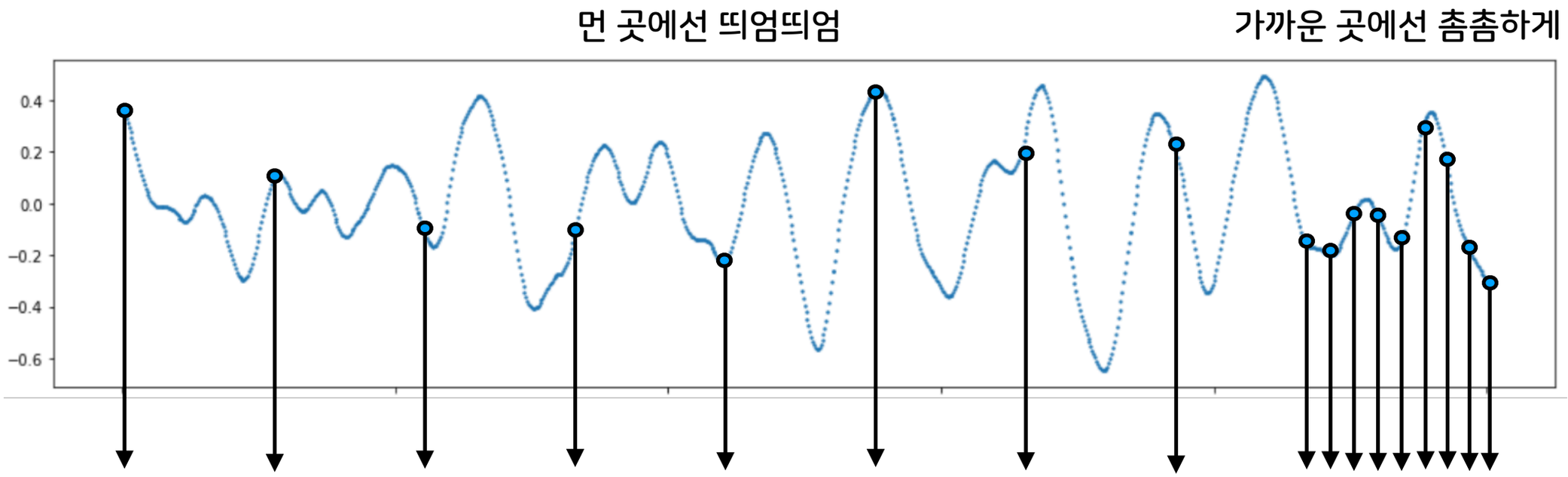
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



2. Autoregressive Models

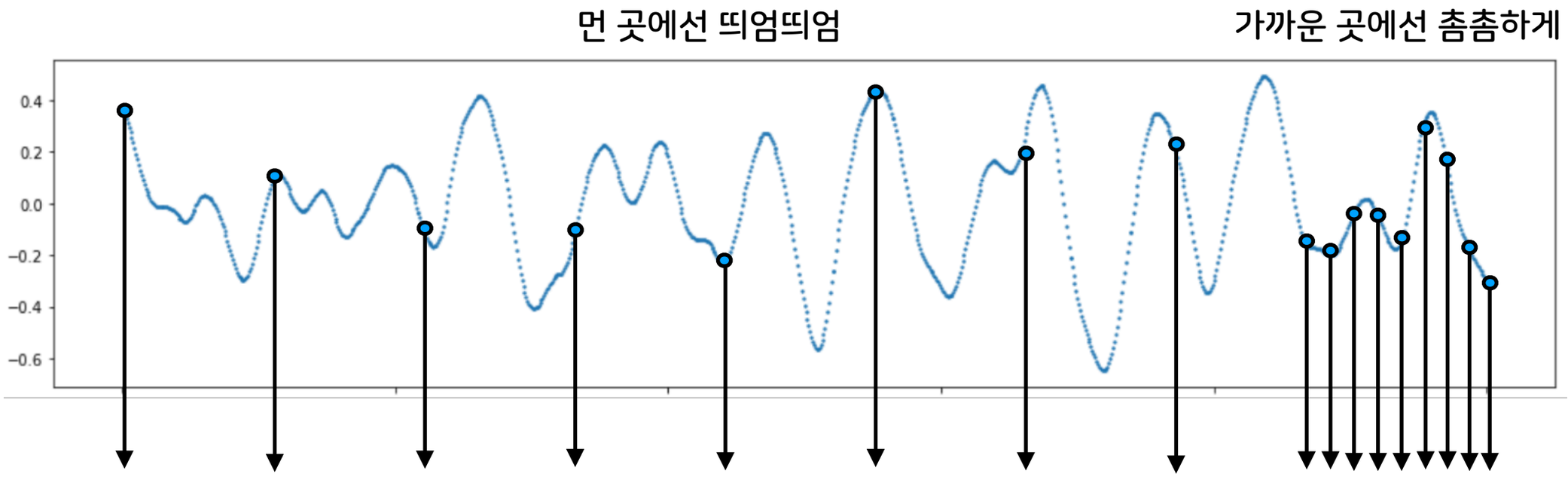
2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



QK	QK	QK	QK	QK	QK	QK	QK	QK	…

2. Autoregressive Models

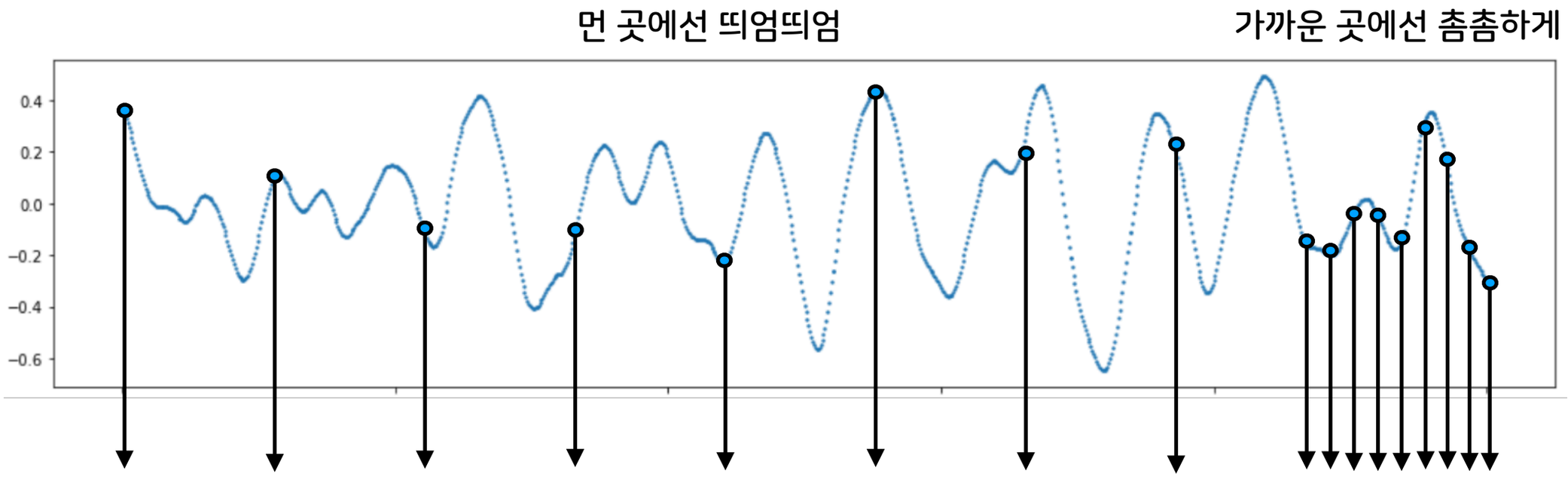
2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



Softmax(QK QK QK QK QK QK QK QK QK …)

2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers

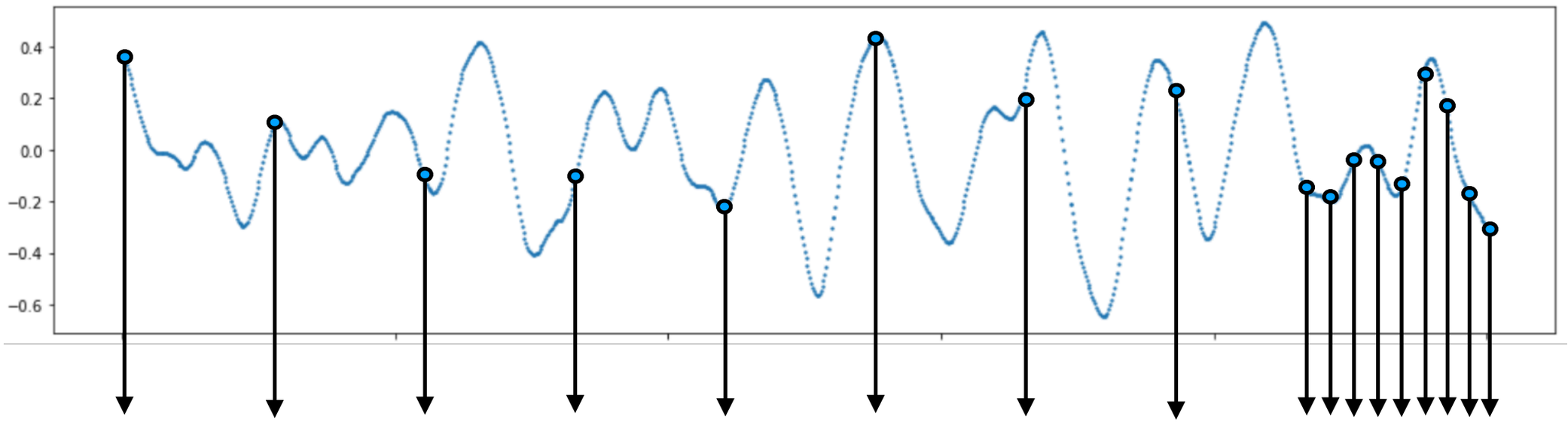


Softmax(QK Value QK Value QK Value QK Value QK Value QK Value QK Value QK Value QK Value …)
…

2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers

먼 곳에선 띄엄띄엄 가까운 곳에선 촘촘하게



Softmax(QK QK QK QK QK QK QK QK QK ...)

Value Value Value Value Value Value Value Value Value ...

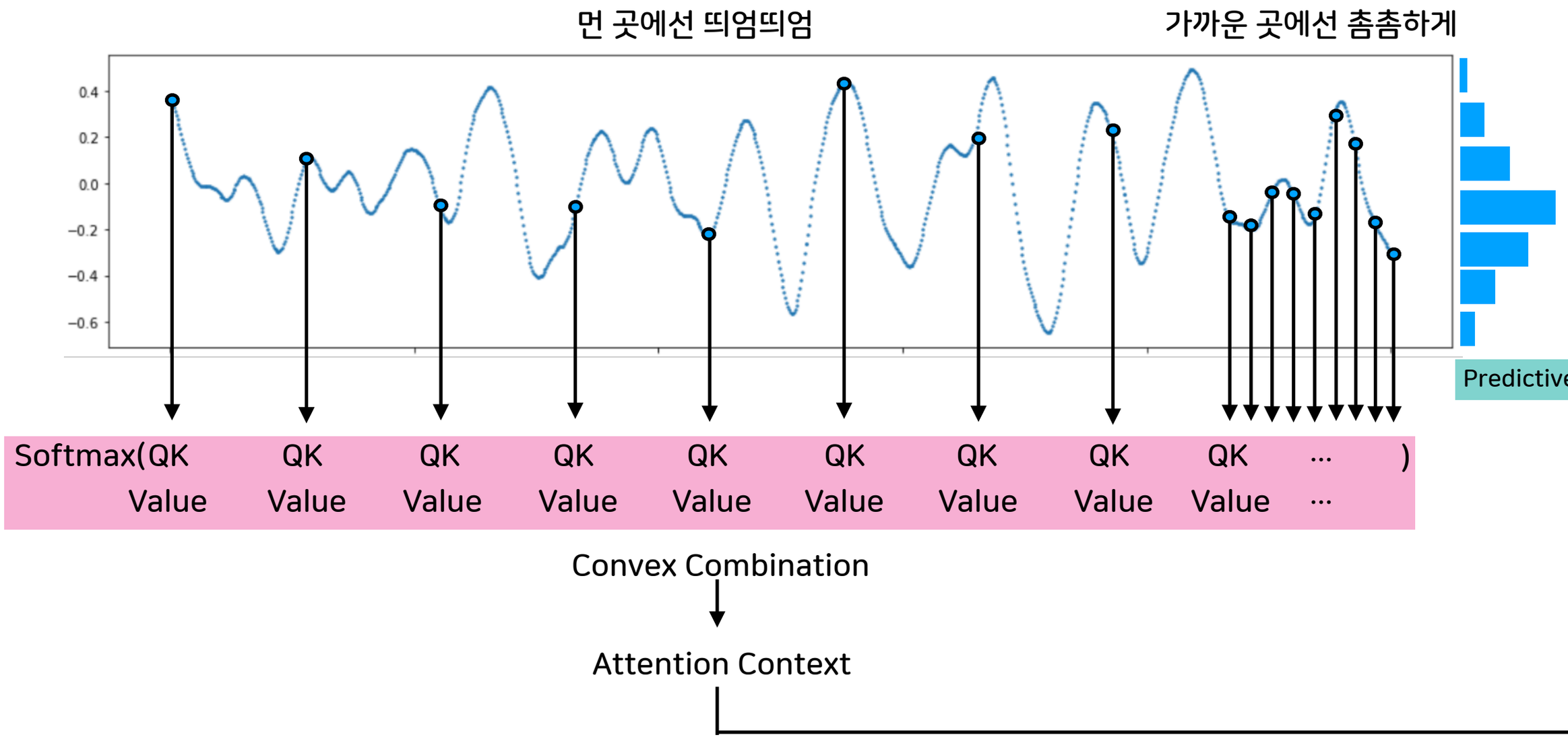
Convex Combination

↓

Attention Context

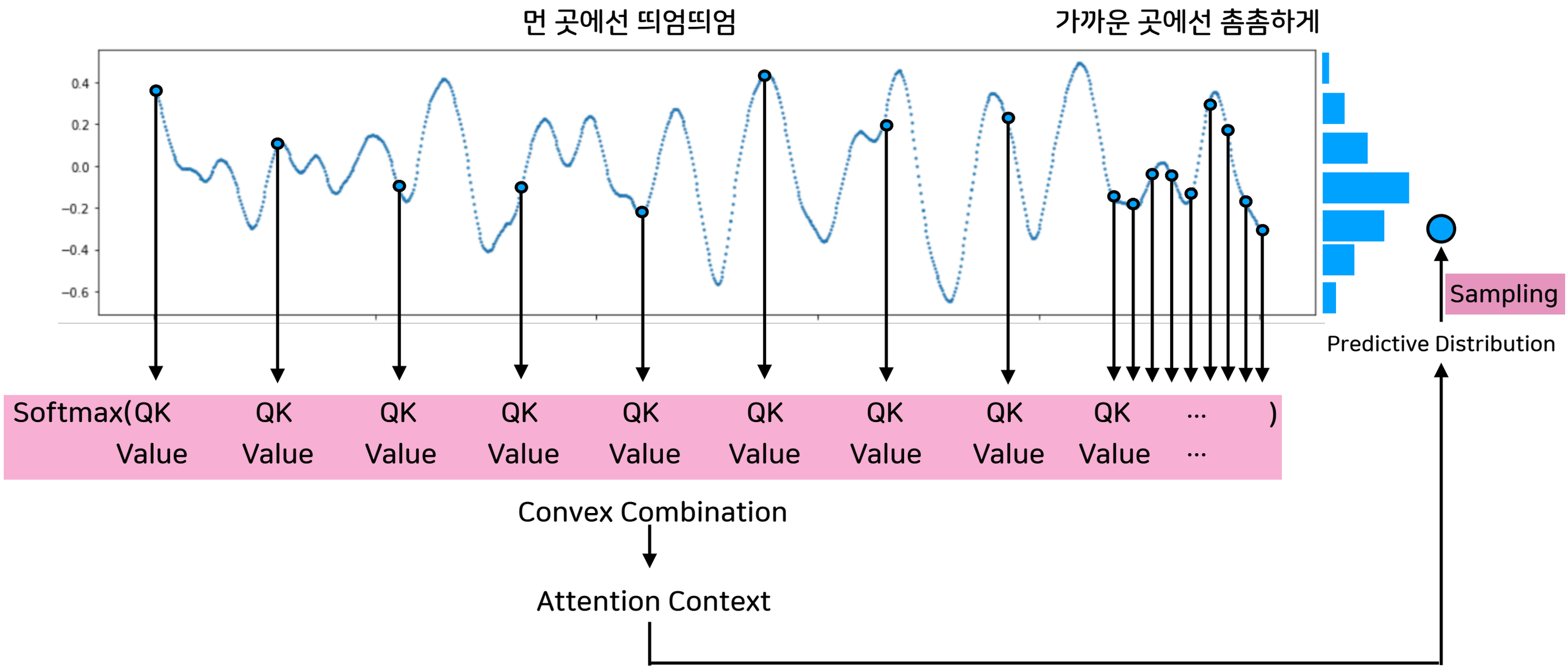
2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



