Whisper, модел за аутоматску транскрипцију и превод

JePTex
Четвртак, 22. 02. 2024

Whisper

Објављен у септембру 2022.

Трениран на аудио материјалу у трајању од 680.000 сати

Број сати на нашем језику: 28

Модел је објављен у различитим величинама (број параметара):

tiny 39 M

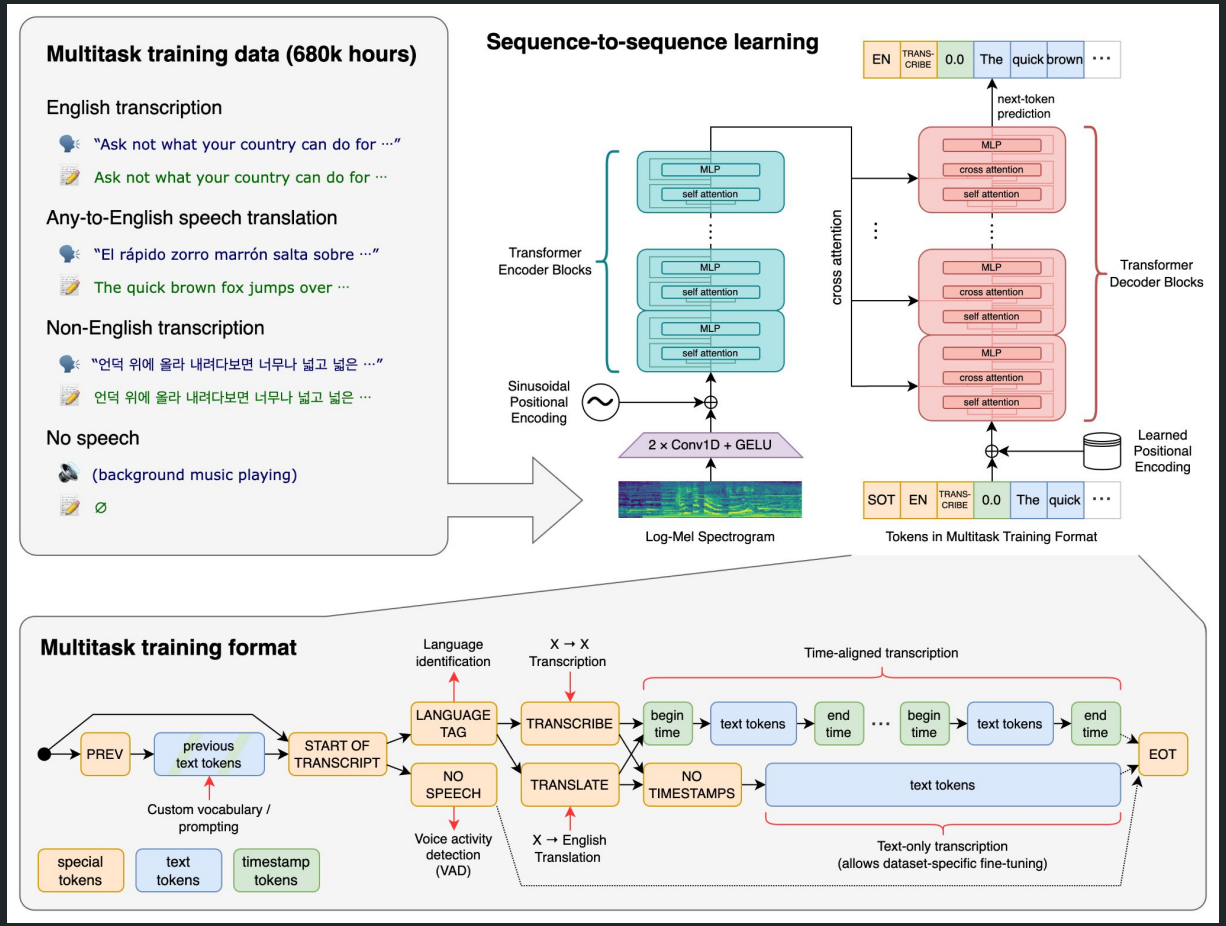
base 74 M

small 244 M

medium 769 M

large 1550 M

Процес обуке



WER (Word Error Rate) за српски језик у %

Whisper tiny 83.7

Whisper base 64.3

Whisper small 42.2

Whisper medium 44.9

Whisper large 29.2

Whisper large-v2 33.9

WER (Word Error Rate) за српски у %

Whisper tiny 83.7

Whisper base 64.3

Whisper small 42.2

Whisper medium 44.9

Whisper large 29.2

Whisper large-v2 33.9

Whisper large-v3 11.6

google/Fleurs

Whisper tiny 106.1

Whisper base 103.0

Whisper small 101.3

Whisper medium 85.6

Whisper large 87.4

Whisper large-v2 70.5

Whisper large-v3 15.7

Mozilla/CommonVoice 9

Извор: https://cdn.openai.com/papers/whisper.pdf

Побољшана подршка за наш језик

У децембру 2022. године портал Hugging Face (HF) орагнизовао је изазов за дообуку Whisper модела.

