

TEKNOLOGI / 4 MINUTTER / 12. JANUAR 2024

# Språkmodeller som ser: GPT-4V og CogVLM

Mange kjenner til mulighetene med ChatGPT. Veldig mange har også testet den. I siste halvdel av 2023 lanserte OpenAI GPT-4V med visuelle funksjoner (GPT-4V), noe som gjorde det mulig å stille *bilde-relaterte spørsmål*; for eksempel: "Hvor mange epler er det på bildet?" eller: "Hva står det på handlelisten?"

Men hvordan fungerer denne modellen i vanlige situasjoner? Og hva med eventuelle konkurrenter? Vi har testet.



Kristoffer Jan Zieba  
Data Scientist

GPT-4V skal klare å håndtere enkle visuelle egenskaper som tidligere kun var mulig med dedikerte datavisjonsmodeller. GPT-4V har en stor fordel sammenlignet med disse modellene – den kan kommunisere med brukeren gjennom naturlig språk, noe som åpner for store muligheter i ulike sektorer og bedrifter.

GPT-4V er imidlertid ikke den eneste modellen som kan håndtere både tekst- og bilde-relaterte spørsmål. Det finnes andre, som Googles Gemini og Alibabas Qwen-VL. Det er også en rekke open-source-modeller tilgjengelig, for eksempel CogVLM.

En betydelig fordel med CogVLM og visse andre open-source-modeller, er evnen til å bli trent i et lukket miljø, for eksempel på egen maskinvare eller i eget skymiljø. Dette er en mulighet som ikke er tilgjengelig for GPT-modeller fra OpenAI. En slik tilnærming eliminerer risikoen for datamisbruk eller lekkasjer. Videre kan disse modellene deployeres lokalt og brukes på enheter uten internettforbindelse, noe som kan være svært nyttig for brukere i områder med begrenset nettdækning. I motsetning til modeller fra OpenAI, unngår man også kostnader forbundet med bruk og tilpasning av modellen til spesifikke bruksområder. En utfordring med noen open-source modeller er imidlertid at de kan være mer krevende å tilpasse og implementere i praksis.

## En typisk arbeidsdag på Kantega

I denne bloggen vil vi utforske hvordan GPT-4V og CogVLM presterer på visuelle oppgaver relatert til en "typisk" arbeidsdag på Kantegas kontor i Trondheim. Vi vil demonstrere noen av de mange mulighetene som språkmodeller med syn tilbyr. Merk at oppgavene vi presenterer er helt fiktive og kun til illustrasjon – de gjenspeiler ikke våre faktiske oppgaver 😊

For å gjøre testene mer realistiske, simulerer vi et scenario der vi har et nettverk av kameraer plassert både inne i og utenfor kontoret. Disse kameraene sender livebilder til modellene (både GPT-4V og CogVLM), slik at de kontinuerlig har tilgang til aktuelle bilder.

## Før vi kommer oss på kontoret

*Hvordan komme seg på jobb i dag, da?*



(Det tenkte) kameraet som er plassert utenfor Bassengbakken, der Kantega holder til, sender dette bildeutsnittet av parkeringsplassen til språkmodellene.

Det er tidlig desembermorgen, og på grunn av dårlig vær vurderer vi hvilken transportmåte vi skal bruke til kontoret. Er det greit å gå, ta bussen, eller skal vi være litt late i dag og kjøre bil? Heldigvis kan vi spørre våre modeller, som har tilgang til live bilder av gaten rundt kontoret vårt mens vi sitter hjemme.

*Vi spør: "Er det vanskelig å gå på gata?" Og får følgende svar:*

Vi spør: "Er det vanskelig å gå på gata?" Og får følgende svar:

**CogVLM:** "Ja, det er vanskelig å gå på gata. Det er lett ut, men det kan være en smule bit slippery etter en snøfall."

**GPT-4V:** "Det ser ut som gaten på bildet kan være vanskelig å gå på. Det er snø på veien og fortauet, og det ser ut til å være litt isete, noe som ofte kan gjøre det glatt og utfordrende å gå uten å skli. Pass på å ha sko med godt grep og være forsiktig når du går på slike overflater."

Vi skjønner at det kan gå fint likevel. Så spør vi videre: "Trenger fortauet brøyting?"

