OCR-modeller for samiske språk

Dette er en samling av modeller for OCR (optical character recognition) av samiske språk. Disse kan brukes til å gjenkjenne tekst i bilder av trykt tekst (scannede bøker, magasiner, o.l) på nordsamisk, sørsamisk, lulesamisk og inaresamisk.

Mer detaljert informasjon om trening og evaluering av modellene kan du lese i artikkelen "Comparative analysis of optical character recognition methods for Sámi texts from the National Library of Norway".
Lenke til preprint: https://arxiv.org/abs/2501.07300

Samlingen består av tre forskjellige typer modeller: Transkribus-modeller, Tesseract-modeller og TrOCR-modeller.

Transkribus-modeller

Transkribusmodellene er tilgjengelige i applikasjonen Transkribus.
Transkribus er et verktøy hvor du kan bruke modeller til å gjenkjenne tekst i bilder, og korrigere eventuelle feilprediksjoner.
Det kan koste penger å bruke Transkribus, avhengig av mengden data som skal transkriberes.

Følgende modeller er tilgjengelige:

- SamiskOCR_smi_ub (modell-id 181605): en modell trent utelukkende på manuelt annotert samisk data
- SamiskOCR_smi (modell-id 181725): print M1-basemodellen (modell-id 39995) fin-tunet på manuelt annotert samisk data
- SamiskOCR_smi_nor (modell-id 182005): print M1-basemodellen fin-tunet på manuelt annotert samisk og norsk data
- SamiskOCR_smi_smipred (modell-id 192137): print M1-basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data
- SamiskOCR_alt (modell-id 179305): print M1-basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data, og manuelt annotert norsk data

Sistnevnte modell er den som fikk best resultater på vårt testdatasett.

Tesseract-modeller

Tesseract er et OCR-verktøy hvor du kan kjøre modeller lokalt.
Se installasjonsguiden for å installere Tesseract.
Så snart du har installert Tesseract, kan du bruke .traineddata-modellfilene i `tesseract_models/` for å kjøre OCR på bildene dine.

Følgende modeller er tilgjengelige:

- ub_smi: en modell trent utelukkende på manuelt annotert samisk data
- smi: den norske basemodellen fin-tunet på manuelt annotert samisk data
- smi_nor: den norske basemodellen fin-tunet på manuelt annotert samisk og norsk data

- smi_pred: den norske basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data
- smi_nor_pred: den norske basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data, og manuelt annotert norsk data
- synth_base: den norske basemodellen fin-tunet på syntetisk* samisk data
- sb_smi: synth_base fin-tunet på manuelt annotert samisk data
- sb_smi_nor_pred: synth_base fin-tunet på manuelt annotert og automatisk transkribert samisk data, og manuelt annotert norsk data

Sistnevnte modell er den som fikk best resultater (av tesseract-modellene) på vårt testdatasett

* syntetisk her betyr at vi har ekte samisk tekst, som vi kan laget bilder av, som skal ligne på scannet tekst. I motsetning til den manuelt transkriberte dataen, som er bøker og aviser som er scannet, og deretter manuelt transkribert.

Tesseract-tips:

- Du må flytte .traineddata-filene til tessdata-området på PC-en din
- For å kjøre OCR med ønsket modell må du spesifisere modellnavnet med -l, eks: `tesseract filnavn.jpg utfil -l smi_nor_pred`. Da vil `utfil.txt` inneholde teksten som modellen fant i bildet
- Bruk --psm 7 hvis du har bilde av en linje med tekst, og ikke en hel side. Se mer info her

TrOCR-modeller

TrOCR er en transformers-basert modellarkitektur for tekstgjenkjenning.
Alle våre TrOCR-modeller ligger på Språkbankens side på huggingface og i `trocr_models/` her
Du kan bruke modellene med transformers-biblioteket

Følgende modeller er tilgjengelige:

- trocr_smi: TrOCR-printed-basemodellen fin-tunet på manuelt annotert samisk data
- trocr_smi_nor: TrOCR-printed-basemodellen fin-tunet on på manuelt annotert samisk og norsk data
- trocr_smi_pred: TrOCR-printed-basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data
- trocr_smi_nor_pred: TrOCR-printed-basemodellen fin-tunet på manuelt annotert og automatisk transkribert samisk data, og manuelt annotert norsk data
- trocr_smi_synth: TrOCR-printed-basemodellen fin-tunet på syntetisk* samisk data, og deretter på manuelt annotert samisk data
- trocr_smi_pred_synth: TrOCR-printed-basemodellen fin-tunet på syntetisk samisk data, og deretter fin-tunet på manuelt annotert og automatisk transkribert samisk data
- trocr_smi_nor_pred_synth: TrOCR-printed-basemodellen fin-tunet på syntetisk samisk data, og deretter fin-tunet på manuelt annotert og automatisk transkribert samisk data, og manuelt annotert norsk

trocr_smi_pred_synth er modellen som fikk best resultater (av TrOCR-modellene) på vårt testdatasett

* syntetisk her betyr at vi har ekte samisk tekst, som vi kan laget bilder av, som skal ligne på scannet tekst. I motsetning til den manuelt transkriberte dataen, som er bøker og aviser som er scannet, og deretter manuelt transkribert.

Modellene fungerer kun med bilder av linjer av tekst. Om du har bilder av hele sider av tekst, må du dermed segmentere teksten i linjer først, for å få nytte av denne modellen.

Brukseksempel med python

```
from transformers import TrOCRProcessor, VisionEncoderDecoderModel
from PIL import Image

# ENTEN
# last ned modellene fra internett
processor =
TrOCRProcessor.from_pretrained("Sprakbanken/trocr_smi_pred_synth")
model =
VisionEncoderDecoderModel.from_pretrained("Sprakbanken/trocr_smi_pred_synth
")

# ELLER
# last inn lokalt lagret modell
processor = TrOCRProcessor.from_pretrained("
<relativ_sti_til>/trocr_models/trocr_smi_pred_synth")
model = VisionEncoderDecoderModel.from_pretrained("
<relativ_sti_til>/trocr_models/trocr_smi_pred_synth")

image = Image.open("path_to_image.jpg").convert("RGB")

pixel_values = processor(image, return_tensors="pt").pixel_values
generated_ids = model.generate(pixel_values)

generated_text = processor.batch_decode(generated_ids,
skip_special_tokens=True)[0]
```

Lisens, bruk og sitering

Modellene i samlingen deles under lisensen CC BY 4.0 (lenke: https://creativecommons.org/licenses/by/4.0/deed.no)

Du kan fritt gjenbruke og modifisere modellene, men skal sitere artikkelen:
"Enstad T, Trosterud T, Røsok MI, Beyer Y, Roald M. Comparative analysis of optical character recognition methods for Sámi texts from the National Library of Norway. Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa) 2025."