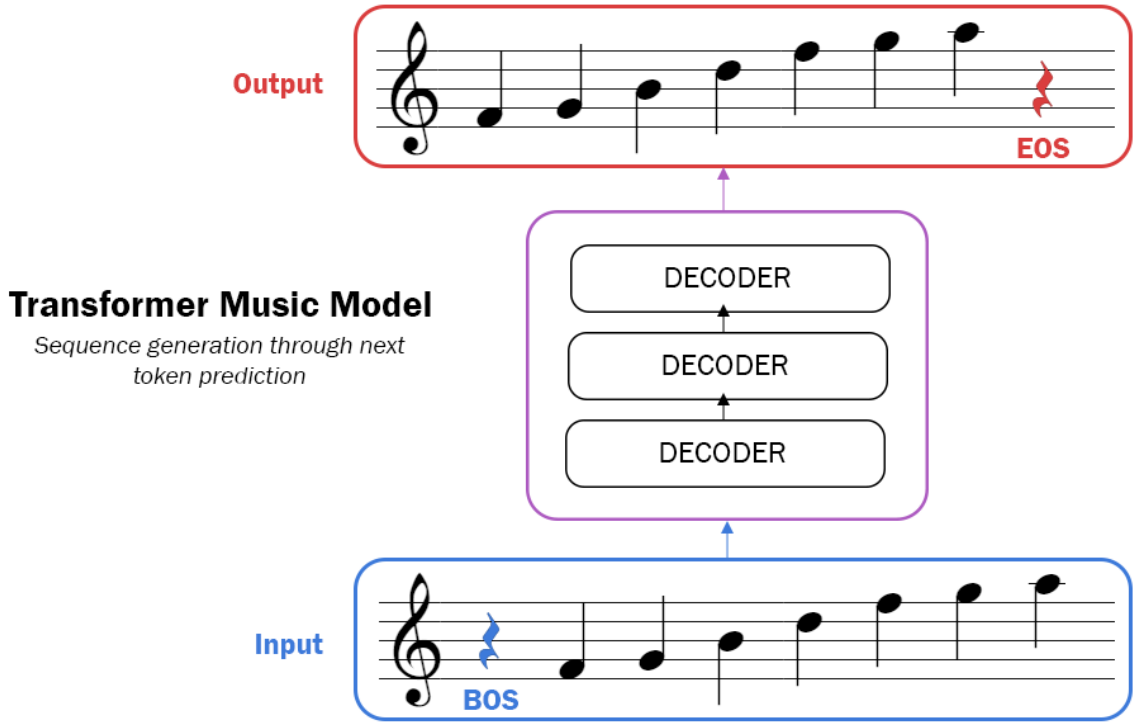
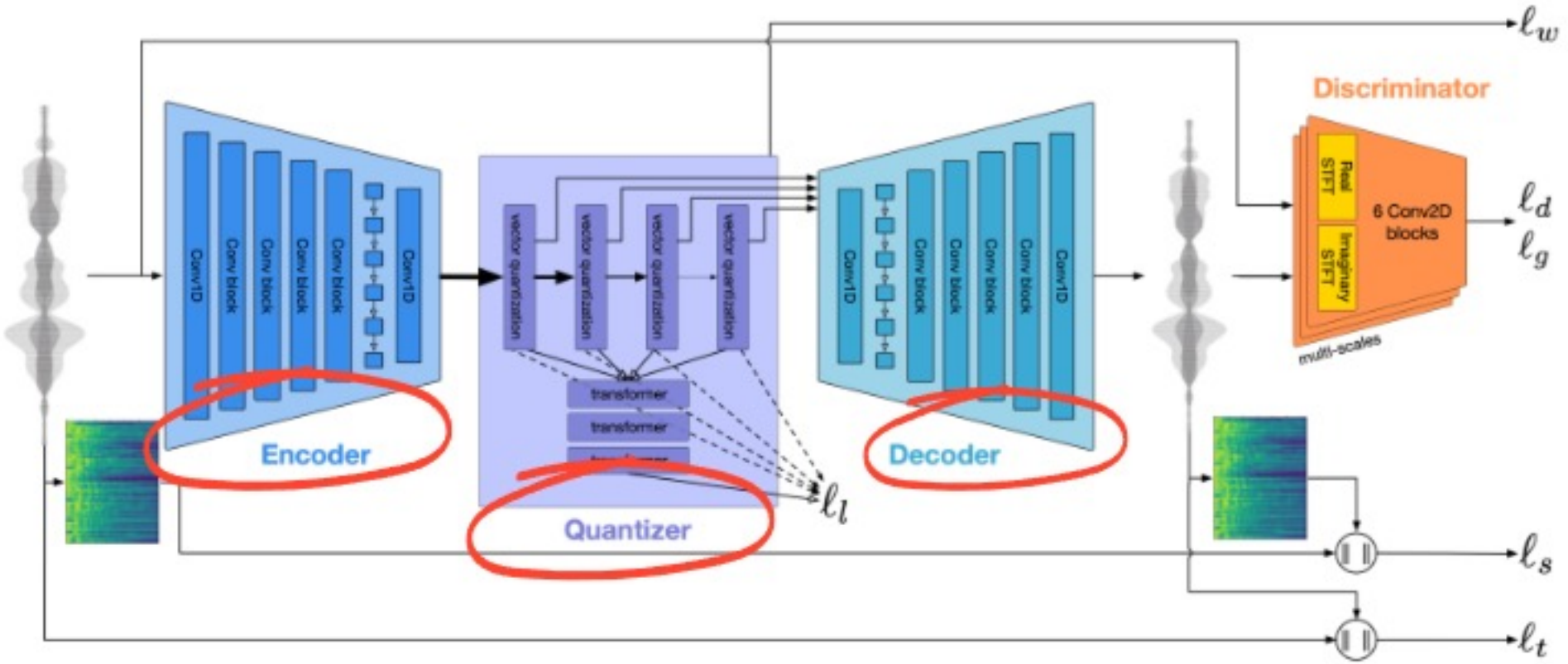
Simple and Controllable Music Generation