**CogVLM:** "Ja, det ser ut til at det er nødvendig å brøyte fortauet på gatehjørnet for å gi biler muligheten til å kjøre fremover."

**GPT-4V:** "Ja, ut fra bildet ser det ut som fortauet kan trenge brøyting. Snøen dekker overflaten og kan gjøre det vanskelig for fotgjengere å gå. Brøyting ville gjøre det tryggere og enklere å ferdes til fots."

Nå er vi klar over at det blir vanskelig å gå. Så vi vurderer å kjøre bil eller

Deretter spør vi: "Hvor mange ledige parkeringsplasser finnes det på bildet?"

**CogVLM:** "Det er tre ledige parkeringsplasser på bildet."

**GPT-4V:** "Det er tre ledige parkeringsplasser synlige på bildet: en rett foran bygningen til venstre og to mot høyre side av bildet, nær der veien svinger."

Det er flott at det fortsatt finnes noen ledige parkeringsplasser. Vi vil også vite om det er mulig å parkere en sykkel, så vi spør: "Finnes det ledige plasser for å parkere en sykkel?"

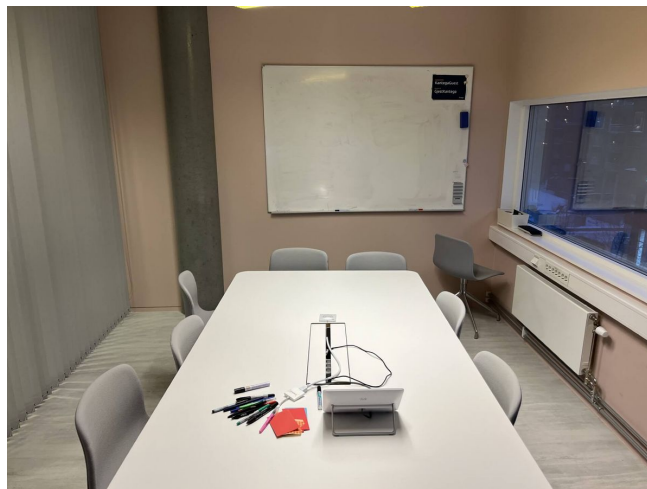
**CogVLM:** "Ja, det ser ut til at det er ledige plass for å parkere en sykkel. Det er en sykkelstang på gatehjørnet."

**GPT-4V:** "Ja, det ser ut til å være ledig plass ved sykkelstativene foran bygningen til venstre på bildet. Stativene ser ikke fullt utnyttet ut, så det burde være plass til å parkere minst én sykkel der."

Så det blir sykling til jobb i dag!

## Første møte

### Er møterommet presentabelt?



(Det tenkte) kameraet på møterommet sender dette bildeutsnittet av rommet til språkmodellene

Vi sitter fortsatt hjemme før avreise til jobb. Dagen starter med en workshop med en kunde. Vi vet at det ble holdt et arrangement i rommet i går ettermiddag, så vi lurar på om rommet er klart for

workshop eller om vi må forberede rommet og reise noen minutter tidligere.

Heldigvis kan vi bruke bildestrøm fra møterommet og spørre modellene: "Er rommet ryddig?"

**CogVLM:** "Ja, rommet ser ut til å være ryddig."

**GPT-4V:** "Ja, rommet ser ryddig ut. Det er en ren tavle på den ene vegg og en organisert utstilling av markører, blyanter, og andre kontorrekvisita på møtebordet. Stolene er pent plassert rundt bordet, og det ser ikke ut til å være rot eller uorden i rommet."

Vi trenger noen workshop-materialer, så vi sjekker: "Finnes det post-it lapper i rommet?"

**CogVLM:** "Ja, det er post-it lapper på skrivbordtavlen."

**GPT-4V:** "Ja, det ser ut til å være post-it lapper på bordet nær hjørnet nærmest kameraet. De er delvis dekket av pennen og markørene som er plassert over dem."

Flott, da trenger vi ikke dra tidligere til jobb.

## Lunch

Er det nok ketchup til pommesen?