2. Autoregressive Models

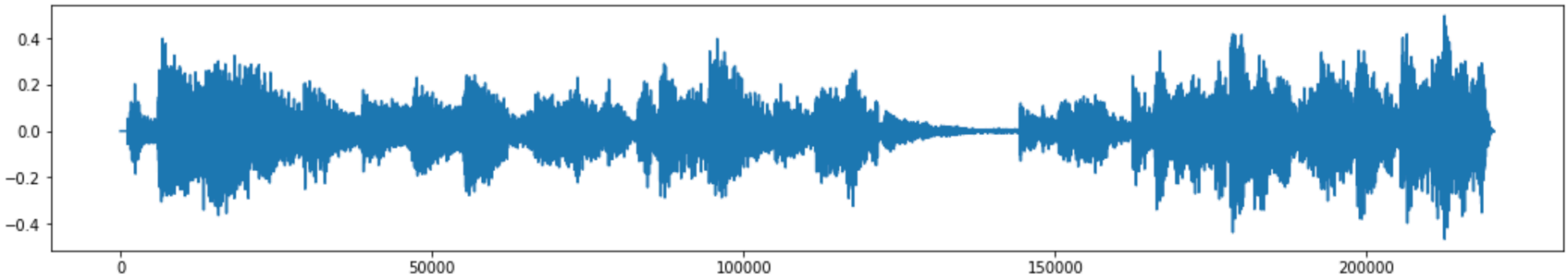
2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers



2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers

Generated Music 1

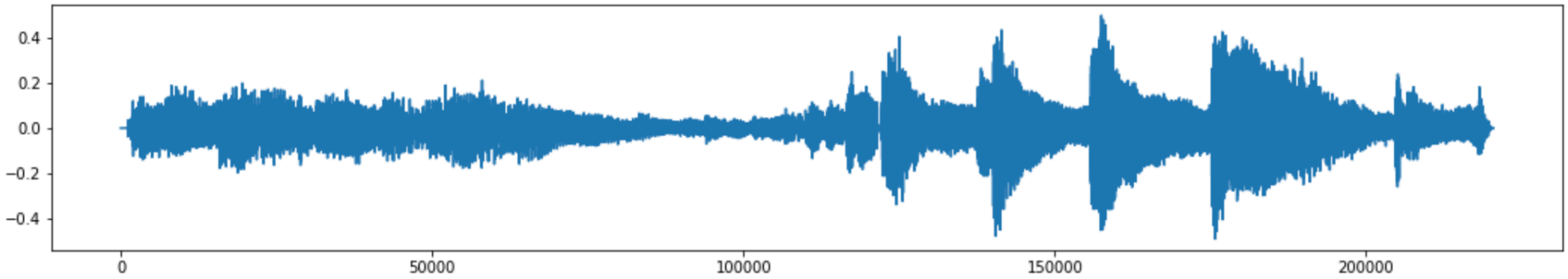


Audio Reference : https://openai.com/blog/sparse-transformer/

2. Autoregressive Models

2.4 Sparse Transformer : Generating Long Sequences with Sparse Transformers

Generated Music 2

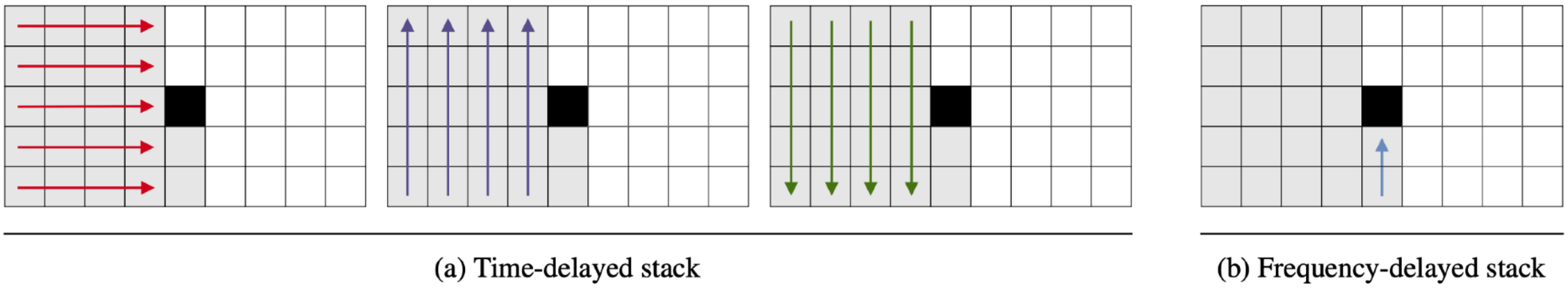


Audio Reference : https://openai.com/blog/sparse-transformer/

Melnet

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain



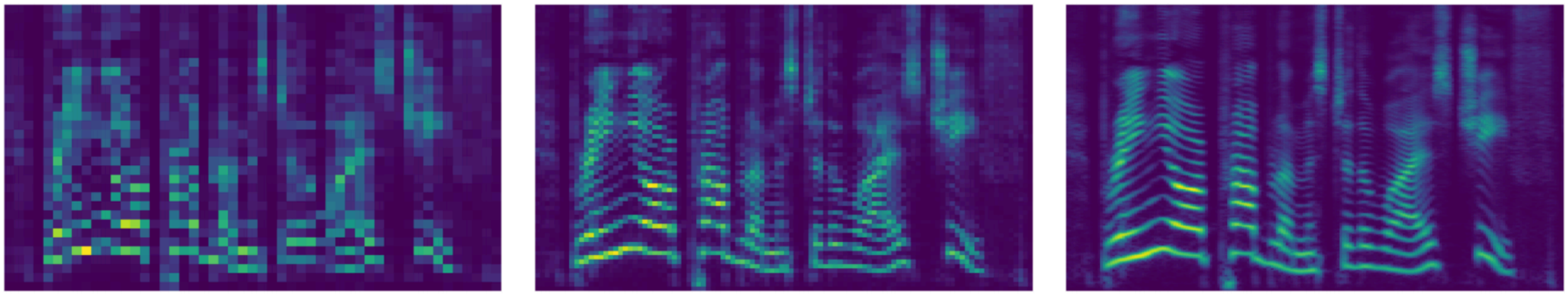
(a) Time-delayed stack

(b) Frequency-delayed stack

Figure credit: Sean Vasquez and Mike Lewis.
MelNet: A generative model for audio in the frequency domain.

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain



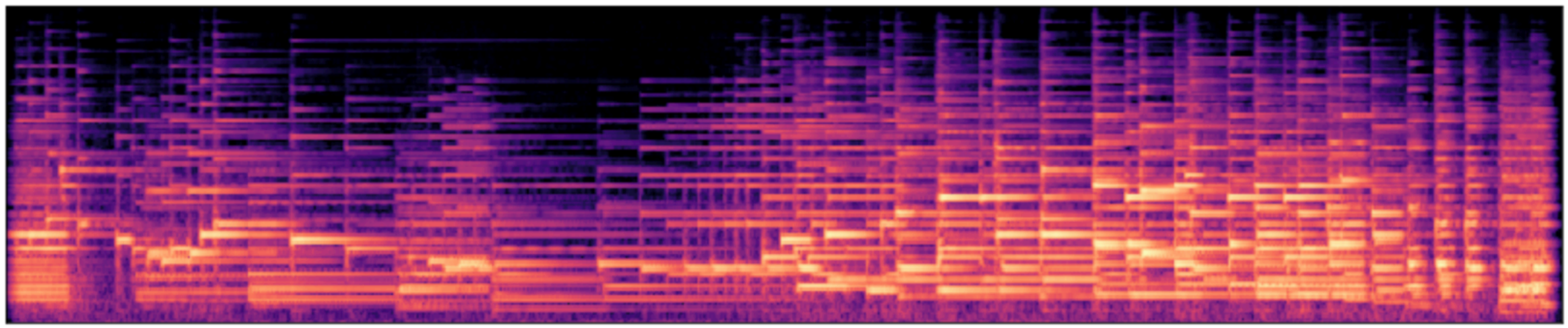
(a) Tier 1 (32×50) (b) Tiers 1–3 (64×100) (c) Tiers 1–6 (256×200)

Figure credit: Sean Vasquez and Mike Lewis.
MelNet: A generative model for audio in the frequency domain.