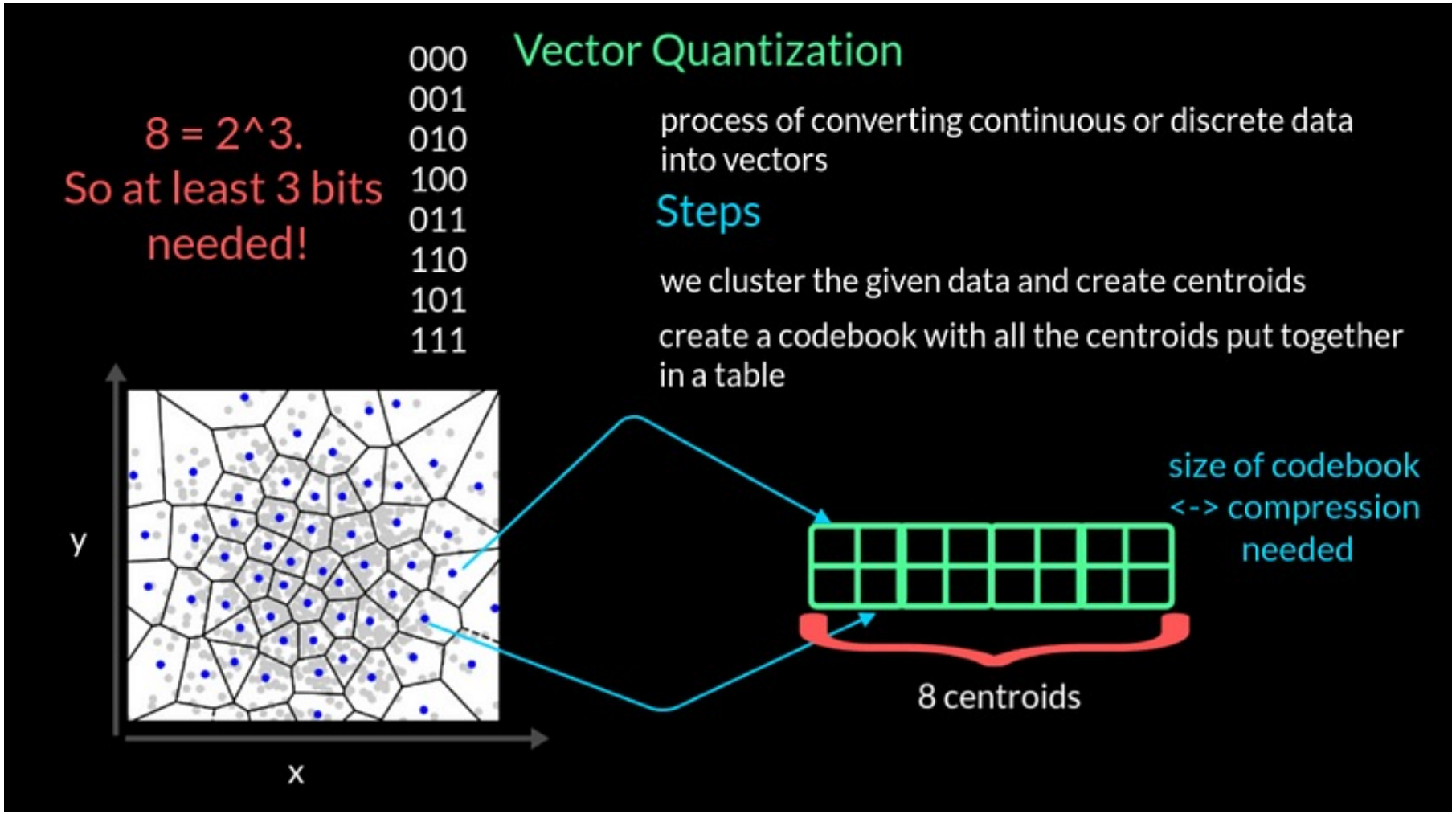
Обзор задачи генерации музыки на основе текста



