EXAMENSARBETE Image Captioning Using Multimodal LLMs
STUDENT Nathalie Tiet
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Hur bra är multimodala språkmodeller på att generera bildtexter?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Nathalie Tiet**

Utvecklingen av multimodala språkmodeller (MLLMs) har förbättrat kvaliteten inom bildtextgenerering. I detta arbete undersöker vi prestanda samt skillnader i fokus mellan olika MLLMs, det vill säga vad modellerna väljer att lyfta fram i en bild när de beskriver den.

I takt med att stora språkmodeller blivit mer och mer avancerade har också utvecklingen inom bildtextgenerering gjort stora framsteg. Det är ett område som har många användningsområden. Ett exempel är inom e-handel där bildtextgenerering möjliggör automatiska produktbeskrivningar. Ett annat är generation av metadata till produktkataloger, vilket skulle kunna optimera sökning och lagring av data. För synskadade kan bildtextgenerering hjälpa genom att generera beskrivningar av omgivningen som sedan läses upp, vilket gör det enklare för människor att navigera sin omgivning. Slutligen kan det även användas inom medicinsk diagnostik som ett stöd för att analysera och förstå medicinska bilder.

I mitt examensarbete har jag undersökt skillnader i prestanda och samt fokus hos fem olika modeller, **BLIP-2**, **GIT**, **LLaVA**, **Qwen2-VL** och **Gemini**. Jag har utvärderat modellerna med hjälp av automatiserade mått samt gjort en mänsklig utvärdering. För jämföra fokus hos modellerna användes POS-taggning samt NRC Emotion Lexicon.

Resultaten visar att BLIP-2 presterar bäst enligt automatiserade mått. Däremot visar den mänskliga utvärderingen att Gemini rankas högst av människor, mycket på grund av hur detaljerad

BLIP-2: *A group of people sitting on a hill.*
GIT: *A group of people standing on top of a mountain.*
LLaVA: *A group of people are standing on a cliff overlooking a valley.*
Qwen2-VL: *Hikers ascend a mountain trail under a colorful sky.*
Gemini: *Tourists enjoy the sunset view from a cliff in the Atacama Desert.*

Reference: *A few people standing on top of a mountain taking pictures.*

Figure 1: Exempelbild från datasetet Flickr8k och fem bildtexter från modellerna samt en referens bildtext från datasetet.

modellen är. Detta tydliggör en diskrepans mellan vad de automatiserade måtten indikerar och hur människor faktiskt uppfattar kvaliteten på bildtexter. Vi ser att automatiserade mått ofta har svårt för att ge en rättvis bedömning av längre och mer detaljerade bildtexter som ofta avviker från de korta och koncisa referenserna.

Vidare visar resultaten att BLIP-2 och GIT fokuserar mer på objekten i en bild när de ska ge en bekrivning medan Gemini and Qwen2-VL är bättre på att även inkludera beskrivningar av atmosfär och stämning. De två sistnämnda modellerna utmärker sig även genom deras förmåga att fånga detaljer, då de exempelvis besitter förmågan att identifiera hundraser och salladssorter.