Энкодеры

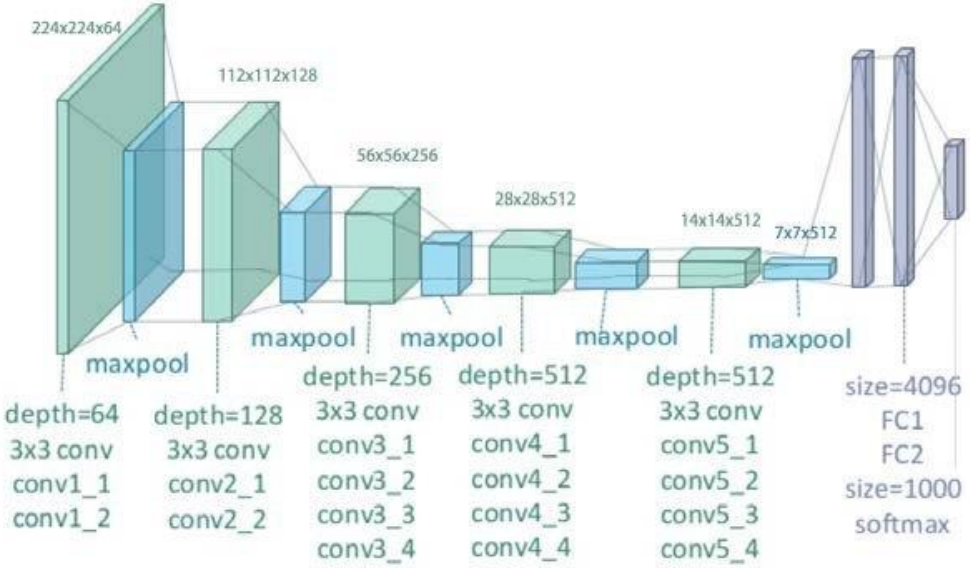
Что такое энкодеры

Энкодер – модель или компонент модели, которая преобразует входные данные в векторное представление.

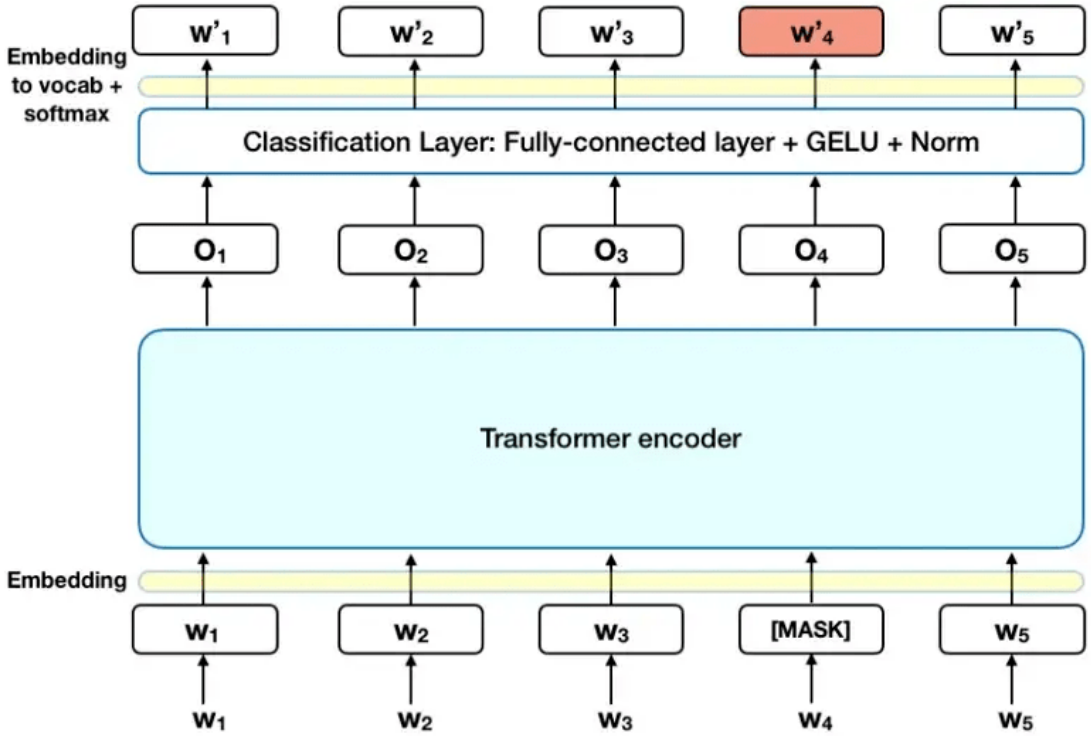
Какие бывают энкодеры

Скрытый слой обученной модели. Например, последний слой ResNet
Обучение специальной модели энкодера. Например, BERT
Обучение автокодировщика, например VAE
Обученный энкодер seq2seq модели. Например, Whisper или FridaT5

