Més sobre Whisper i Google Colab

Tipus: [1]
Autor: Gutiérrez Ferrerías, Fernando [2]
Creació: Publicat per Fernando Gutiérrez Ferrerías [2] el 16/11/2023 - 14:35 | Última modificació: 16/11/2023 - 17:23
Etiquetes: accessibilitat
Etiquetes: vídeo
Etiquetes: youtube
Etiquetes: subtítols
Etiquetes: WCAG
4 adjunts [3]

Si esteu fent servir l'IA Whisper per subtitular els vostres vídeos amb el mètode que exposava en un article anterior [4], potser haureu advertit que la mida del programari ha variat lleugerament de 2,87 GB a 2,88 GB. La raó és que hi ha un nou model d'entrenament de la IA presentat fa uns dies per OpenAi en el seu Dev Day.

Si recordeu l'article anterior, l'ordre per executar Whisper era aquesta:

!whisper "/content/audio.mp3" --task transcribe --model large --language=ca --verbose False --output_format srt --output_dir audio_transcription

Els models d'entrenament de Whisper

Doncs bé, els models d'entrenament disponibles en Whisper que s'introdueixen amb el paràmetre --model, són: tiny, base, small, medium, large (o large-v1), large-v2 i el nou presentat ara anomenat large-v3. Les diferències són la quantitat de paràmetres amb que estan entrenats aquests models -des dels 39 milions del més petit fins els 1.550 milions dels més grans- que es correlacionen directament amb el seu nivell de precisió i inversament de rapidesa.

El(s) model(s) large

En la pràctica i pel cas que ens ocupa, en català només recomano fer servir qualsevol dels models large doncs són els únics que proporcionen resultats satisfactoris. Les diferències entre les tres versions del model large, segons la meva experiència són:

--model large, és adequat per a vídeos principalment en català o bilingües en català/castellà. És també el model més gramaticalment acurat, incorpora millor els signes de punctuació, majúscules, etc. a les frases generades, és més editorial si voleu.
--model large-v2, és adequat per a vídeos principalment en català o trilingües en català/castellà i un tercer idioma, com l'anglès.
--model large-v3, suposadament és un avenç en precisió i velocitat respecte a large-v2, en el qual es basa. L'he provat en alguns vídeos i diria que al·lucina un xic més que els seus germans petits.

Comparativa de models large

Prenem com a exemple aquesta entrevista a Theodor Kallifatides [5], un escriptor suec d'origen grec que, en el vídeo s'expressa en anglès i castellà, mentre que entrevistadora i traductora ho fan en català i castellà principalment. Adjunt trobareu la comparativa dels subtítols generats per Whisper large sense indicació d'idioma (és a dir sense el paràmetre --language ca), large amb indicació d'idioma i large-v2 durant els primers 15 minuts. En groc marcats els canvis més rellevants, veureu que tot és més o menys igual fins al subtítol 85-95 en que l'entrevistadora formula una pregunta en castellà i large tradueix la resposta d'en Theodor (subtítols 96 a 115) al català, mentre que large-v2 reconeix el canvi de llengua i la transcriu en anglès.

Formats de sortida, no només subtítols

Ja que entrem a disseccionar Whisper, comentar també que no només pot generar subtítols en format SRT, adequats per a YouTube, sinó que Whisper ens ofereix fins a cinc possibilitats, a saber: srt, txt, tsv, vtt i json. La transcripció TXT serà en text pla sense marques de temps. Si us cal algun d'aquests formats, podeu demanar-li variant aquest paràmetre de la instrucció (es generaran cinc arxius diferents a la carpeta de sortida):

--output_format all

Optimitzar l'ús de Google Colab per evitar penalitzacions

Com us deia en l'anterior article, Google Colab és un entorn colaboratiu de computació compartida que Google possa a l'abast de manera gratuïta però amb certes limitacions, entre d'altres el temps màxim d'execució d'un quadern és de 12 hores i es desconnecta automàticament si detecta temps d'inactivitat. En aquests casos potser que després d'unes quantes vegades et penalitzi i et deixi sense accés al que denomina unitats de computació. M'ha passat. Per tant, sempre que no estigueu fent servir el Colab, tanqueu l'entorn d'execució (Runtime > Disconnect and delete runtime) per evitar penalitzacions.