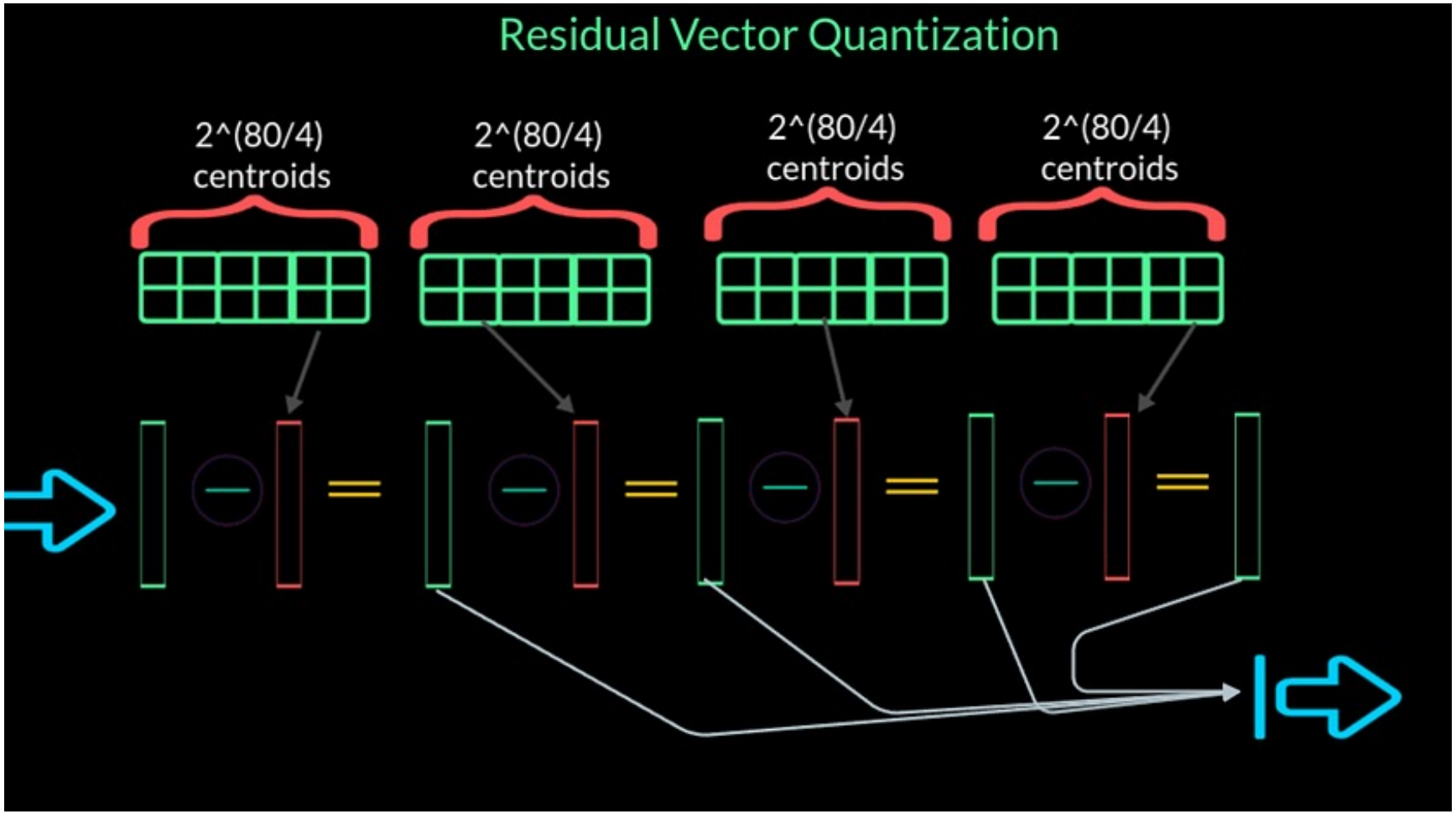
Encoder


ℓ_w
Discriminator
Real STFT
Imaginary STFT
6 Conv2D blocks
multi-scales
ℓ_d
ℓ_g
Conv1D
Conv block
vector quantization
transformer
Encoder
Quantizer
Decoder
ℓ_l
ℓ_s
ℓ_t