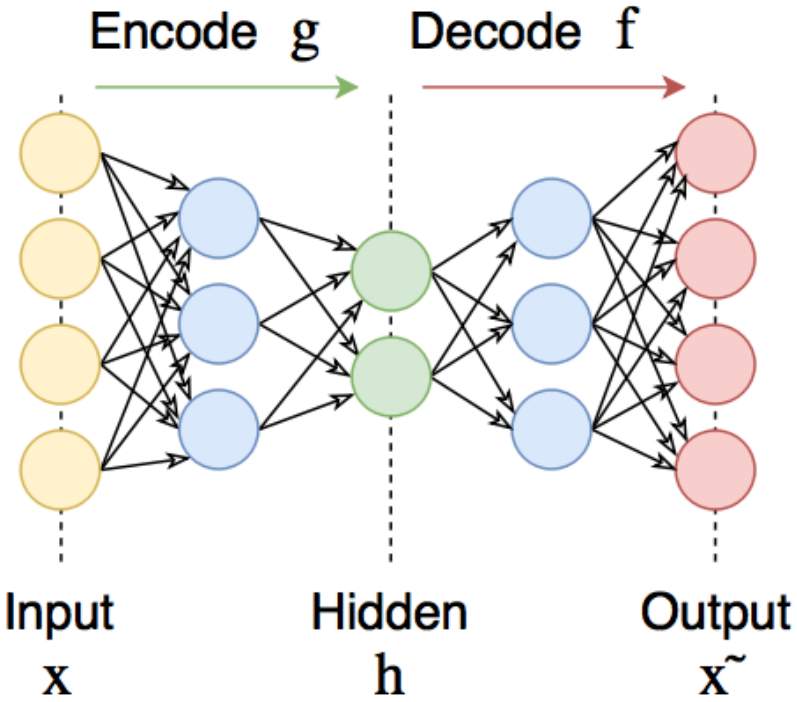
скрытый слой



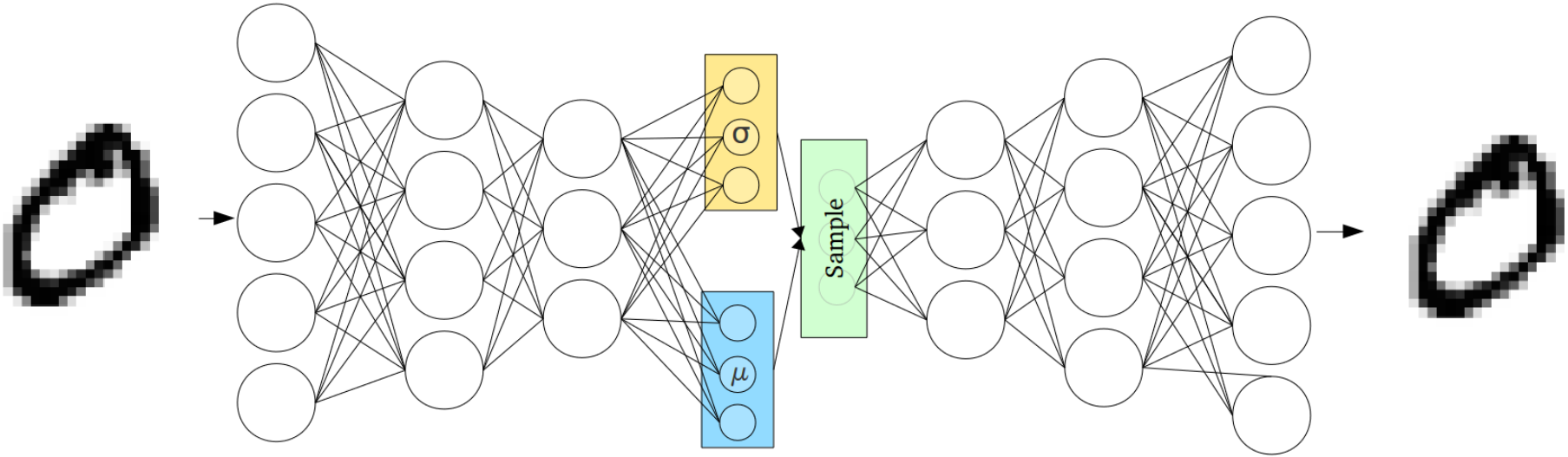
BERT


w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Автокодировщик