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain

Generated Music. 1

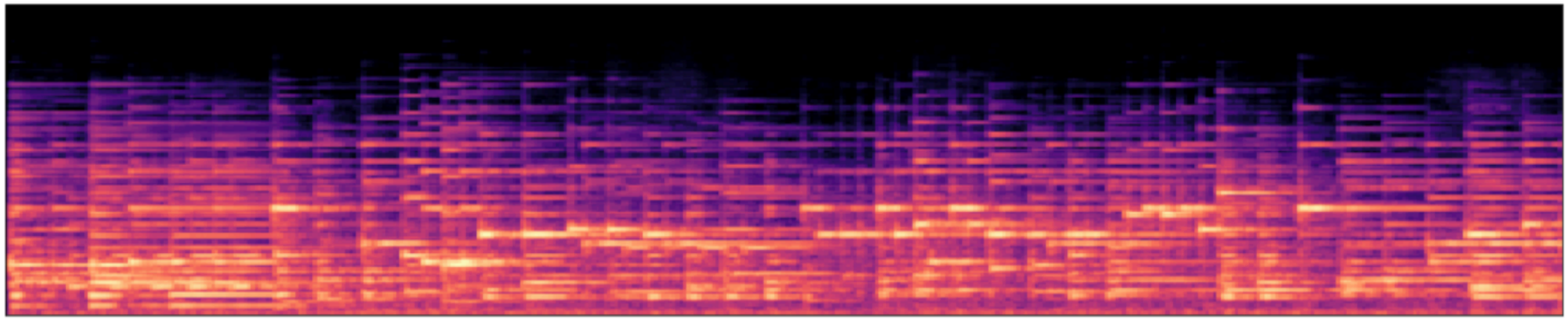


Audio Reference : https://sjvasquez.github.io/blog/melnet/

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain

Generated Music. 2

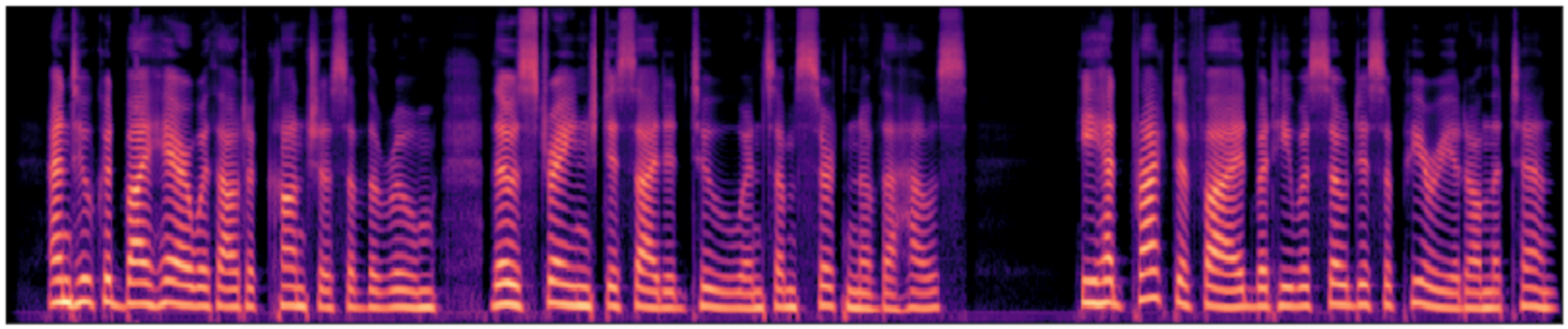


Audio Reference : https://sjvasquez.github.io/blog/melnet/

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain

Generated Speech. 1

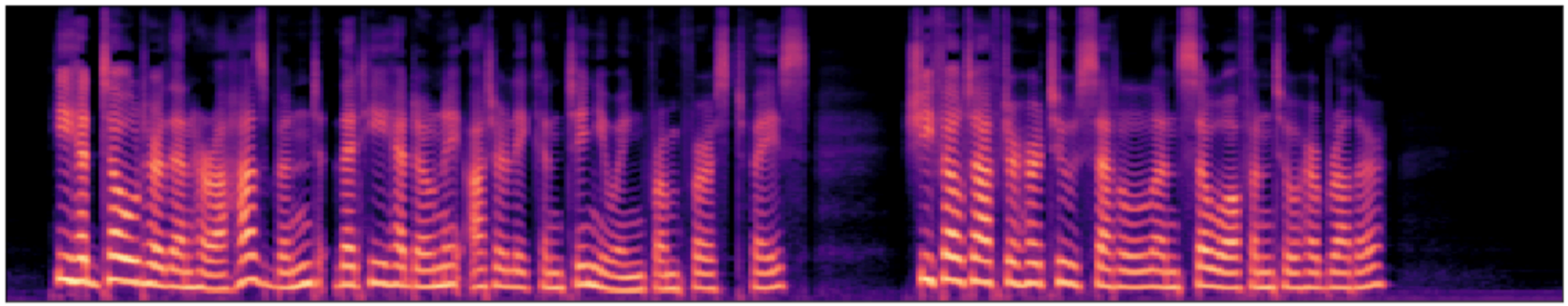


Audio Reference : https://sjvasquez.github.io/blog/melnet/

2. Autoregressive Models

2.3 Melnet : A Generative Model for Audio in the Frequency Domain

Generated Speech. 2



Audio Reference : https://sjvasquez.github.io/blog/melnet/

Flow-Based Models

3. Flow-Based Models

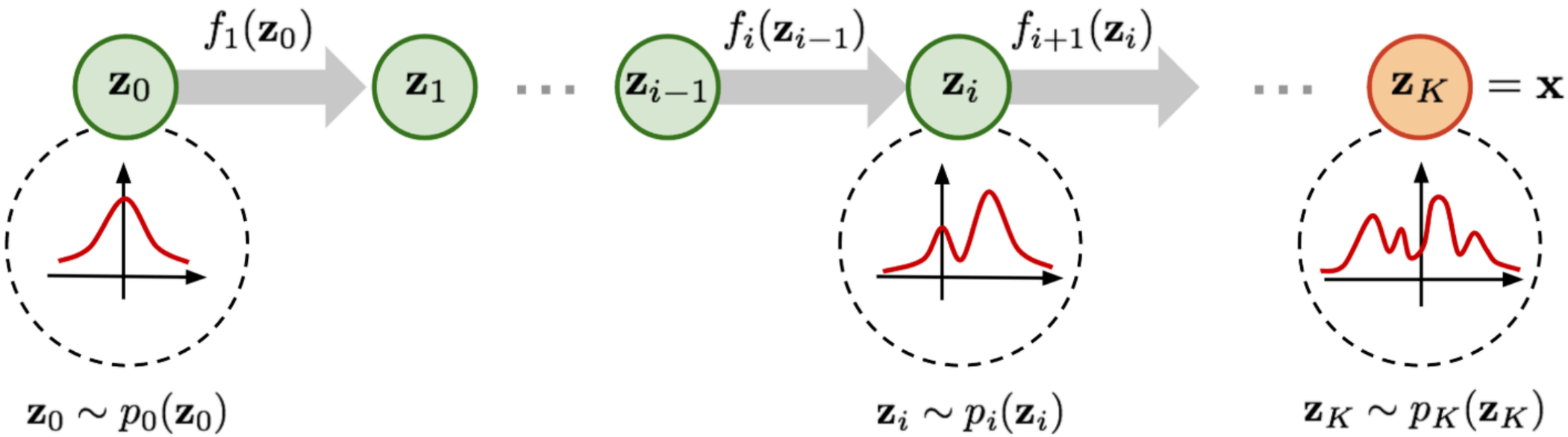
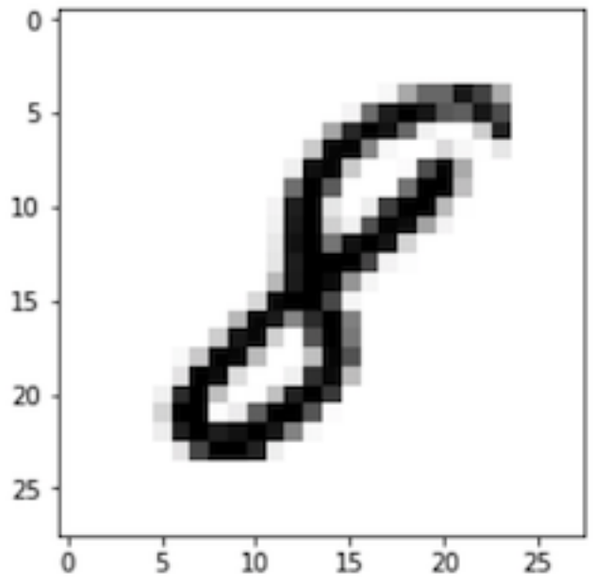


Figure credit: Lillian Weng.
lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

NICE & RealNVP & Glow

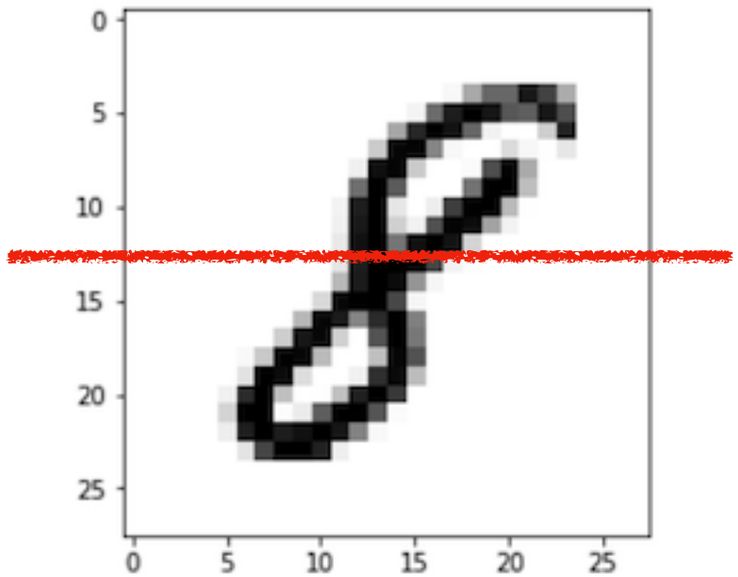
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



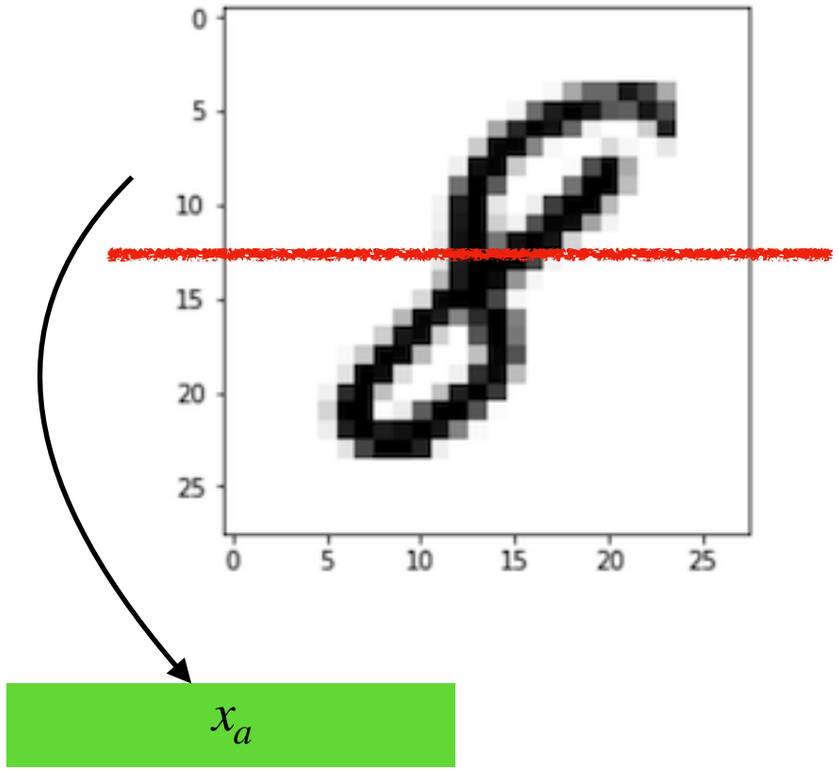
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



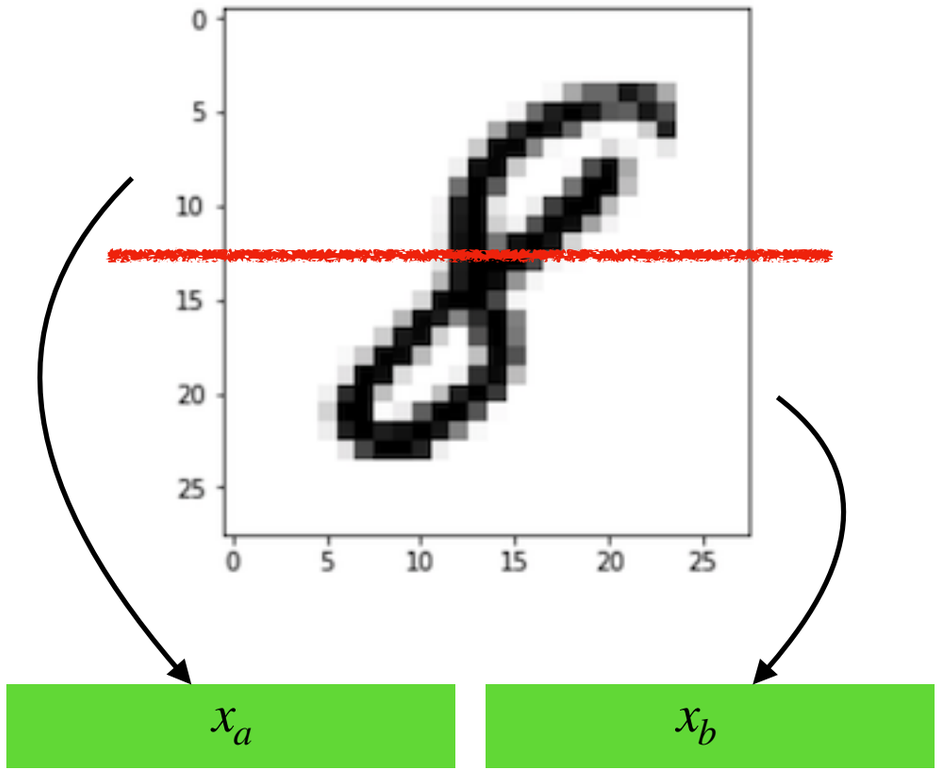
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



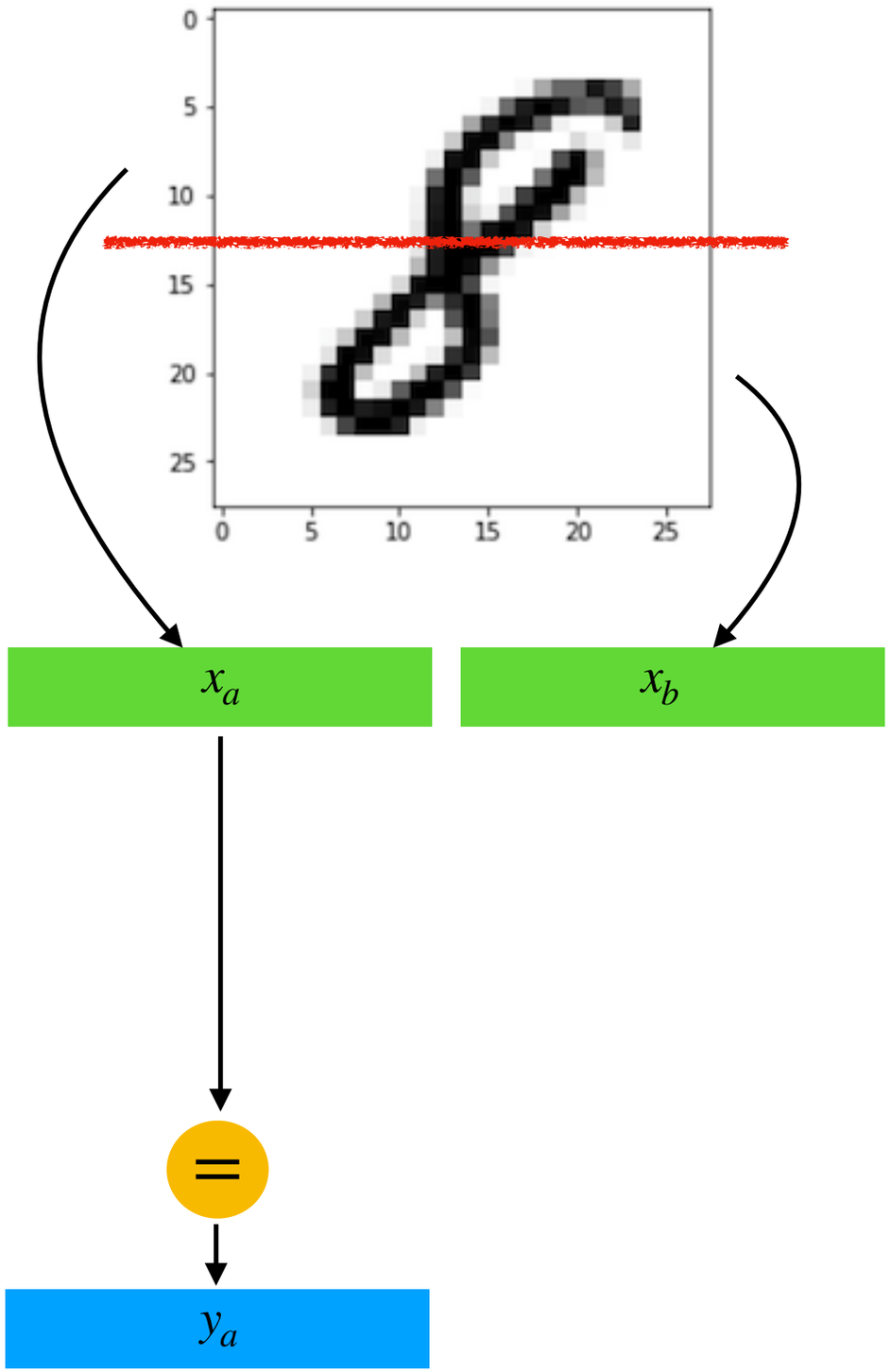
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



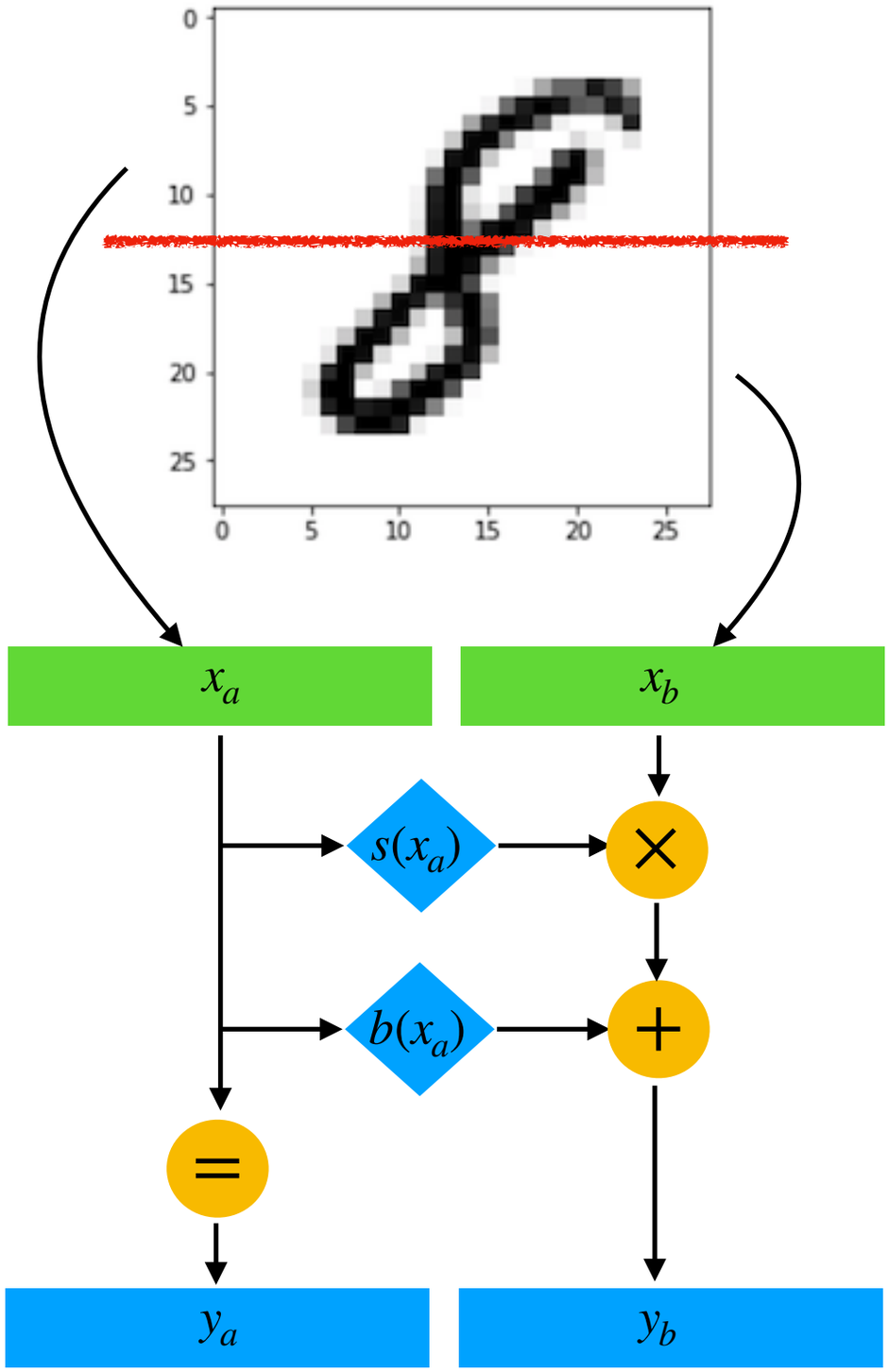
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