Com rebre una notificació sonora quan finalitzi la transcripció

En el cas de vídeos llargs, parlem de més d'una hora, no és gaire pràctic estar pendent de quan acaba la transcripció i si no estàs atent, pot ser que es desconnecti el Colab, perdis l'arxiu i hagis de tornar a començar. Per evitar-ho, i poder seguir fent altres tasques mentre Whisper fa la seva, podeu executar aquest codi a continuació de la cel·la de Whisper:

```
# Reprodueix automàticament el fitxer d'àudio un cop acabada la transcripció
from IPython.display import Audio, display
display(Audio('/content/audio.mp3', autoplay=True))
```

En cas de vídeos curts s'iniciarà la reproducció de l'àudio transcrit, però en el cas de vídeos llargs, tot i que a mi em va funcionar en un primer moment, després d'uns quants sembla ser que Google Colab penalitza d'alguna manera la reproducció d'streams llargs i deixa de funcionar. Per aquest motiu actualment estic fent servir un petit arxiu generat per Audacity que carrego manualment a l'espai d'usuari del Colab:

```
# Reprodueix automàticament el fitxer d'àudio un cop acabada la transcripció
from IPython.display import Audio, display
display(Audio('/content/dtmf.mp3', autoplay=True))
```

Us el deixo adjunt també, podeu fer servir aquest si voleu, o un efecte de Pixabay [6] (cha-cha-ender) o la cinquena de Beethoven [7].

Efemèride: 25 anys de les WCAG

Res més, només recordar que "*1. Provide equivalent alternatives to auditory and visual content.*" era la primera de les primeres Pautes d'Accessibilitat del Contingut Web (WCAG 1.0 [8]), publicades allà per l'any 1999 (per tant sis anys abans del naixement de YouTube). A veure si aconseguim satisfer ni que sigui aquest punt amb motiu de la propera efemèride dels 25 anys de la publicació de les primeres WCAG que tindrà lloc el proper 5 de maig de 2024. Actualment però, la versió vigent que hauriem de satisfer són les WCAG 2.1 [9].
Categories: Punt web
Etiquetes: accessibilitat
Etiquetes: vídeo
Etiquetes: youtube

Etiquetes: subtítols
Etiquetes: WCAG

Adjunt	Mida
theodor_kalifatides_asr_15m.docx [10]	49.62 KB
cha-cha-ender-162072.mp3 [11]	154.29 KB
beet5mov1bars1to5.ogg [12]	73.8 KB
dtmf.mp3 [13]	8.93 KB

- [14]

URL d'origen: https://puntweb.diba.cat/blogs/2023/11/16/mes-sobre-whisper-google-colab

Enllaços:
[1] https://puntweb.diba.cat/
[2] https://puntweb.diba.cat/members/gutierrezff
[3] https://puntweb.diba.cat/blogs/2023/11/16/mes-sobre-whisper-google-colab
[4] https://puntweb.diba.cat/blogs/2023/10/16/subtitulacio-de-videos-catala-amb-whisper
[5] https://www.youtube.com/watch?v=p6zM2DJt-E4
[6] https://pixabay.com/sound-effects/search/creative-commons/
[7] https://commons.wikimedia.org/wiki/File:Beet5mov1bars1to5.ogg
[8] https://www.w3.org/TR/WAI-WEBCONTENT/
[9] https://www.w3.org/TR/WCAG21/
[10] https://puntweb.diba.cat/sites/puntweb.diba.cat/files/theodor_kalifatides_asr_15m.docx
[11] https://puntweb.diba.cat/sites/puntweb.diba.cat/files/cha-cha-ender-162072.mp3
[12] https://puntweb.diba.cat/sites/puntweb.diba.cat/files/beet5mov1bars1to5.ogg
[13] https://puntweb.diba.cat/sites/puntweb.diba.cat/files/dtmf.mp3
[14] https://puntweb.diba.cat/node/1409