Вариационный автокодировщик



Энкодер seq2seq модели

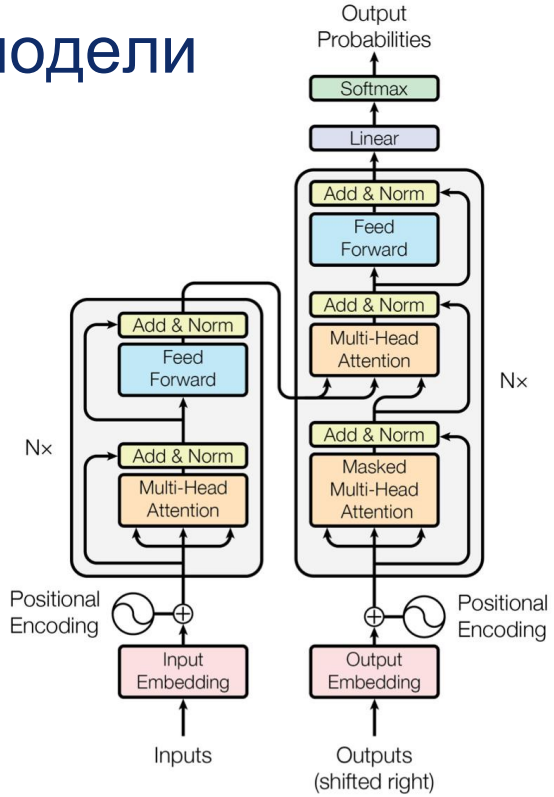
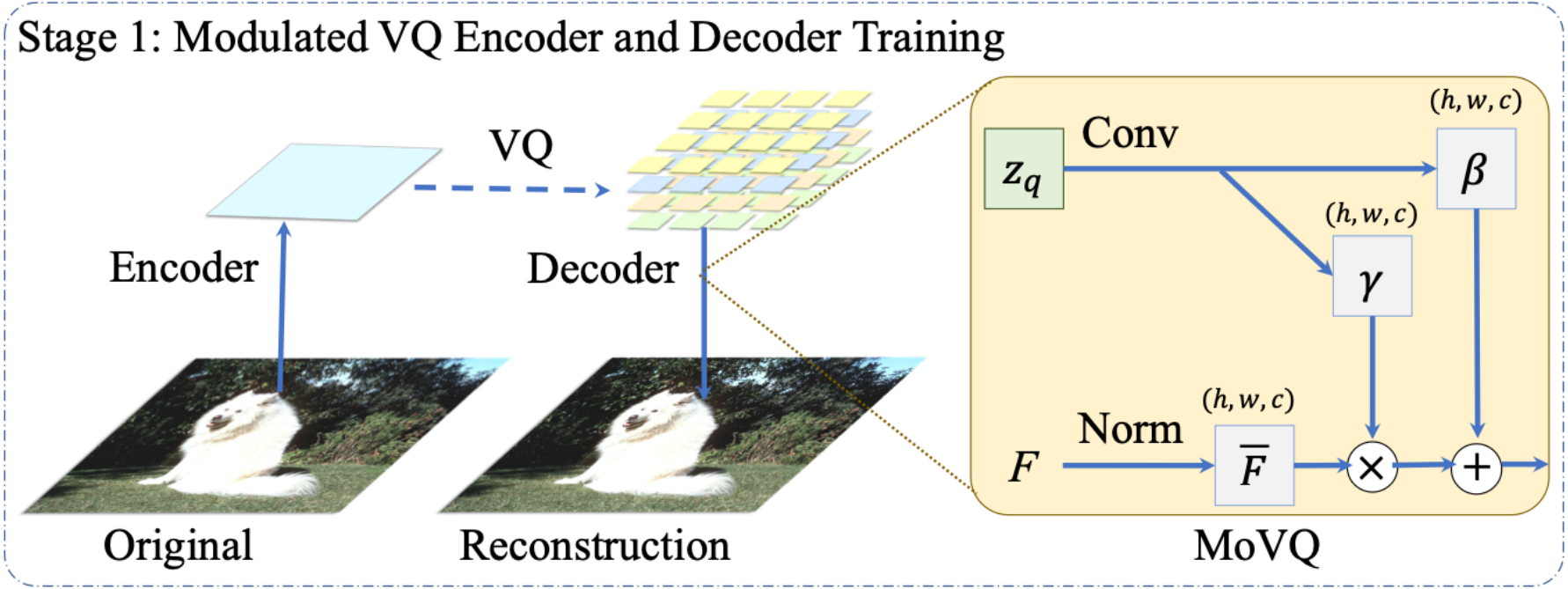


Figure 1: The Transformer - model architecture.