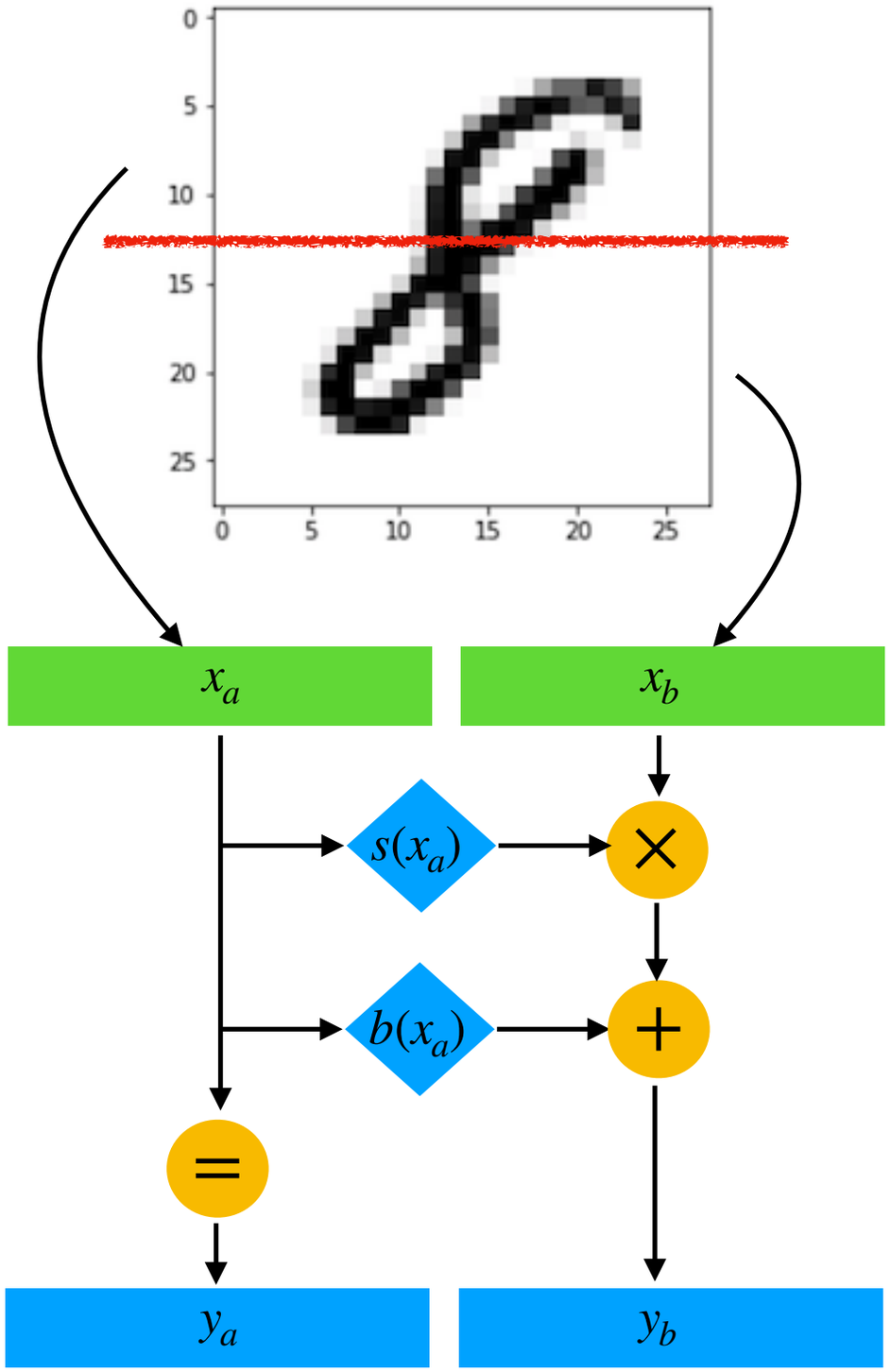
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

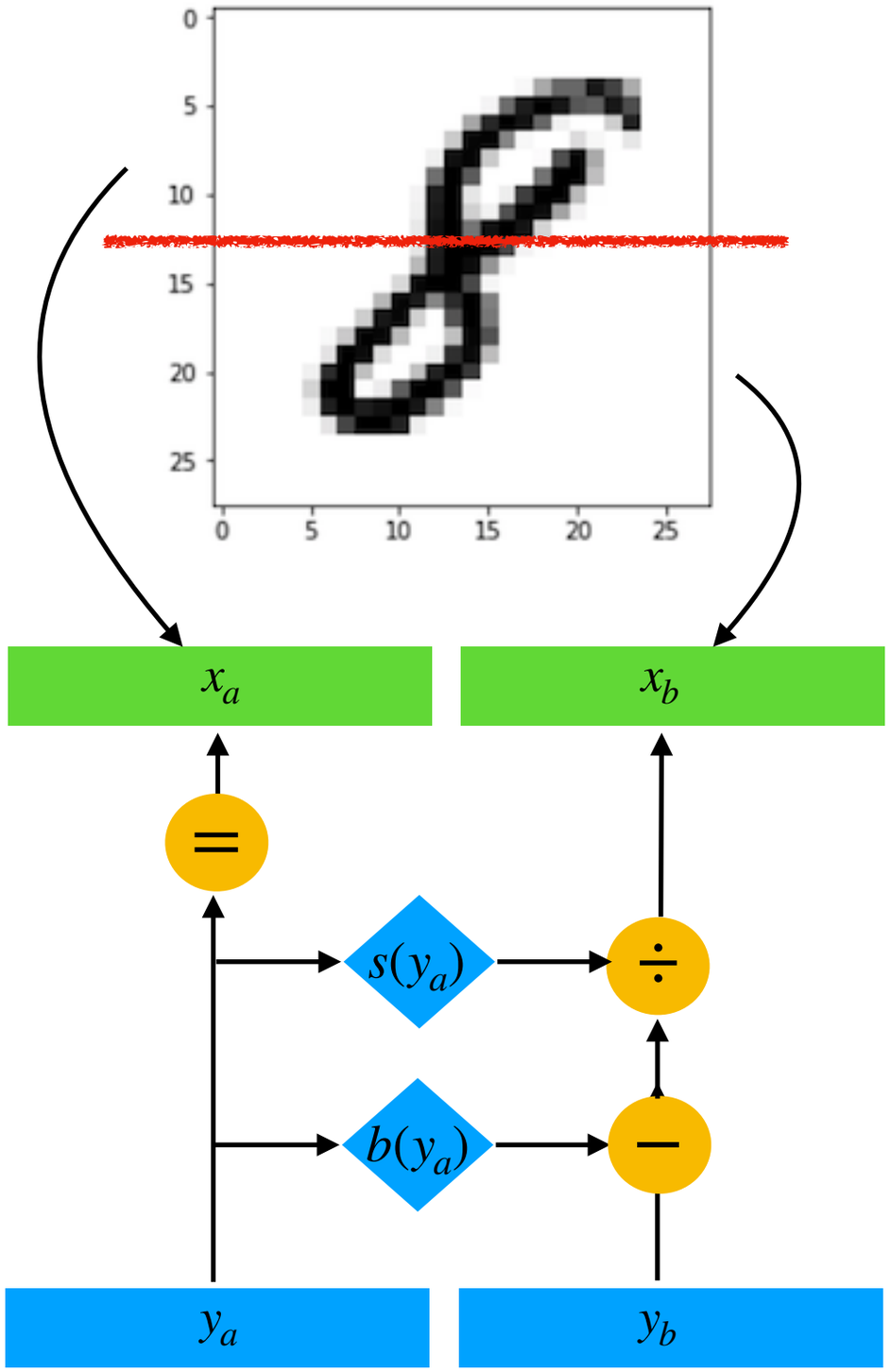


Coupling Layer (Forward)

$$y_a = x_a$$
$$y_b = x_b \times s(x_a) + b(x_a)$$

3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions



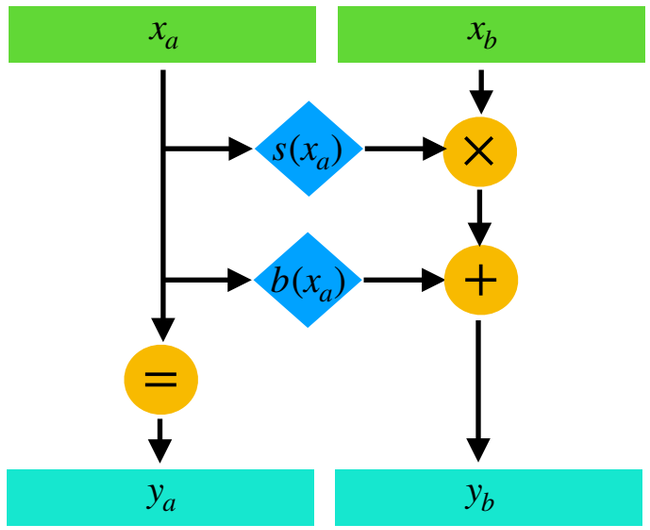
Coupling Layer (Backward)