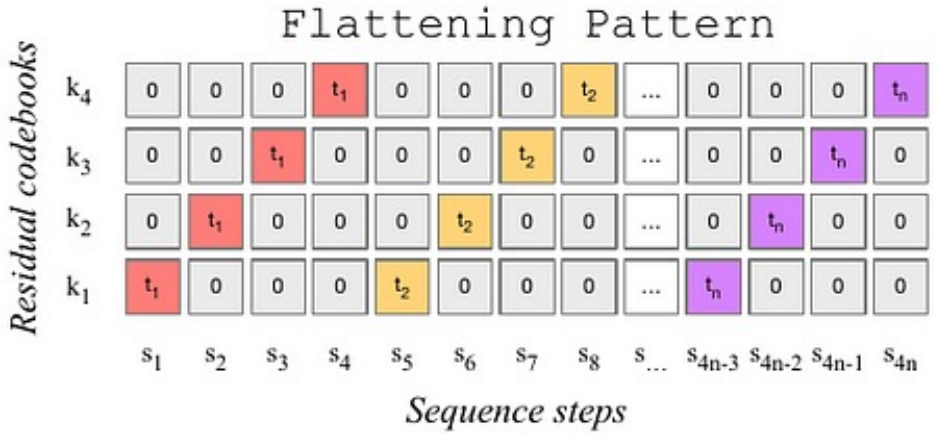
Vector Quantization

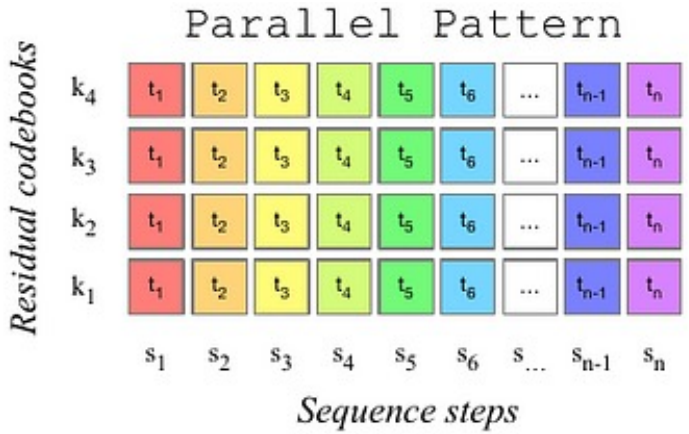


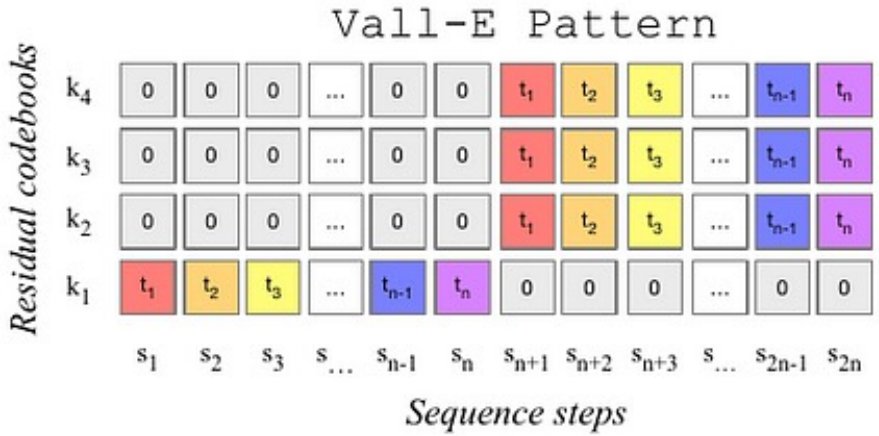
Residual Vector Quantization

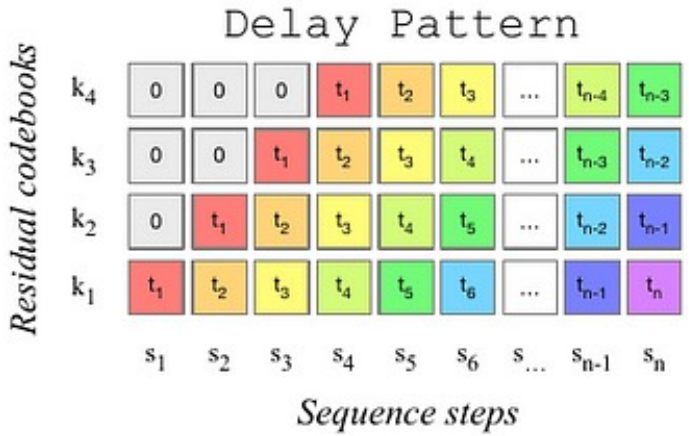


Interleaving Patterns