Энкодеры изображений

image to vector: ResNet, AE, VaE
image to latent: VQGAN
Image to sequence: ViT

VQGAN



https://habr.com/ru/companies/sberbank/articles/740624

VQGAN

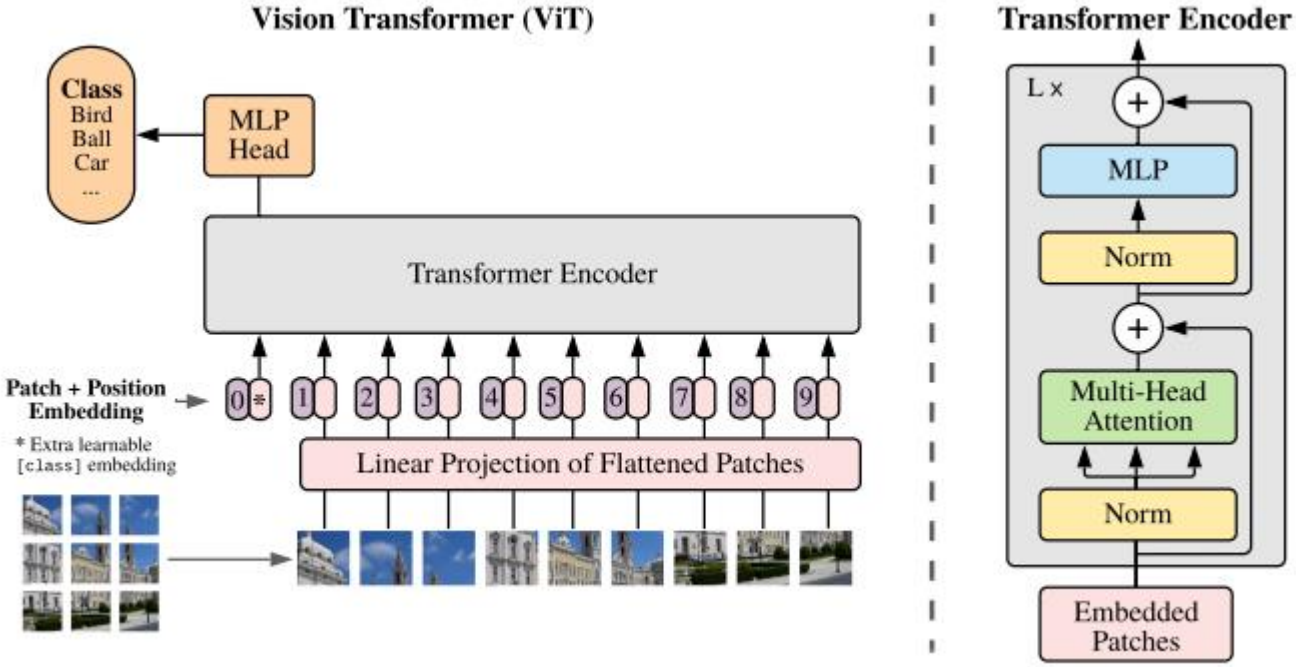
отображение входного изображения в скрытое представление
квантизация скрытого представления
реконструкция изображения из многомерного массива токенов