Til og med kamera i kjøleskapet, gitt! Som sender dette (tenkte) bildeutsnittet til språkmodellene

Kantineansvarlig Berit er syk, og vi har blitt bedt om å steppe inn. Men har vi nok til å servere varmmat med pommes frites i dag? Heldigvis har vi et kjøleskap med kamera 😊.

Uten å gå til kantina chatter vi med modellene: "Hvor mange flasker ketchup finnes i kjøleskapet?"

**CogVLM:** "Det finnes tre flasker ketchup i kjøleskapet."

**GPT-4V:** "Det finnes tre flasker med ketchup i kjøleskapet."

Som uerfaren kantineansvarlig må vi spørre:

"Blir det nok ketchup for lunch til 86 ansatte som skal spise pommes frites?"

**CogVLM:** "3 flasker ketchup er ikke nok for 86 ansatte som skal spise pommes frites, så flere flasker trenger å kjøpes for lunch."

**GPT-4V:** "Det avhenger av hvor mye ketchup hver ansatt vanligvis bruker og hvor store flaskene er. Generelt bruker en person ca. 15-20 gram ketchup per porsjon pommes frites. Dersom vi antar at hver flaske inneholder 500 gram ketchup, vil tre flasker inneholde totalt 1500 gram."

For å beregne om dette er nok for 86 ansatte, kan vi gjøre en rask beregning:

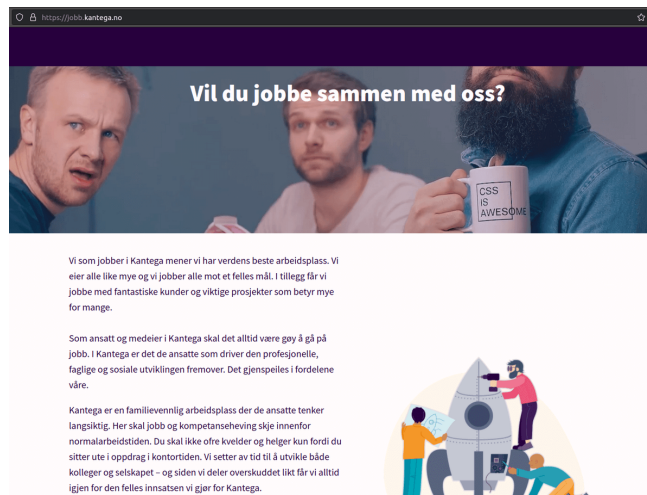
$86 \text{ ansatte} \times 20 \text{ gram per ansatt} = 1720 \text{ gram ketchup}$   
 $86 \text{ ansatte} \times 20 \text{ gram per ansatt} = 1720 \text{ gram ketchup}$

Dersom hver ansatt bruker 20 gram, vil man trenge 1720 gram ketchup. Med bare 1500 gram tilgjengelig, vil det ikke være nok for alle 86 ansatte til å ha 20 gram hver. Man ville trenge minst en flaske til for å dekke dette behovet, forutsatt at alle bruker 20 gram og at hver flaske faktisk inneholder 500 gram. Hvis de bruker mindre, eller flaskene er større, kan det kanskje holde."

Det var et veldig detaljert svar. Men modellene anbefaler oss forskjellige ting, så vi er litt forvirret. Hvem skal vi stole på? OK, vi tar en kjapp handleturn for sikkerhetsskyld 😊

## Jobbe med nettside

Har du noen tips til innholdet vårt?



Skjerm bilde fra Kantegas stillingsannonse-side. Dette sendes til språkmodellene.

I dag skal vi friske opp hovedsiden for ledige Kantega-jobber. Idéen er å forkorte den lange teksten.

Vi åpner en nettleser, tar et skjerm bilde og spør modellene: "Kan du lage en kort oppsummering av tekst på websida?"

**CogVLM:** "Ja, vil du jobbe sammen med oss? Vi har en avslaplig arbeidslagsplass med god arbeidsmiljø, muligheter til personlig utvikling og tilfredsstillende fordeler."

**GPT-4V:** "Siden du har sendt meg et skjerm bilde av en nettside, kan jeg dessverre ikke lese eller analysere tekst direkte fra bildet. Imidlertid, hvis du kan gi meg teksten som ren tekst eller be om spesifikke deler av nettsiden du ønsker oppsummert, kan jeg gi deg en oppsummering basert på det."