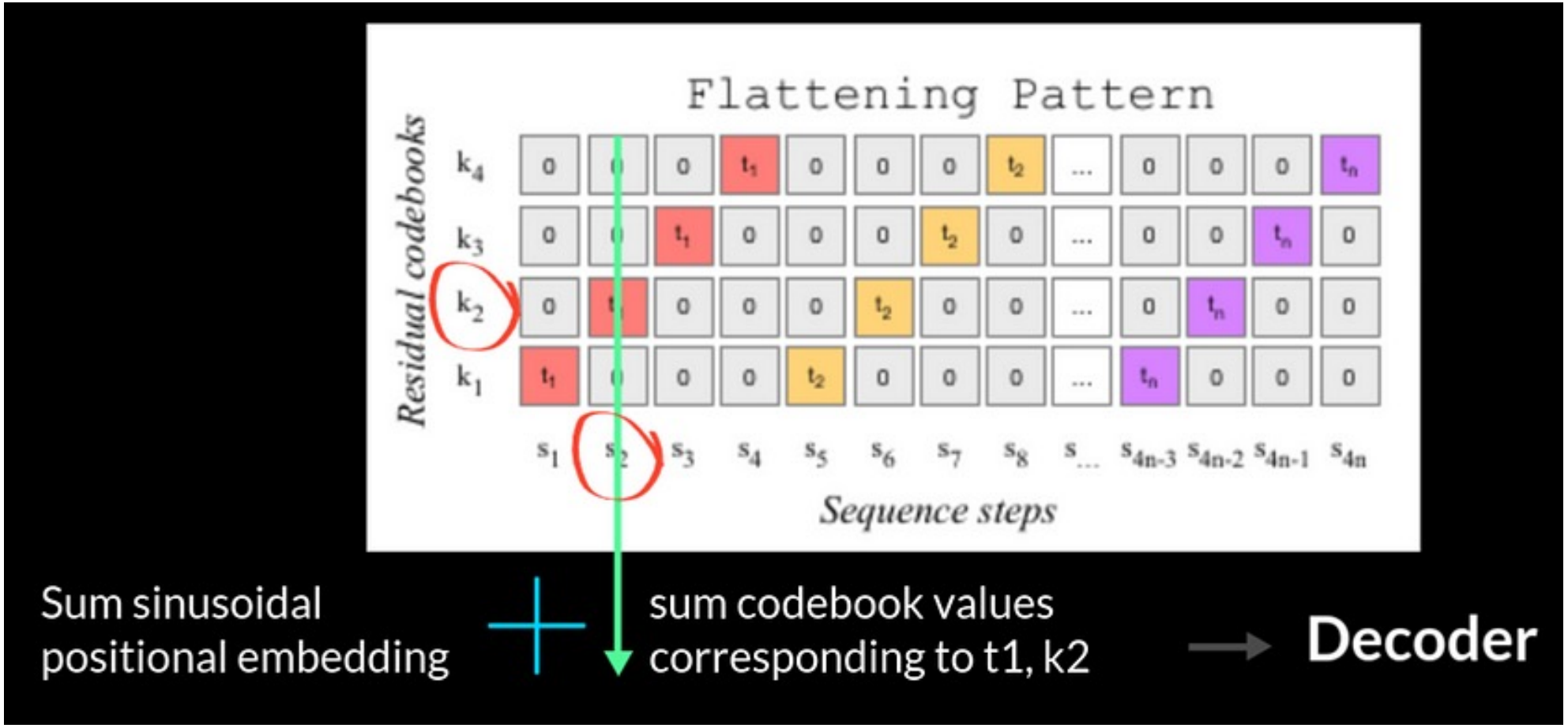








Codebook Projection and Positional Embedding



Model Conditioning with Text and Audio

Text

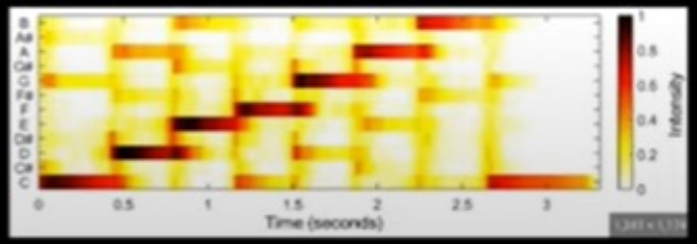
pretrained text encoder T5

FLAN-T5: "Scaling Instruction-Finetuned Language Models"