VQGAN. задачи

генерация изображений

объединение модальностей для генерации изображений

Vision transformer



https://arxiv.org/pdf/2010.11929

Vision transformer

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Vision transformer

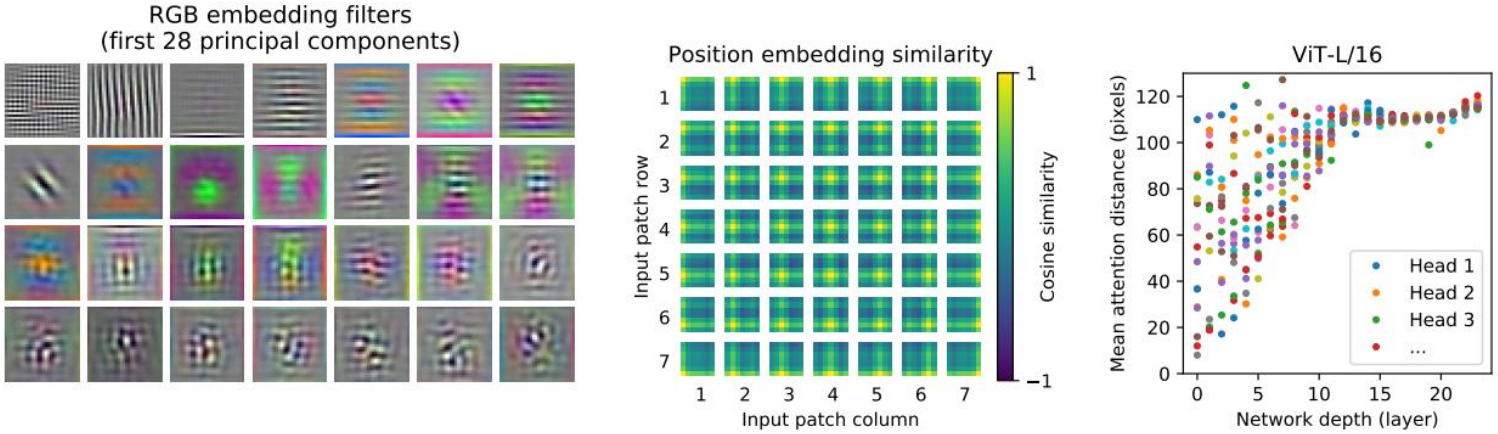


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

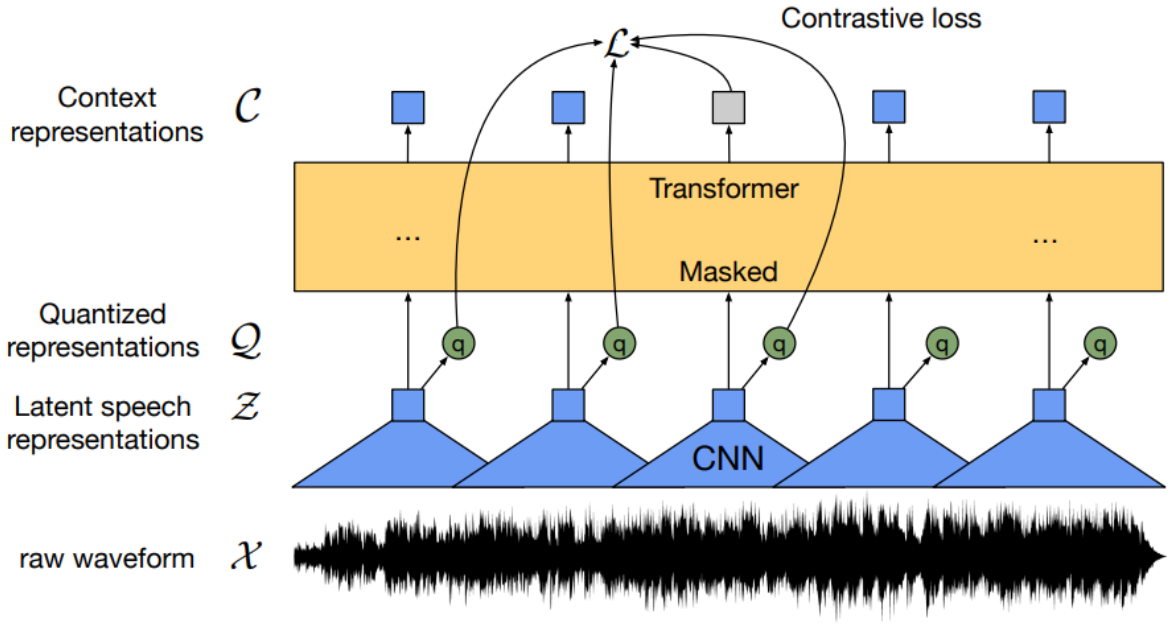
Vision transformer. Задачи

кодирование изображения в последовательность для LLM
классические задачи с изображнениями – классификация, детекция изображений

Аудио энкодеры

энкодеры – спектрограмма, wav2vec, HuBERT
токенизаторы – RVQ, Encodec, Mimi, Xcodec
большие модели - whisper