$$x_a = y_a$$
$$x_b = (y_b - b(y_a)) \div s(x_a)$$

3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

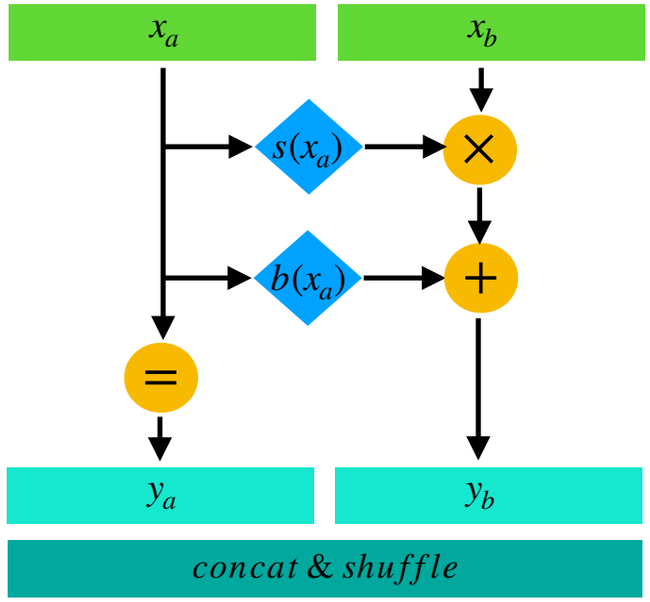
Training Time



3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

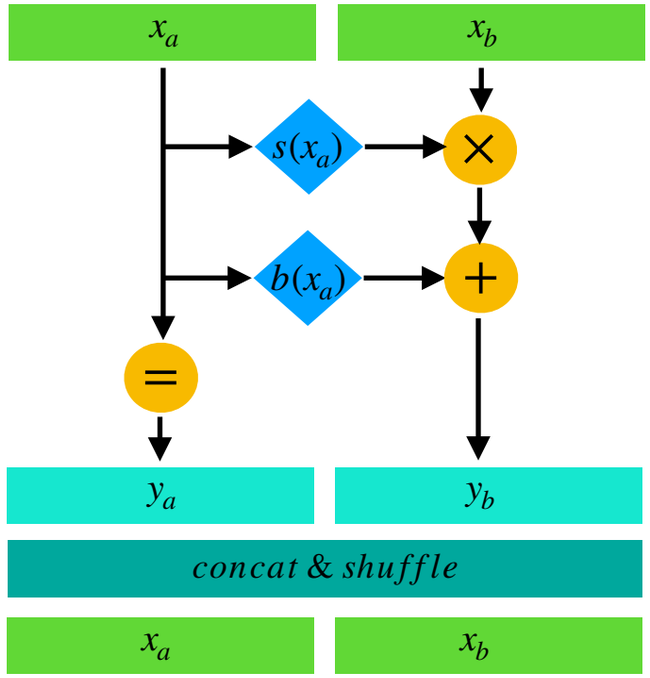
Training Time



3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

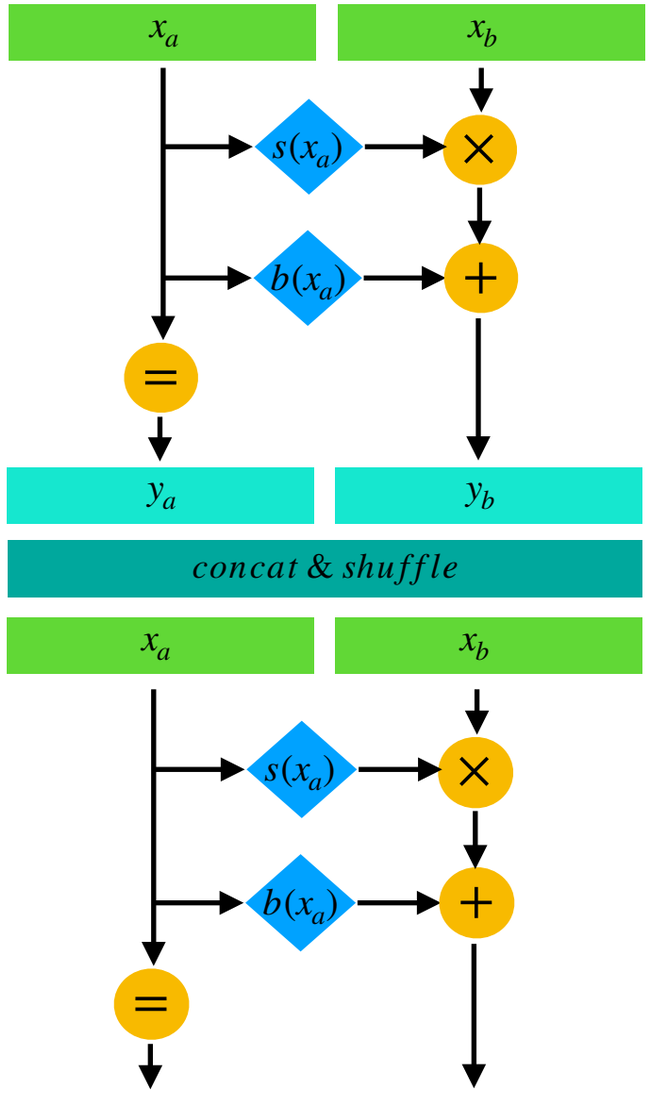
Training Time



3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

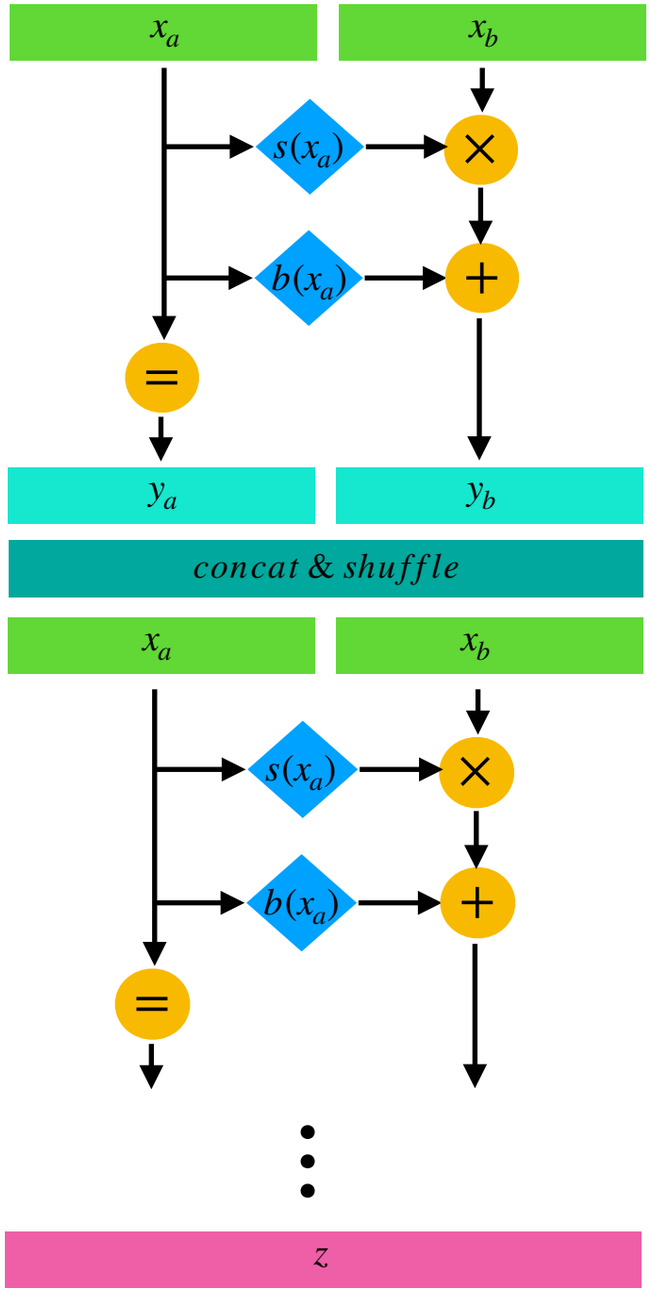
Training Time



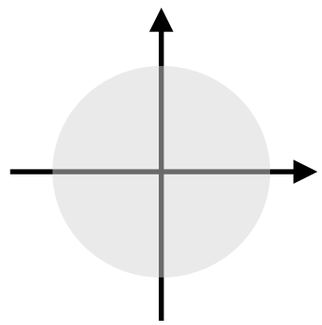
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

Training Time



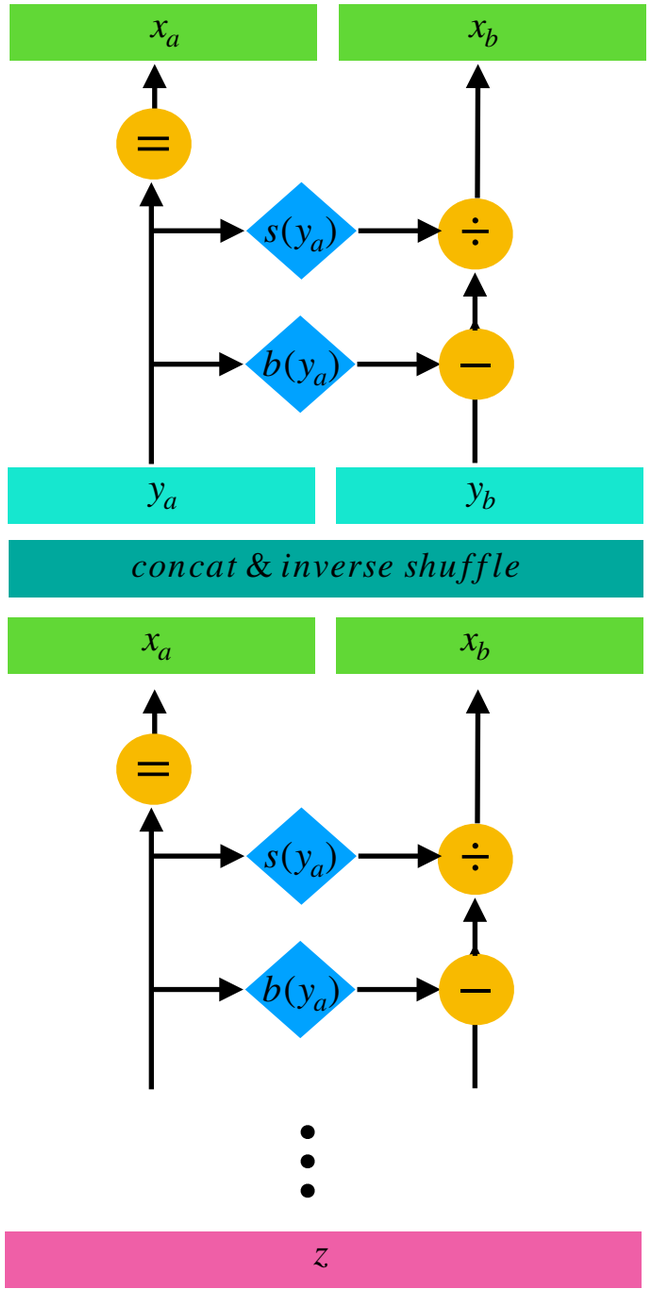
Maximum Likelihood
In Isotropic Gaussian



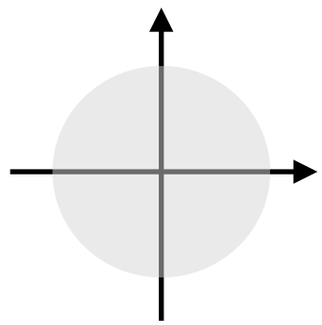
3. Flow-Based Models

3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

Inference Time



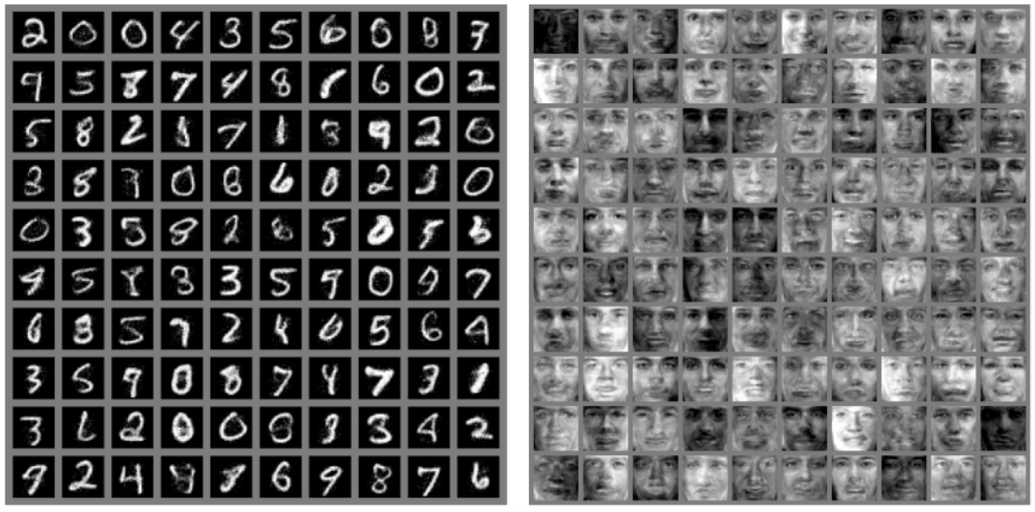
Sampling



3. Flow-Based Models

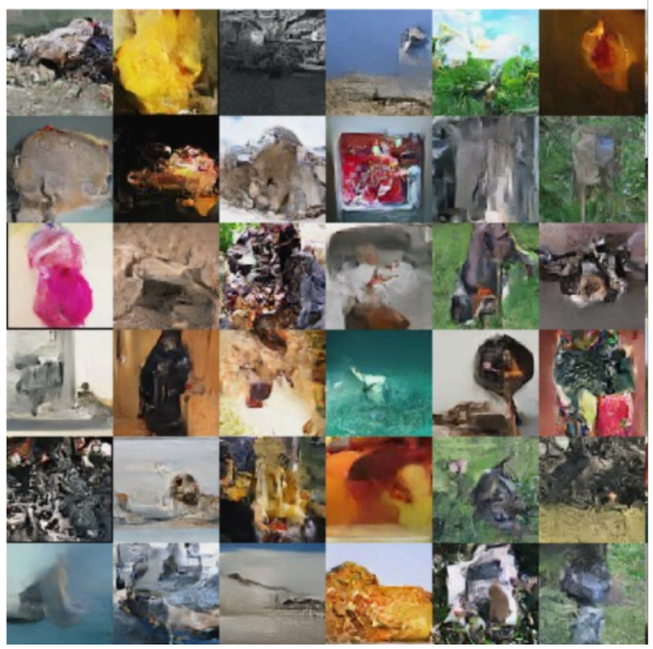
3.1 NICE : Non-linear Independent Components Estimation
RealNVP : Density estimation using Real NVP
Glow : Generative Flow with Invertible 1×1 Convolutions

NICE



Glow

RealNVP



FlowSeq

3. Flow-Based Models

3.3 FlowSeq

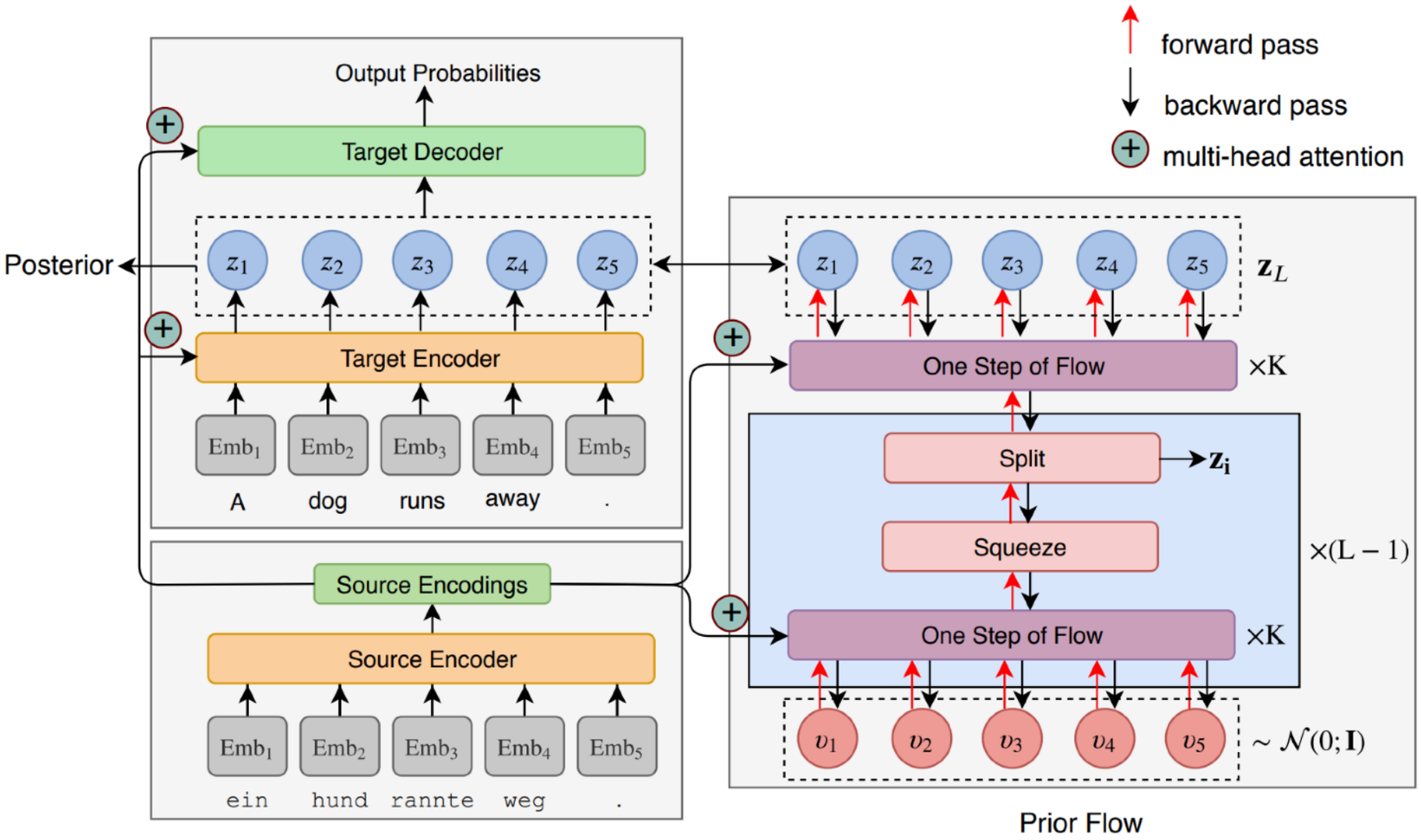
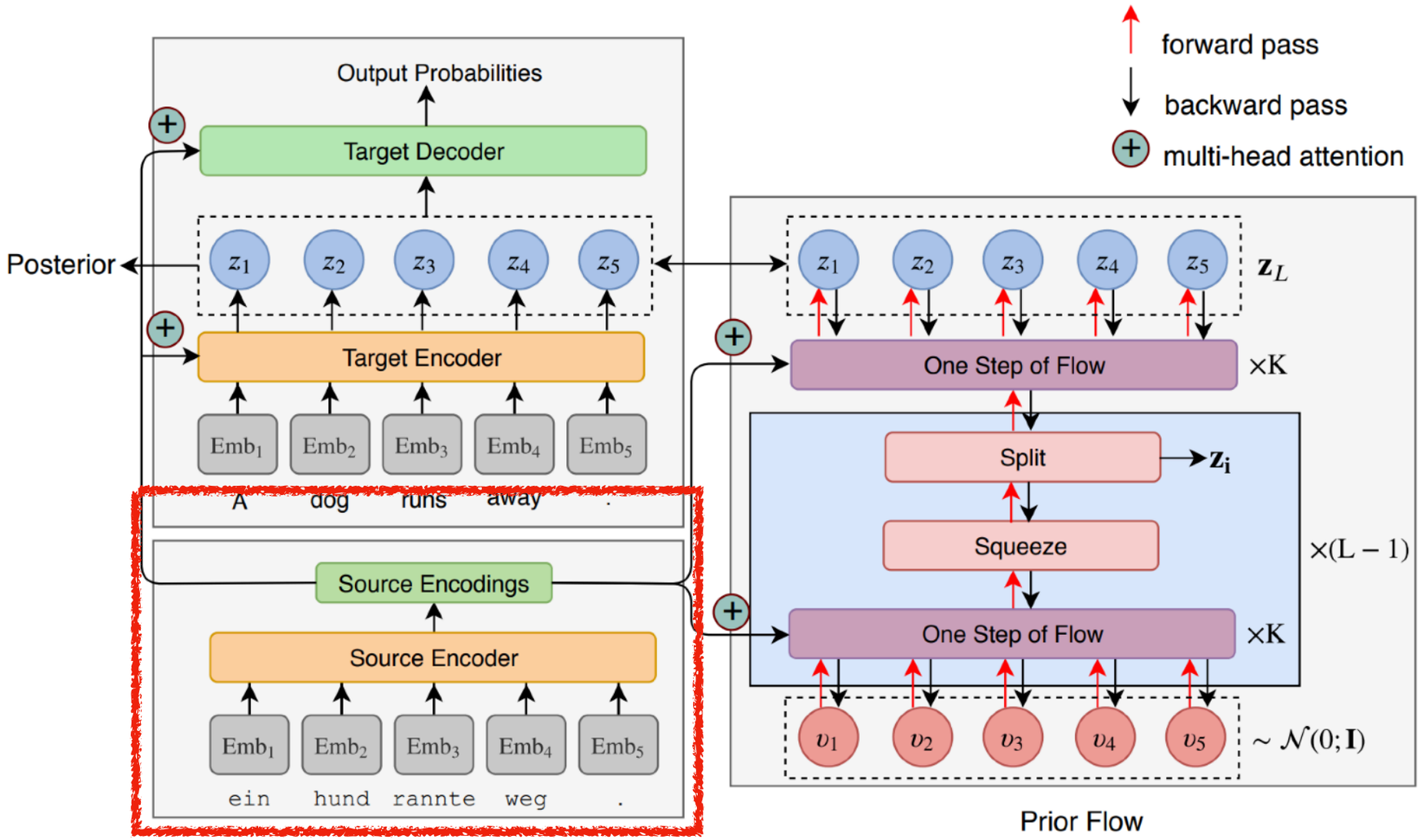


Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Training Time



1. Encode the source

Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

2. VAE to the target **Training Time**

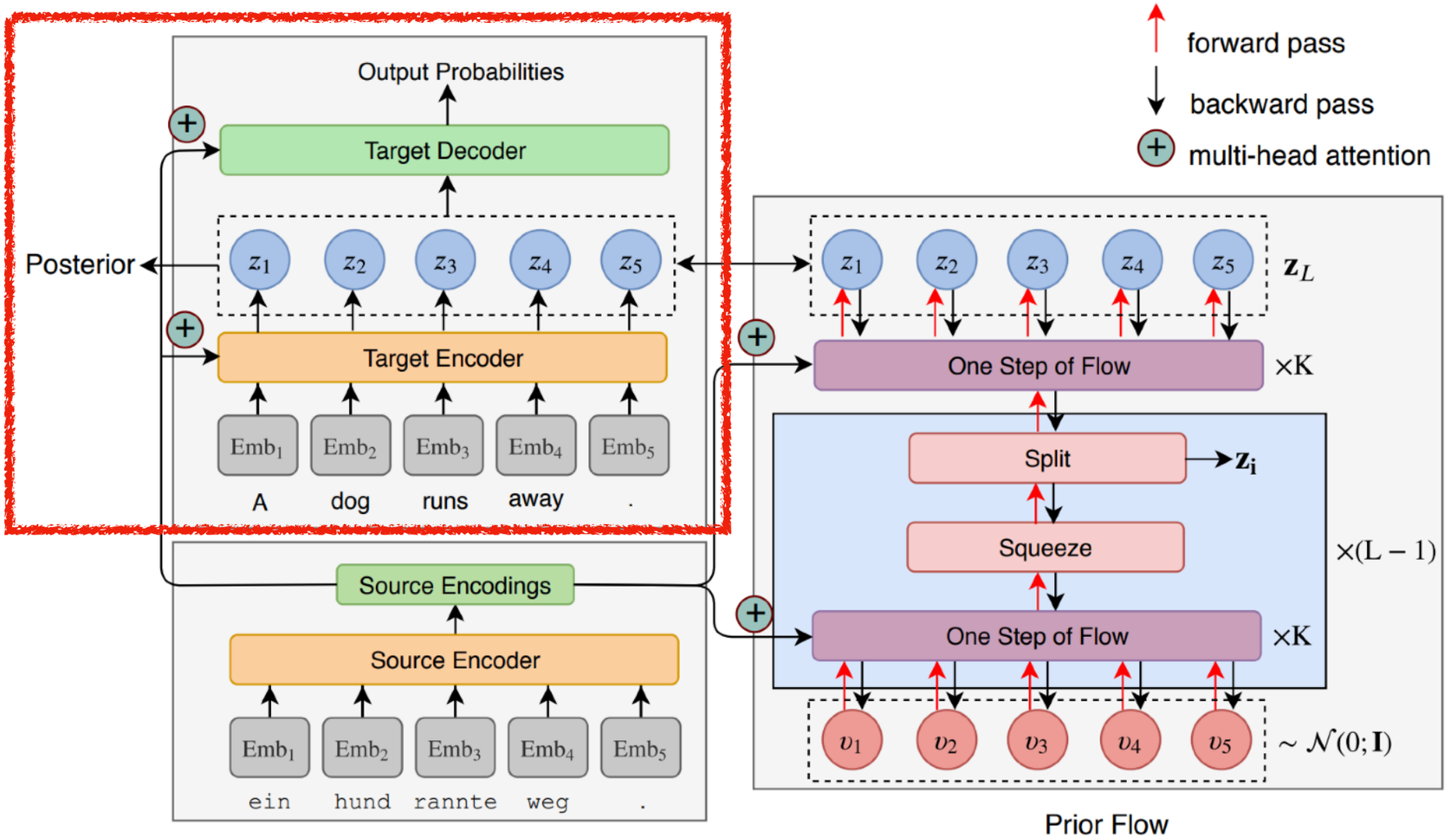
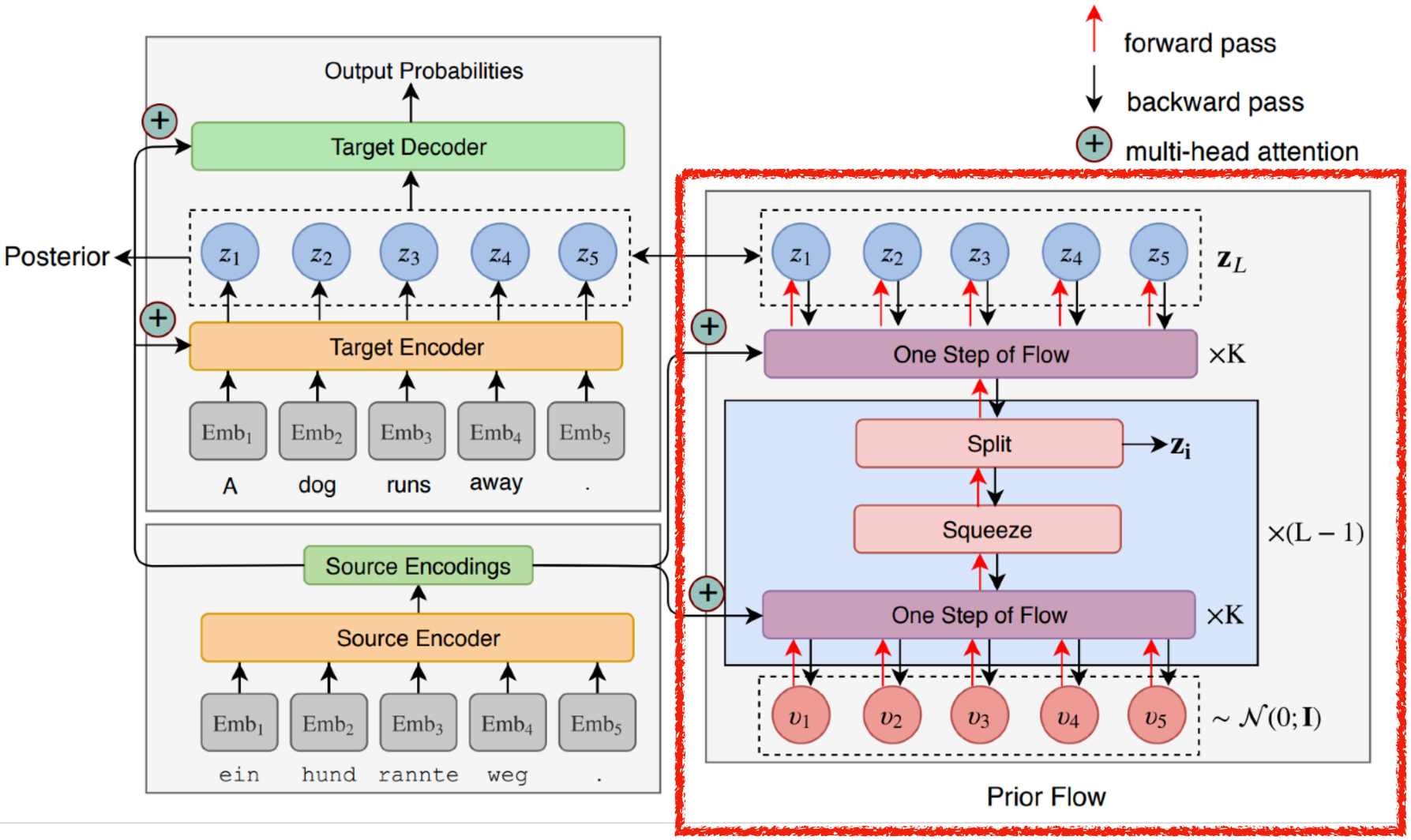


Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Training Time



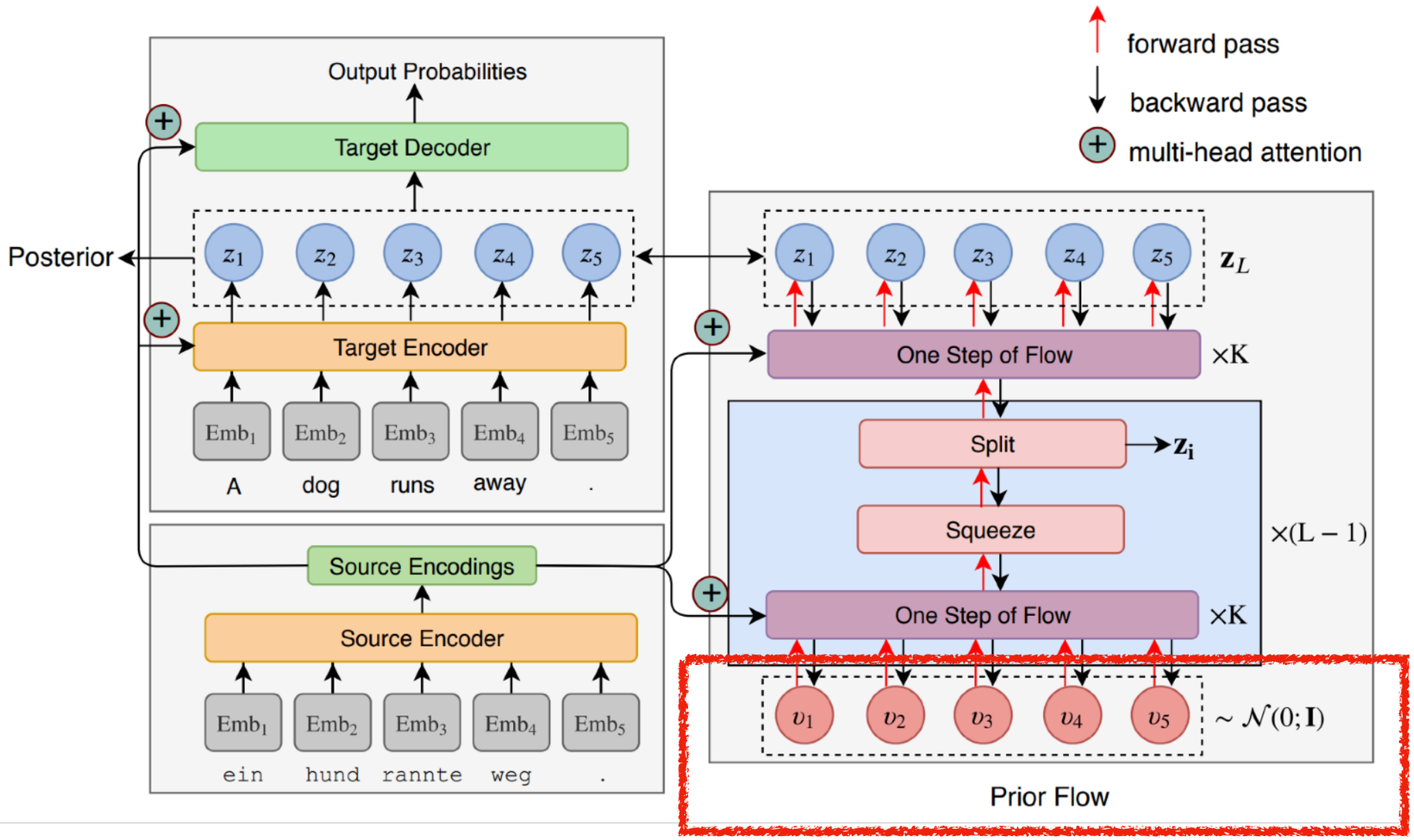
3. Normalize the posterior

Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Inference Time



1. Sample from the prior

Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Inference Time

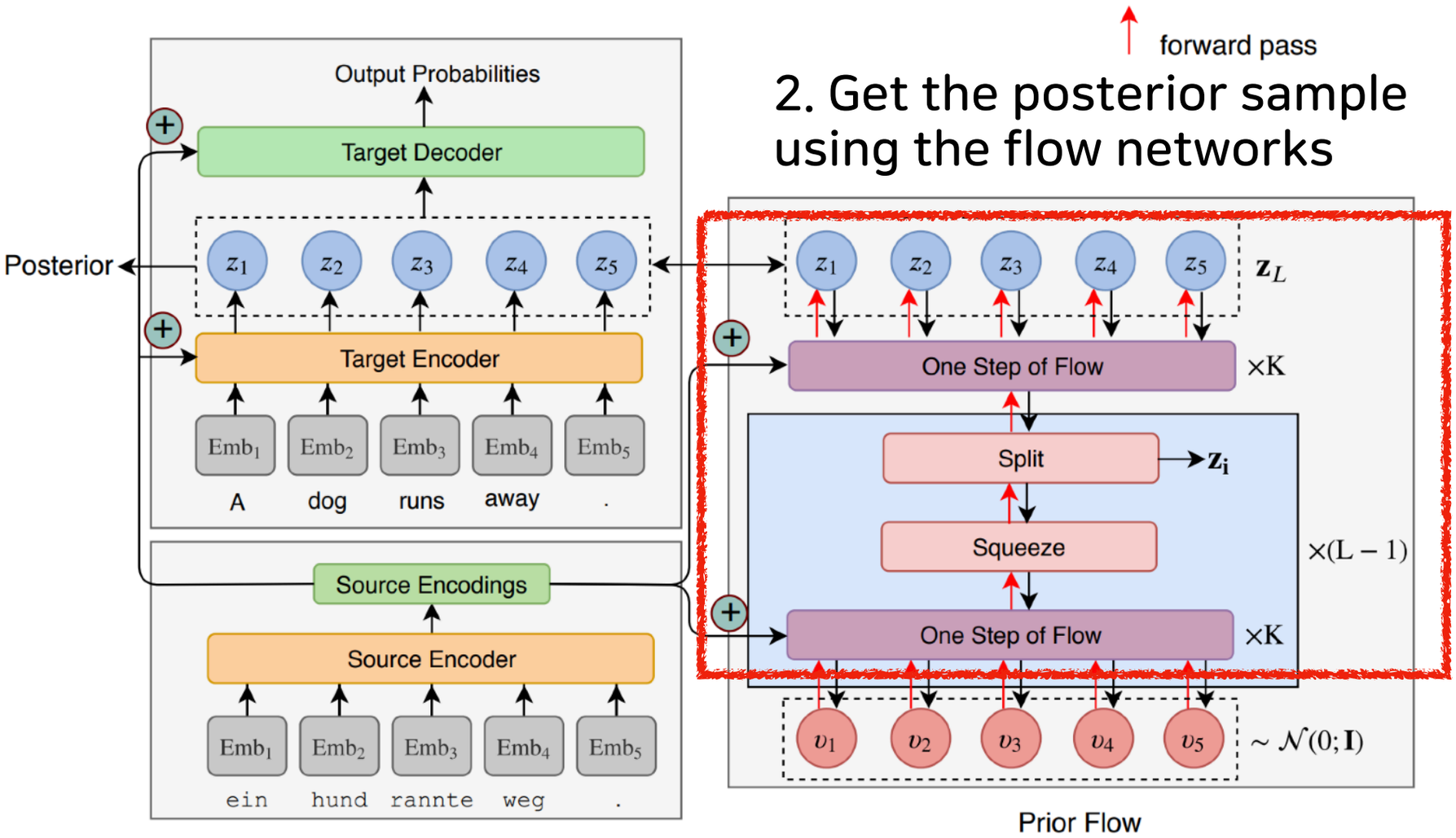
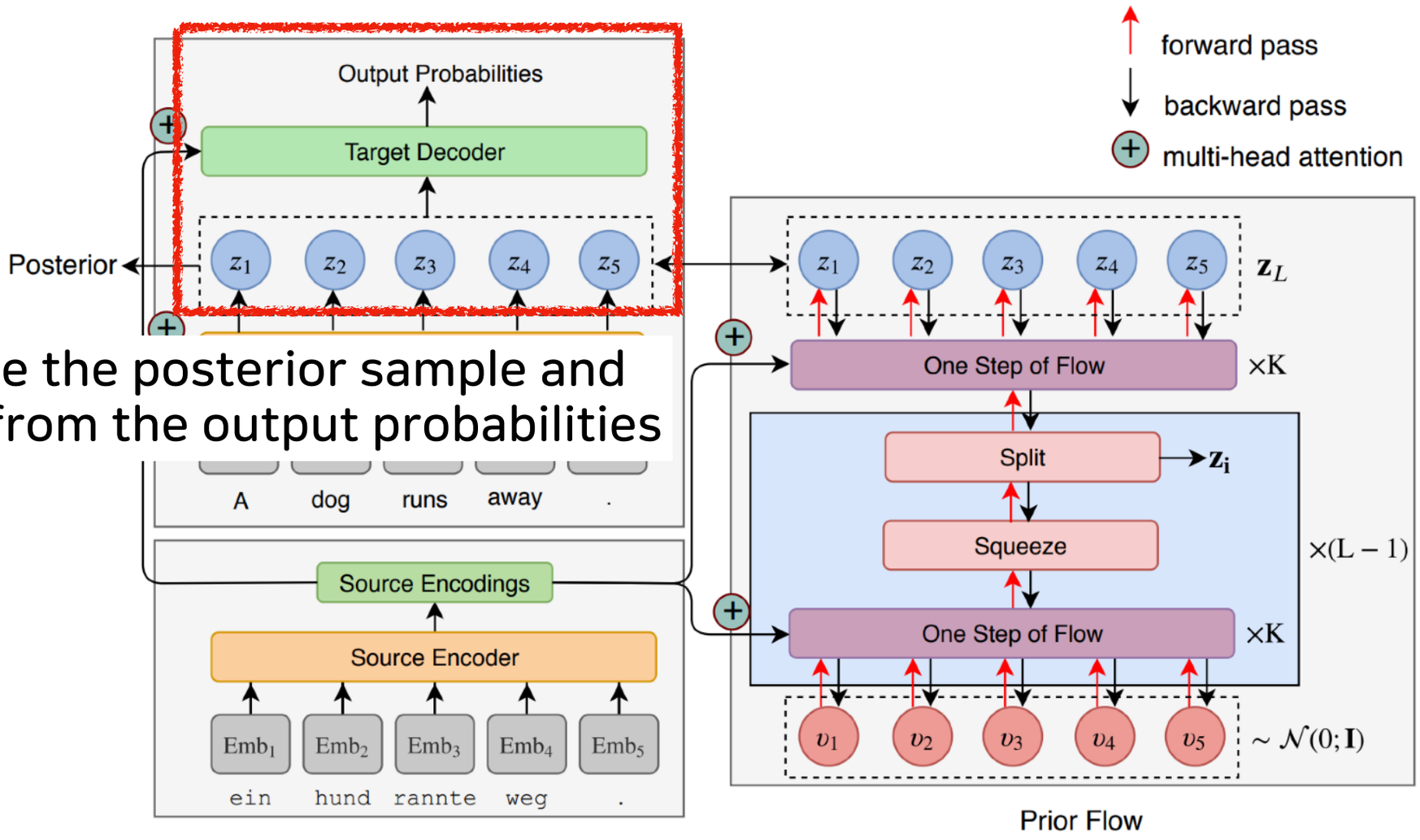


Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Inference Time



3. Decode the posterior sample and Sample from the output probabilities

Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

3. Flow-Based Models

3.3 FlowSeq

Source	Grundnahrungsmittel gibt es schlielich berall und jeder Supermarkt hat mittlerweile Sojamilch und andere Produkte.
Ground Truth	There are basic foodstuffs available everywhere , and every supermarket now has soya milk and other products.
Sample 1	After all, there are basic foods everywhere and every supermarket now has soya amch and other products.
Sample 2	After all, the food are available everywhere everywhere and every supermarket has soya milk and other products.
Sample 3	After all, basic foods exist everywhere and every supermarket has now had soy milk and other products.
Source	Es kann nicht erklären, weshalb die National Security Agency Daten ber das Privatleben von Amerikanern sammelt und warum Whistleblower bestraft werden, die staatliches Fehlverhalten offenlegen.
Ground Truth	And, most recently, it cannot excuse the failure to design a simple website more than three years since the Affordable Care Act was signed into law.
Sample 1	And recently, it cannot apologise for the inability to design a simple website in the more than three years since the adoption of Affordable Care Act.
Sample 2	And recently, it cannot excuse the inability to design a simple website in more than three years since the adoption of Affordable Care Act.
Sample 3	Recently, it cannot excuse the inability to design a simple website in more than three years since the Affordable Care Act has passed.

Figure credit: Xuezhe Ma et. al., FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow

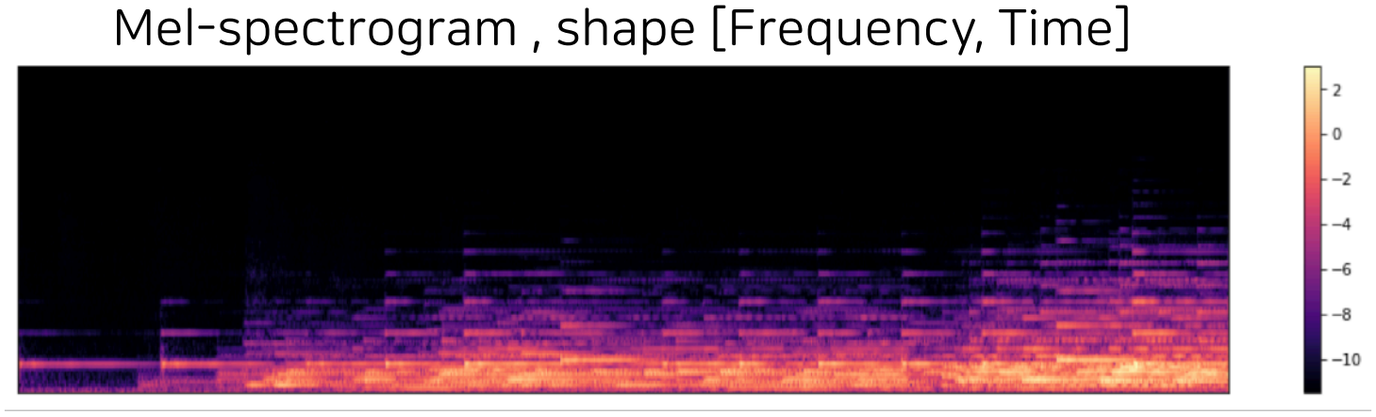
MelFlow

영감이란 교향곡 전체가 한꺼번에 떠오르는 것이다.

Arnold Schoenberg, ‘Style and Idea’