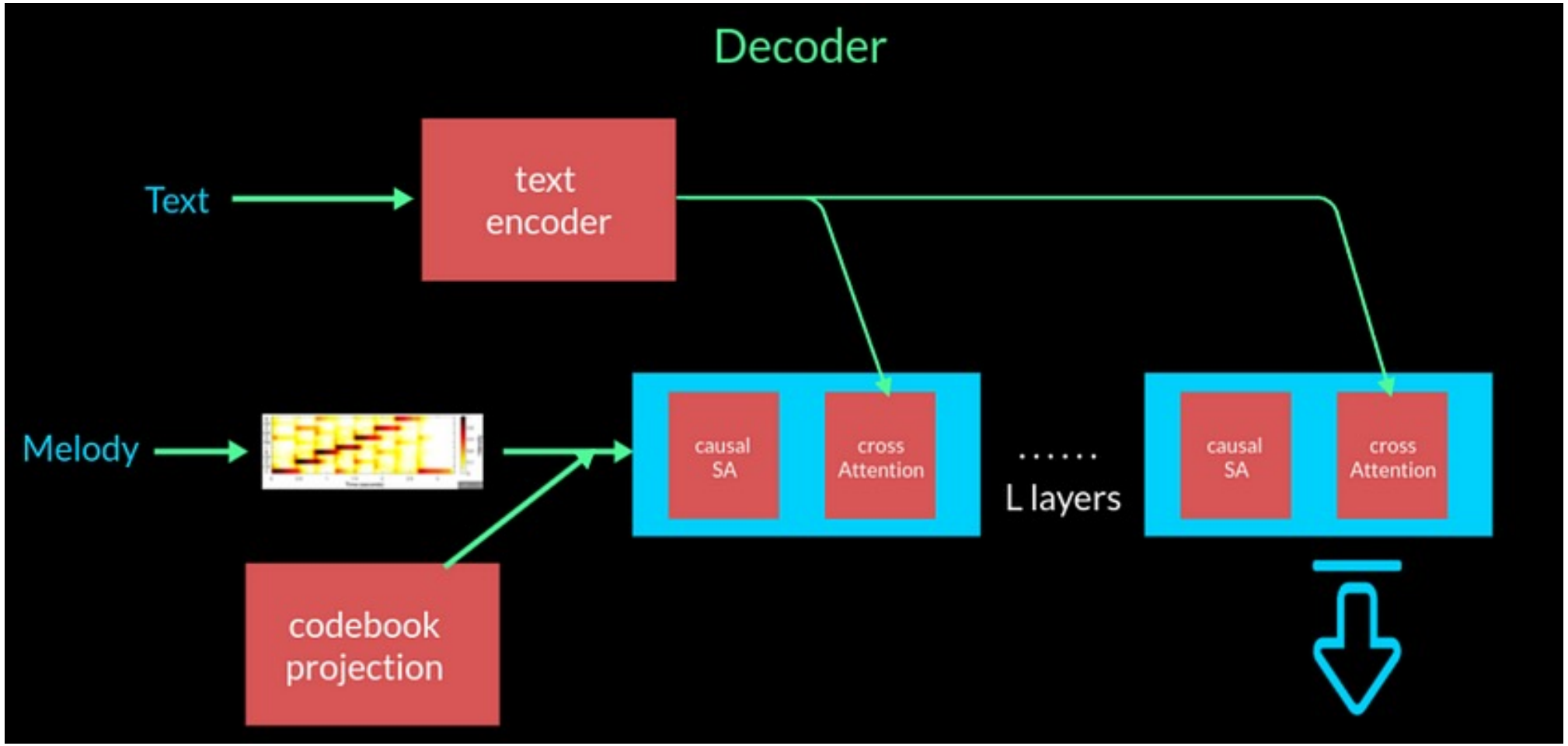
CLAP: Learning Audio Concepts From Natural Language Supervision

Melody

convert to chromogram and
suppress the dominant
frequency



Decoder


Decoder
Text
text encoder
Melody
causal SA
cross Attention
.....
L layers
causal SA
cross Attention
codebook projection

Experimental setup

Модель аудиотокенизации:

- Используется модель EnCodec с пятью слоями для монофонического аудио с частотой дискретизации 32 кГц.
- Шаг составляет 640, что дает частоту кадров 50 Гц.
- Начальный размер скрытого слоя - 64, удваивается на каждом из пяти слоев модели.
- Вложения квантуируются с помощью RVQ с четырьмя квантизаторами, каждый из которых имеет размер кодовой книги 2048.
- Обучение модели происходит на односекундных аудиосегментах, выбранных случайным образом из аудиопоследовательности.

Experimental setup

Модель трансформера:

- Обучаются авторегрессивные модели трансформера разных размеров: 300M, 1.5B, 3.3B параметров.
- Используется эффективный по памяти Flash attention из пакета xFormers для улучшения скорости и использования памяти с длинными последовательностями.
- Обучение моделей происходит на 30-секундных аудиосегментах, выбранных случайным образом из полного трека.
- Используется оптимизатор AdamW с параметрами batch size 192, $\beta 1 = 0.9$, $\beta 2 = 0.95$, весом декоррелированного распределения 0.1 и ограничением градиента 1.0.
- Для модели с 300M параметрами используется D-Adaptation для улучшения сходимости модели.
- Применяется косинусное расписание обучения с разминкой в 4000 шагов.
- Используется экспоненциальное скользящее среднее со затуханием 0.99.
- Обучение моделей с 300M, 1.5B и 3.3B параметрами проводится с использованием 32, 64 и 96 GPU соответственно, с смешанной точностью float16.
- Для выборки используется top-k выборка с оставлением лучших 250 токенов и температурой 1.0.