Wav2Vec2

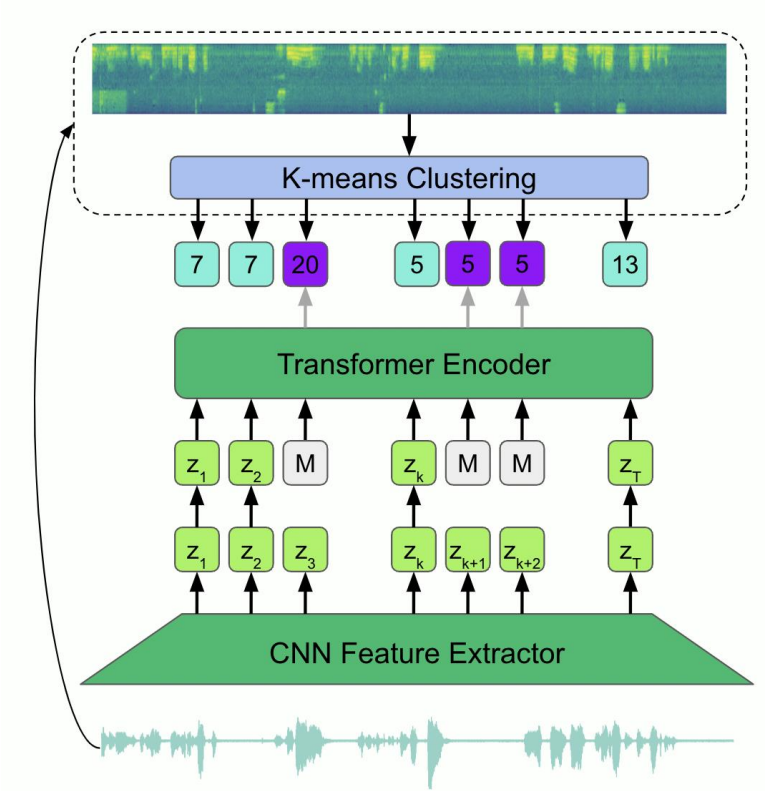


https://arxiv.org/pdf/2006.11477

Wav2Vec2

кодирует общую информацию об аудио
подходит для универсальных задач (captioning, classification, etc)

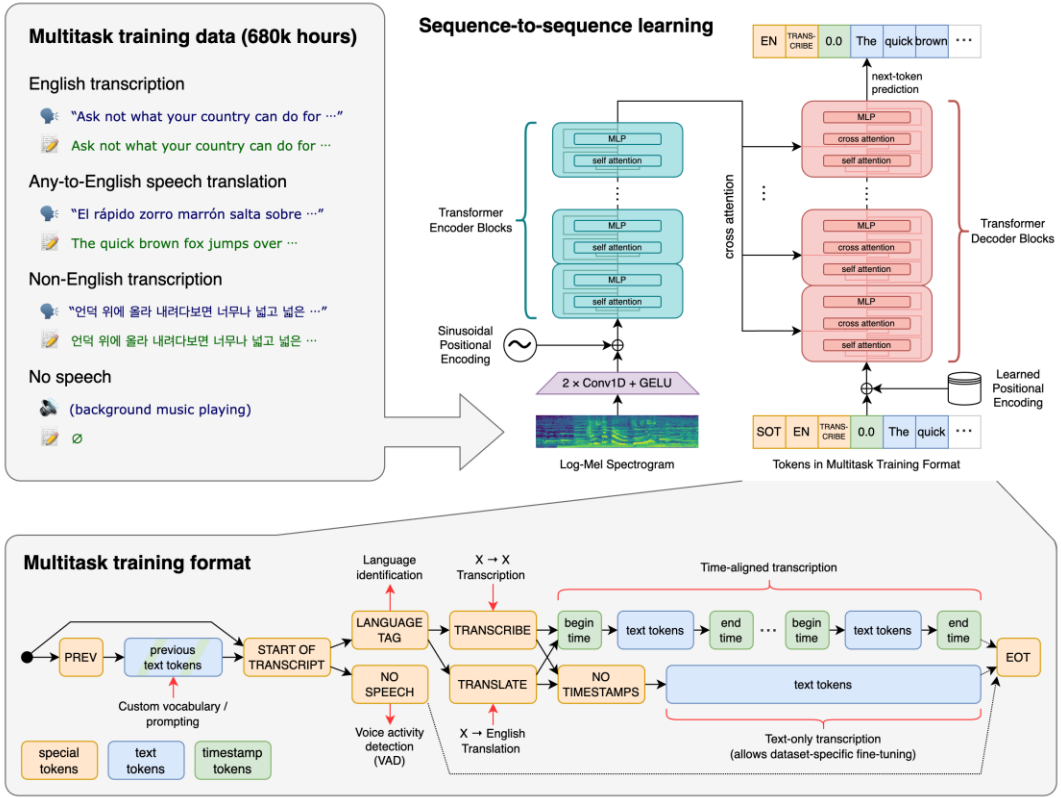
HuBERT


K-means Clustering
7
7
20
5
5
5
13
Transformer Encoder
z1
z2
M
zk
M
M
zT
z1
z2
z3
zk
zk+1
zk+2
zT
CNN Feature Extractor