Један од победничких модела био је и:

https://huggingface.co/DrishtiSharma/whisper-large-v2-serbian

Wer: 10.7649

mozilla-foundation/common_voice_11_0

Како је све почело?

У дигиталној колекцији библиотеке "Милутин Бојић" постоји аудио колекција Михаила Миљковића која се састоји од 14 снимака.

https://zavicajna.digitalna.rs:443/jsp/RcWebBrowse.jsp?browse_cid=a65de25a-262f-4d67-ab36-3ce82db3521a

Прва 3 снимка транскрибована су користећи поменути модел.

Нов пројекат МК за 2023!

Добијена подршка од МК.

Пројекат

Набавка рачунара која ће служити за дообучавање модела:

ThinkStation P3 Tower

i9 13900x
RAM 128G
Nvidia RTX A5500 (24G)
NVME SSD 2T
NVME SSD 4T
HDD 6T

Припрема и обука

У јулу 2023. HF је организовао нов аудио курс, део тог курса је садржао и дообуку Whisper модела.

Изазов

Доступност отворених аудио датасетова на српском
google/Fleurs
mozilla/common-voice

Потом сам открио и датасет од 55 сати “Јужне вести” на Clarin репозиторијуму
https://www.clarin.si/repository/xmlui/handle/11356/1679

Пре неколико недеља Clarin је објавио нових 850 сати аудио датасета као део пројекта “Parla Speech”
https://www.clarin.si/repository/xmlui/handle/11356/1834
Serbian Parliament Corpus

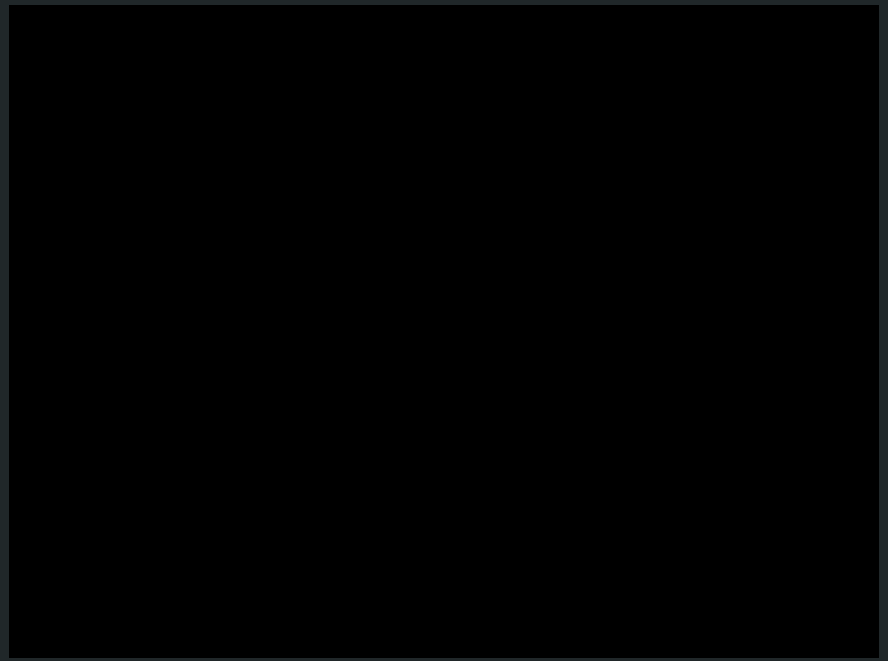
Резултати

Дообучене верзије Whisper модела за наш језик на спојеним датасетовима:
Јужне вести + Fleurs + CommonVoice

https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb
Wer: 4.15%

https://huggingface.co/Sagicc/whisper-medium-sr-cmb
Wer: 6.58%

Демо



https://gitlab.com/aadnk/whisper-webui/-/tree/main

ХВАЛА!

andrija.sagic@gmail.com