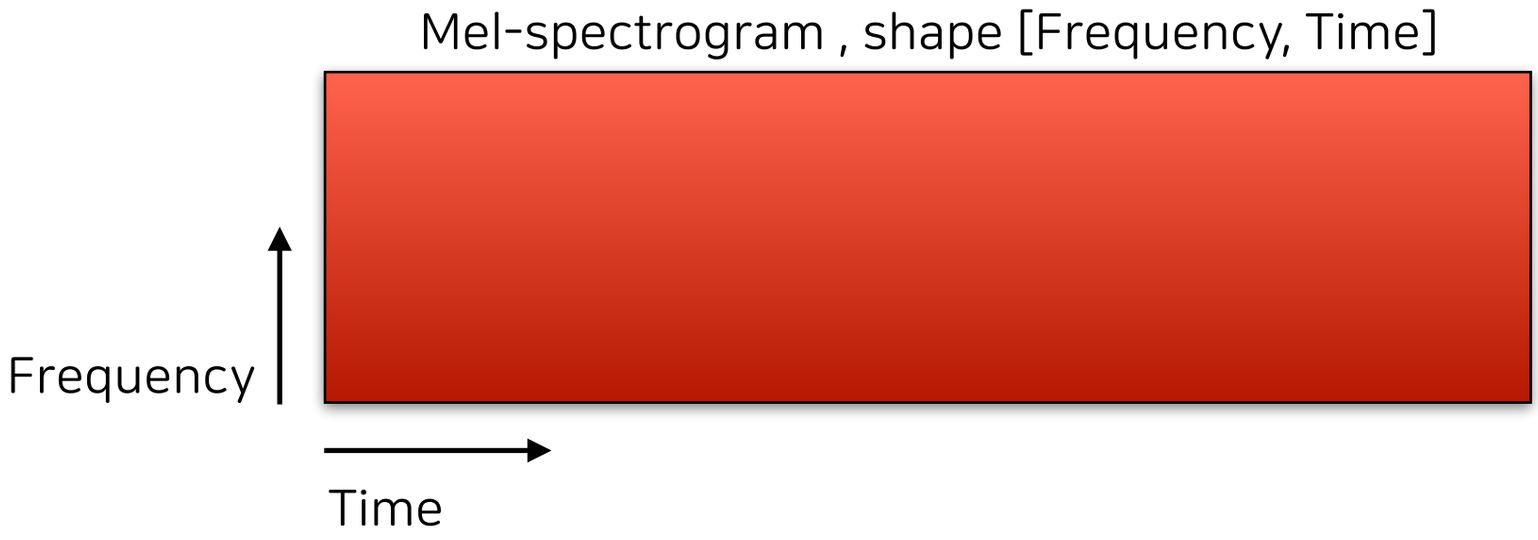
3. Flow-Based Models

3.4 MelFlow



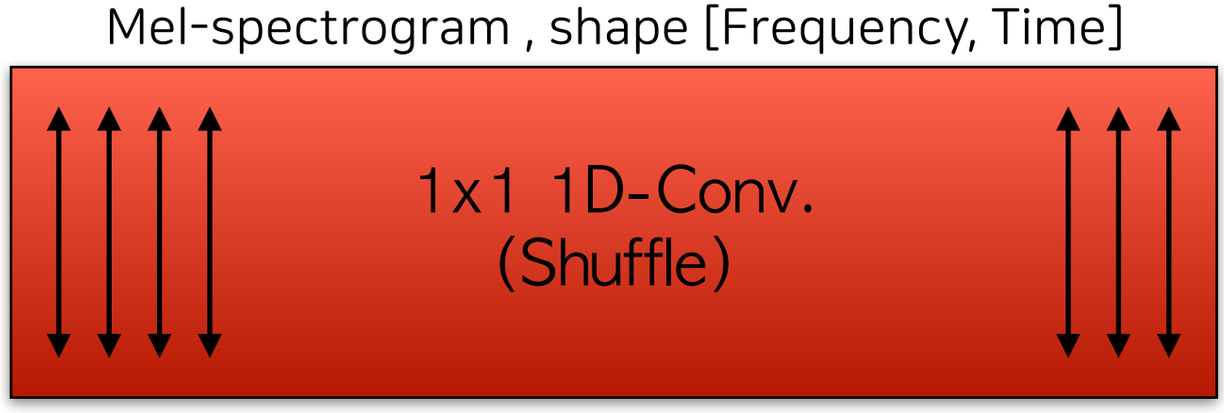
3. Flow-Based Models

3.4 MelFlow



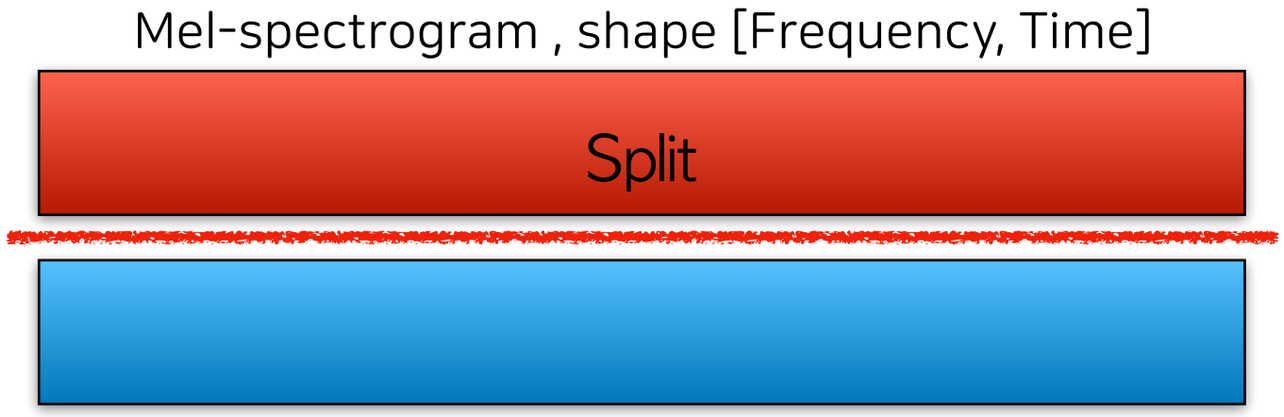
3. Flow-Based Models

3.4 MelFlow



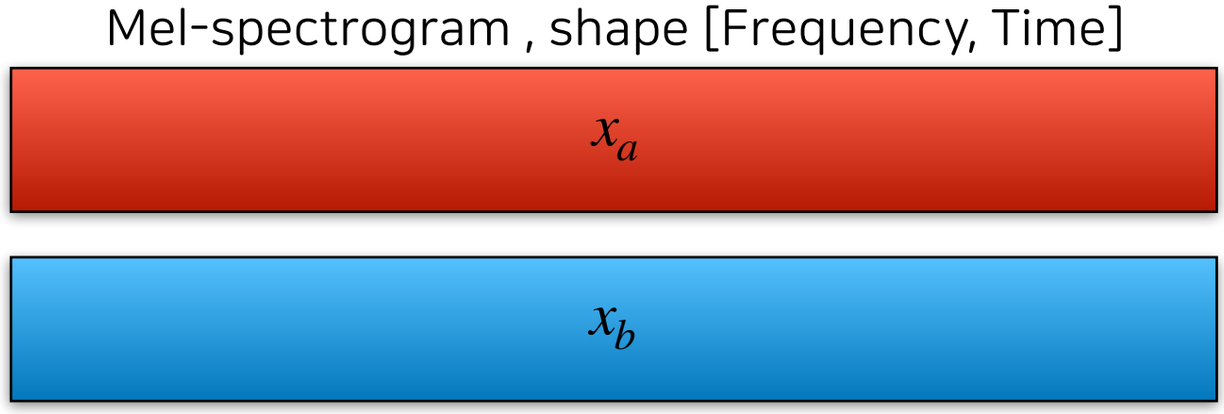
3. Flow-Based Models

3.4 MelFlow



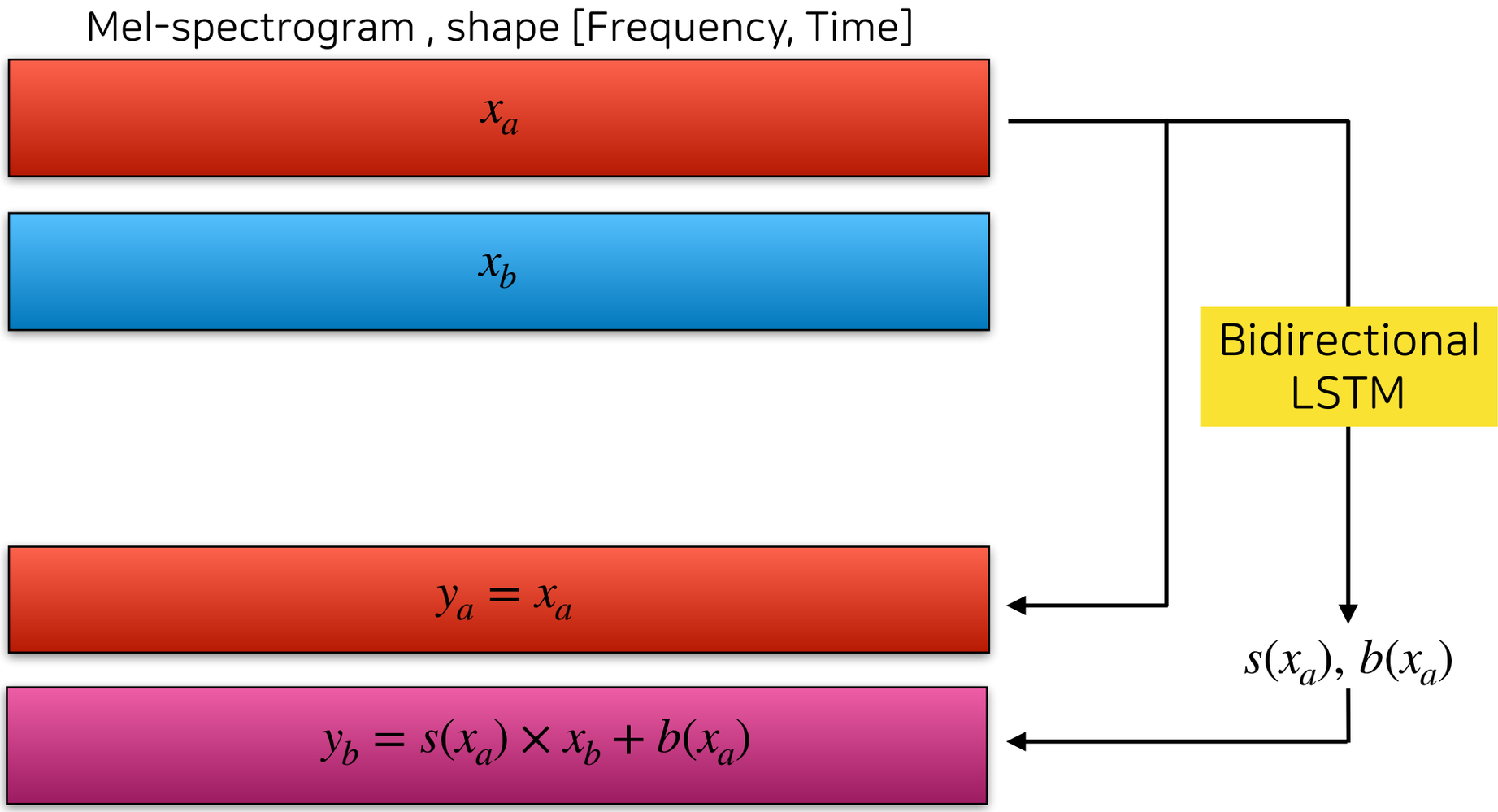
3. Flow-Based Models

3.4 MelFlow



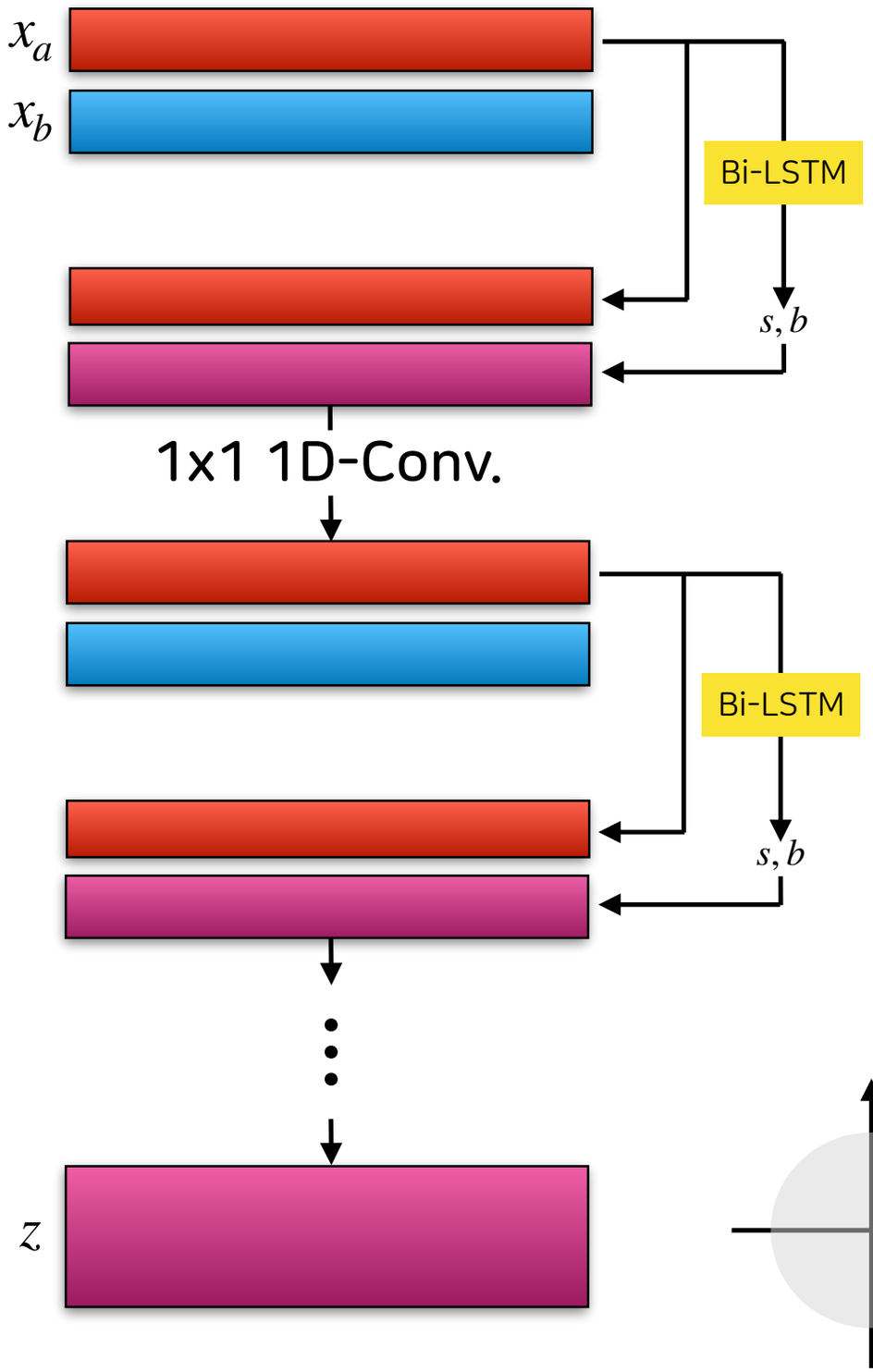
3. Flow-Based Models

3.4 MelFlow

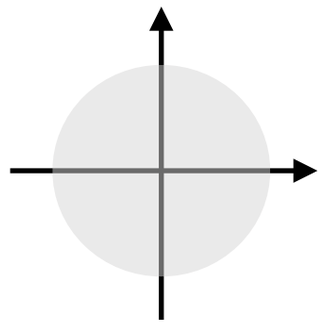


3. Flow-Based Models

3.4 MelFlow

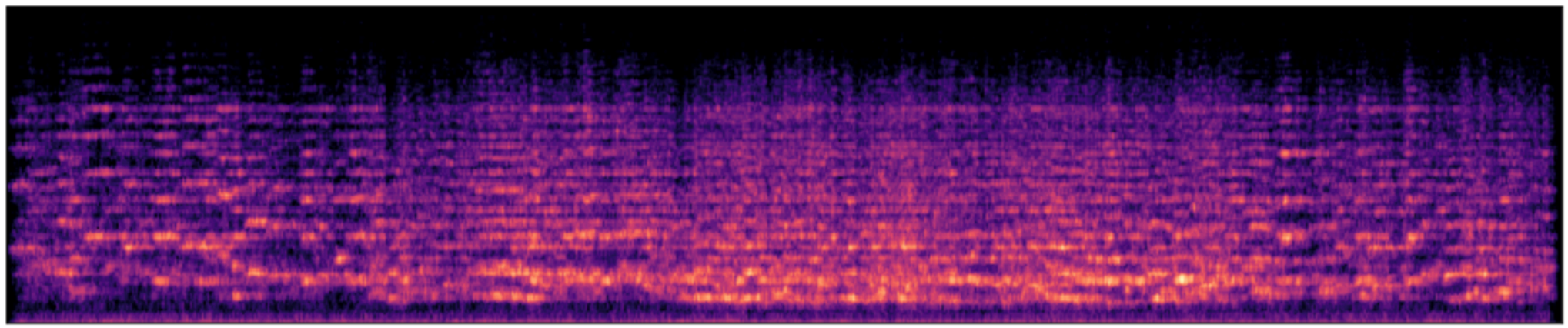


Maximum Likelihood
In Isotropic Gaussian



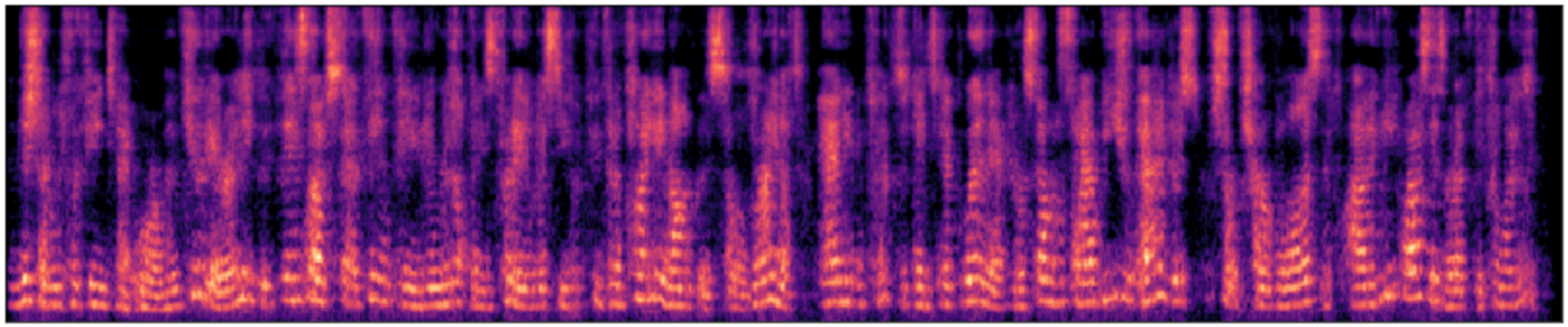
3. Flow-Based Models
3.4 MelFlow

Generated Music Using Bach Violin Partitas and Sonatas



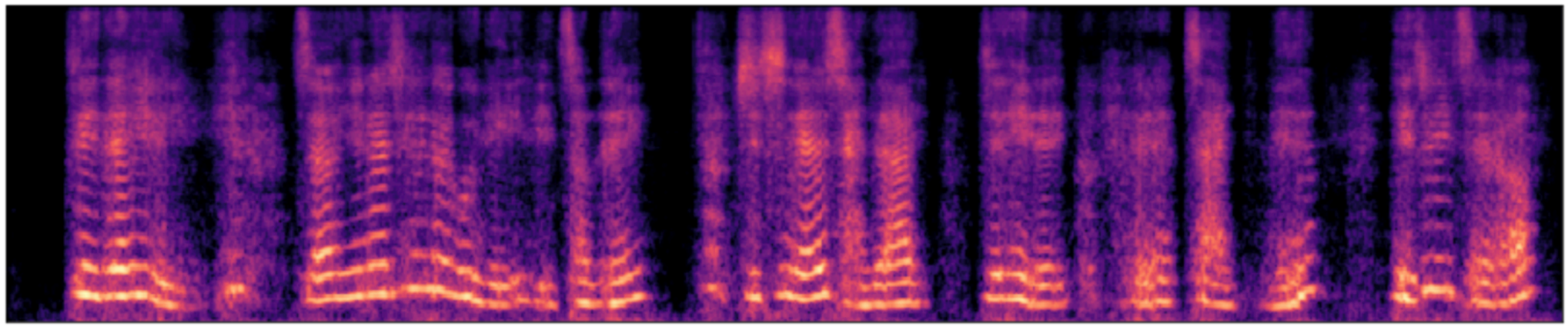
3. Flow-Based Models
3.4 MelFlow

Generated Speech Using LJSpeech



3. Flow-Based Models
3.4 MelFlow

Generated Speech Using Korean Single Speaker Speech Dataset



요약&결론

4. 요약&결론

1. Wave 데이터

장점: 구하기 쉽고, 데이터량 풍부
단점: MIDI에 비해 복잡하고, 다루기 어려움

4. 요약&결론

1. Wave 데이터

장점: 구하기 쉽고, 데이터량 풍부
단점: MIDI에 비해 복잡하고, 다루기 어려움

2. Spectrogram, Mel-spectrogram

Wave를 주파수 영역에서 분석하여 인간의 인지방식과 적합

4. 요약&결론

1. Wave 데이터

   장점: 구하기 쉽고, 데이터량 풍부
   단점: MIDI에 비해 복잡하고, 다루기 어려움

2. Spectrogram, Mel-spectrogram

   Wave를 주파수 영역에서 분석하여 인간의 인지방식과 적합

3. Autoregressive Models

   Wavenet: Wave데이터 이용
   VQ-VAE: Z variable 도입
   Sparse Transformer: Transformer 접목
   Melnet: Mel-spectrogram 이용

4. 요약&결론

1. Wave 데이터

장점: 구하기 쉽고, 데이터량 풍부
단점: MIDI에 비해 복잡하고, 다루기 어려움

2. Spectrogram, Mel-spectrogram

Wave를 주파수 영역에서 분석하여 인간의 인지방식과 적합

3. Autoregressive Models

Wavenet: Wave데이터 이용
VQ-VAE: Z variable 도입
Sparse Transformer: Transformer 접목
Melnet: Mel-spectrogram 이용

4. Flow-Based Models

NICE&RealNVP: Coupling Layer 도입
FlowSeq: 병렬적 번역 모델 시도
MelFlow: 병렬적 음악 생성 모델 시도

감사합니다