HuBERT

кодирует псевдофонетическую информацию
подходит для речевых задач (ASR, translation, etc)

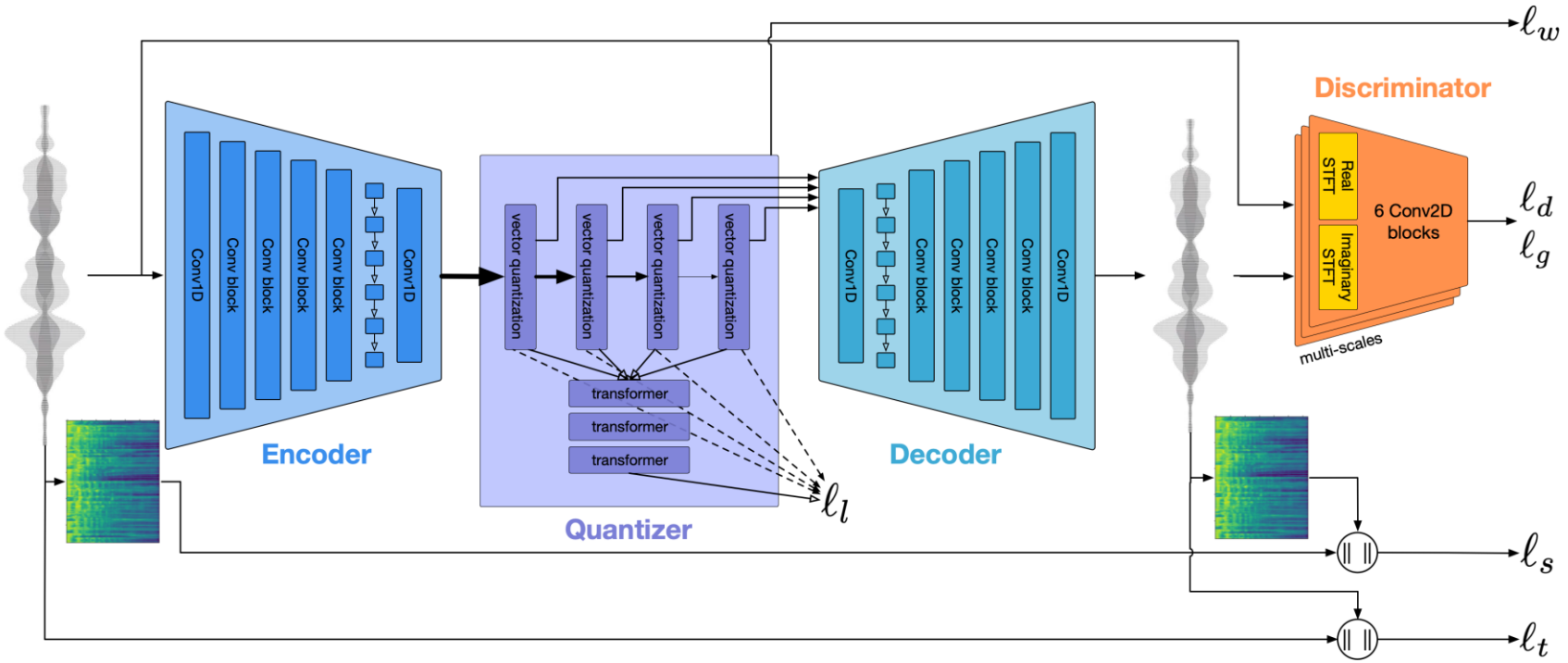
Whisper


Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
Transformer Encoder Blocks
MLP
self attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
cross attention
Transformer Decoder Blocks
next-token prediction
Learned Positional Encoding
EN
TRANS-CRIBE
0.0
The
quick
brown
SOT
Tokens in Multitask Training Format
Multitask training format
PREV
previous text tokens
Custom vocabulary / prompting
START OF TRANSCRIPT
Language identification
LANGUAGE TAG
NO SPEECH
Voice activity detection (VAD)
X → X Transcription
TRANSCRIBE
TRANSLATE
X → English Translation
Time-aligned transcription
begin time
text tokens
end time
NO TIMESTAMPS
Text-only transcription (allows dataset-specific fine-tuning)
EOT
special tokens
text tokens
timestamp tokens