Så synd at GPT-4V ikke kan gi oss en oppsummering. Vi spør CogVLM: "Hva ville du endret på websida?"

**CogVLM:** 1. "Inkluder en mer personlig og engasjerende tone i teksten. 2. Fokuser mer på de fordeler som jobbene tilbyr, som for eksempel tilfredsstillende arbeidsmiljø og muligheter for personlig utvikling. 3. Fjern de tidligere referanser til selskapets størrelse og utvikling, og fokuser mer på hva de tilbyr til sine ansatte."

Takk for tilbakemelding CogVLM 😊 Men tja, vet ikke helt hvor godt det der traff.

## Tid for selvutvikling

Hvilke bøker har vi egentlig i biblioteket på jobb?



Hva et tenkt kamera, som filmer mot bokhylla på biblioteket vårt, kan sende til språkmodellene

I Kantega bruker vi en del av arbeidstiden til egenutvikling. Etter en eventyrlig dag på jobb tenkte vi å bruke noen timer på slutten av dagen til å lese. Heldigvis filmer et kamera en bokhylle fylt med bøker.

Bare av nysgjerrighet spør jeg: "Hvor mange bøker finnes det på bildet?" (vi telte at det er 18 stykker).

**CogVLM:** "Det er 19 bøker på bildet."

**GPT-4V:** "Det er 24 synlige bokrygger på bildet."

I det siste har vi vært nysgjerrige på psykologi og visste at det fantes en bok i det fagfeltet. Så spør jeg: "Hva er boktittel av en bok om psychology som ligger på bokhylla?"

**CogVLM:** "En av bøkene som ligger på bokhylla er "Social Psychology" av David Myers."

**GPT-4V:** "Boktittelen som omhandler psykologi på bokhyllen er "Social Psychology".

## Hva har vi lært?

Både GPT-4V og CogVLM presterte godt på komplekse oppgaver som krever forståelse av bilder, og de kunne til og med gi anbefalinger på et nivå som er passende for voksne.

Begge modellene kommuniserer effektivt med brukeren. GPT-4V viste seg å ha litt bedre kunnskap i norsk, mens CogVLMs norsk virket noe mer kronglete. Ofte svarte de på engelsk, så vi måtte be dem om å svare på norsk innimellom.

Vi observerte at CogVLM generelt er bedre enn GPT-4V på typiske oppgaver innen maskinsyn. For eksempel klarte ikke GPT-4V å gjenkjenne tekst i et skjermbilde og gjorde noen feil i å telle bøker. Dette stemmer overens med funn i andre tester (<https://blog.roboflow.com/gpt-4-vision-alternatives/>). Men, GPT-4V virker å være mer avansert når det gjelder å argumentere, som vi kan se



i svaret på spørsmålet om tre flasker ketchup er nok for 86 personer.

Språkmodeller med visuelle evner, ofte kalt multimodale modeller, åpner for spennende muligheter i flere bransjer. Her er noen eksempler:

- **Helsevesenet:** Multimodale språkmodeller kan bidra til å analysere medisinske bilder for diagnostikk, og forklare komplekse medisinske forhold på en enkel måte for pasienter.
- **Utdanning:** Disse modellene kan forbedre interaktiv læring ved å forklare og analysere visuelle læringsmaterialer, og tilpasse læringsinnhold basert på visuelle innganger og studentrespons.
- **Markedsføring og reklame:** Ved å analysere kundenes visuelle innhold og preferanser, kan modellene hjelpe til med å skape målrettet og personlig markedsføringsinnhold.
- **Kundeservice:** Bruk av visuelle språkmodeller i kundeservice kan øke effektiviteten ved å tolke og svare på kundeforhør som inkluderer bilder eller videoer.

### Lenker:

Offisiell CogVLM repo: <https://github.com/THUDM/CogVLM>

GPT-4V: <https://platform.openai.com/docs/guides/vision>

*Kantega*

Tjenester

Hvorfor velge oss

Prosjekter


Ansatteierskap 


[Bloggen](#)

[Dette er oss](#)

[Ta kontakt](#)

[Nyheter](#)

[Sommerjobb](#) 

[Jobb i Kantega](#) 

[Åpenhetsloven](#)

[Bærekraft](#)