Experimental setup

Предобработка текста:

- Используется нормализация текста, включая опущение стоп-слов и лемматизацию.
- Проводится эксперимент с объединением дополнительных аннотаций к тексту, таких как музыкальный ключ, темп, тип инструментов и др.
- Применяется метод "word dropout" для аугментации текста.

Datasets

	MusicCaps Test Set				
Model	$\text{FAD}_{vgg}\downarrow$	KL $\downarrow$	$\text{CLAP}_{scr}\uparrow$	Ovl. $\uparrow$	Rel. $\uparrow$
Riffusion	14.8	2.06	0.19	79.31$\pm$1.37	74.20$\pm$2.17
Mousai	7.5	1.59	0.23	76.11$\pm$1.56	77.35$\pm$1.72
MusicLM	4.0	-	-	80.51$\pm$1.07	82.35$\pm$1.36
Noise2Music	**2.1**	-	-	-	-
MusicGen w.o melody (300M)	3.1	1.28	0.31	78.43$\pm$1.30	81.11$\pm$1.31
MusicGen w.o melody (1.5B)	3.4	1.23	**0.32**	80.74$\pm$1.17	**83.70**$\pm$1.21
MusicGen w.o melody (3.3B)	3.8	**1.22**	0.31	**84.81**$\pm$0.95	82.47$\pm$1.25
MusicGen w. random melody (1.5B)	5.0	1.31	0.28	81.30$\pm$1.29	81.98$\pm$1.79

Evaluation

- MUSICGEN сравнивается с двумя бейзлайнами для генерации музыки из текста: Riffusion и Mousai.
- Riffusion используется для вывода результатов, а Mousai обучается на предоставленном датасете для справедливого сравнения.
- Помимо этого, при возможности также проводится сравнение с MusicLM и Noise2Music.
- Для оценки используются объективные и субъективные метрики.
- Объективные метрики включают в себя Fréchet Audio Distance (FAD), Kullback-Leibler Divergence (KL) и CLAP score (CLAP).

Results

		In Domain Test Set			
TRAIN CONDITION	TEST CONDITION	SIM. ↑	MEL. ↑	OVL. ↑	REL. ↑
Text	Text	0.10	64.44±0.83	82.18±1.21	81.54±1.22
Text+Chroma	Text	0.10	61.89±0.96	81.65±1.13	**82.50**±0.98
Text+Chroma	Text+Chroma	**0.66**	**72.87**±0.93	**83.94**±1.99	80.28±1.06

	MUSICCAPS Test Set				
MODEL	FAD_{vgg} ↓	KL ↓	$CLAP_{scr}$ ↑	OVL. ↑	REL. ↑
Riffusion	14.8	2.06	0.19	79.31±1.37	74.20±2.17
Mousai	7.5	1.59	0.23	76.11±1.56	77.35±1.72
MusicLM	4.0	-	-	80.51±1.07	82.35±1.36
Noise2Music	**2.1**	-	-	-	-
MUSICGEN w.o melody (300M)	3.1	1.28	0.31	78.43±1.30	81.11±1.31
MUSICGEN w.o melody (1.5B)	3.4	1.23	**0.32**	80.74±1.17	**83.70**±1.21
MUSICGEN w.o melody (3.3B)	3.8	**1.22**	0.31	**84.81**±0.95	82.47±1.25
MUSICGEN w. random melody (1.5B)	5.0	1.31	0.28	81.30±1.29	81.98±1.79

Обзор источников

- Слайды: https://www.youtube.com/watch?v=cbAa7kart-4
- Оригинальная статья: https://arxiv.org/abs/2306.05284