Whisper

понимание языка
мультиязычные задачи

Encodec



https://arxiv.org/pdf/2210.13438

Encodec

сжатие аудио
генерация аудио

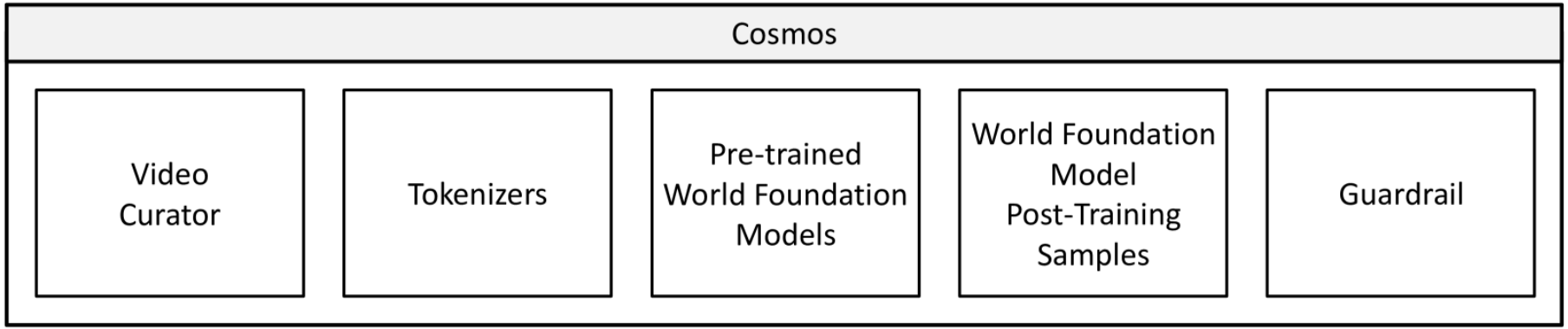
Видео

кодирование изображения и звука отдельно и объединеие (SALMONN, VideoLLaVa etc)
Специальные энкодеры (Cosmoss)

Видео

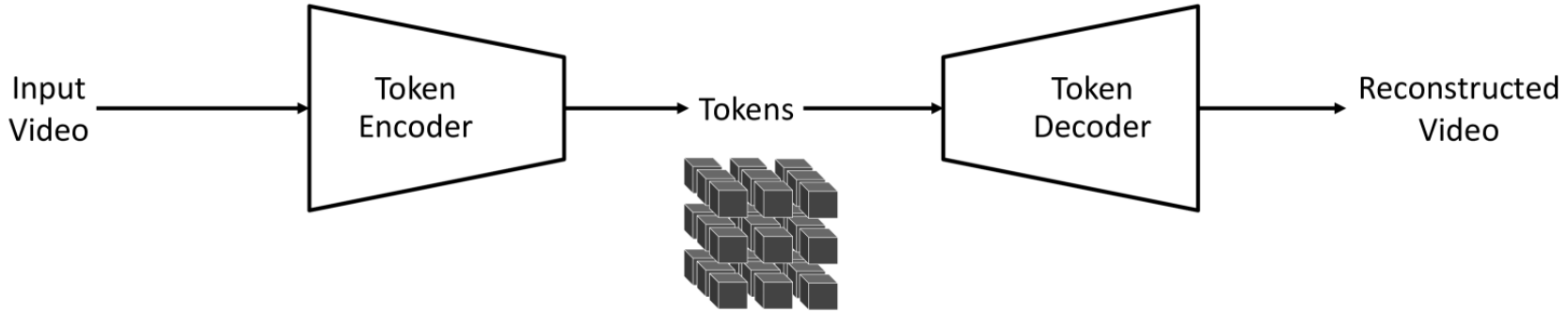
кодирование изображения и звука отдельно и объединеие (SALMONN, VideoLLaVa etc)
Специальные энкодеры (Cosmoss)

Видео. Cosmos



https://arxiv.org/html/2501.03575v1

Видео. Cosmos



https://arxiv.org/html/2501.03575v1


ВСЁ!