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 ⋯"
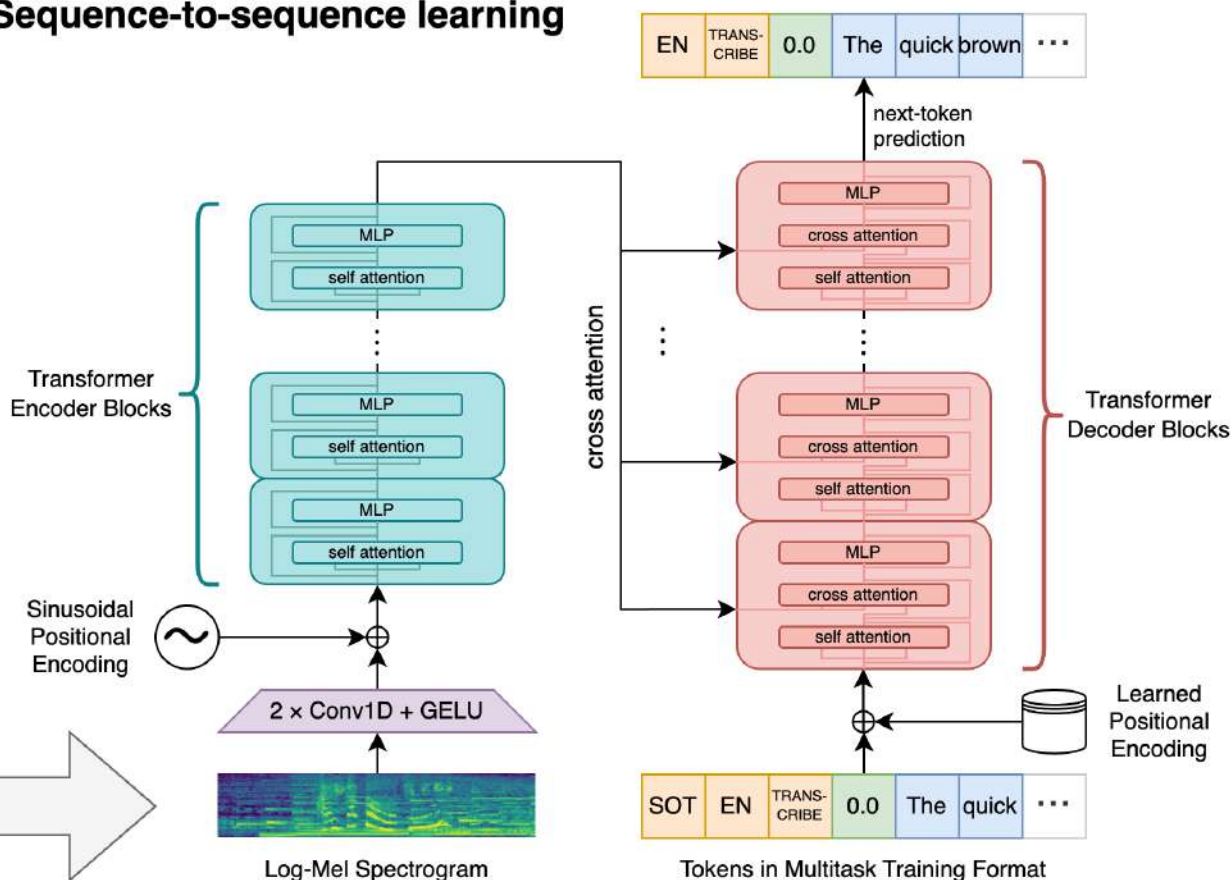
📝 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

🔊 (background music playing)

📝 ∅

Sequence-to-sequence learning

| EN | TRANS-CRIBE | 0.0 | The | quick | brown | ⋯ |

next-token prediction

MLP
cross attention
self attention

MLP
cross attention
self attention

MLP
cross attention
self attention

Transformer Decoder Blocks

cross attention

MLP
self attention

MLP
self attention

MLP
self attention

Transformer Encoder Blocks

Sinusoidal Positional Encoding

2 × Conv1D + GELU

Log-Mel Spectrogram

Learned Positional Encoding

| SOT | EN | TRANS-CRIBE | 0.0 | The | quick | ⋯ |

Tokens in Multitask Training Format

Multitask training format

Language identification

X → X Transcription

Time-aligned transcription

PREV → previous text tokens → START OF TRANSCRIPT → LANGUAGE TAG → TRANSCRIBE → begin time → text tokens → end time → ⋯ → begin time → text tokens → end time → EOT

Custom vocabulary / prompting

NO SPEECH

TRANSLATE

NO TIMESTAMPS → text tokens

Voice activity detection (VAD)

X → English Translation

Text-only transcription (allows dataset-specific fine-tuning)

special tokens

text tokens

timestamp tokens