Multimodální AI

Multimodální AI je typ systému umělé inteligence, který dokáže zpracovávat a propojovat informace z různých typů datových zdrojů, neboli **modalit**. Mezi tyto modality patří text, obrázky, video, zvuk, řeč nebo dokonce data ze senzorů.

Namísto toho, aby model viděl svět jen jako posloupnost písmen, dokáže „vidět“ obrázek a „slyšet“ tón hlasu, přičemž chápe vztahy mezi těmito vstupy.

Jak multimodální AI funguje?

Základem je schopnost převést různé typy dat do společného matematického prostoru (tzv. **embeddingy**).

- **Kodéry (Encoders):** Každá modalita má svůj vlastní „přijímač“ (např. Vision Transformer pro obrázky).
- **Fúze (Fusion):** Systém spojí informace z různých kodérů do jednoho celku.
- **Dekodér (Decoder):** Na základě pochopeného kontextu vygeneruje odpověď (textovou, obrazovou či zvukovou).

Srovnání: Unimodální vs. Multimodální AI

Vlastnost	Unimodální AI (např. GPT-3)	Multimodální AI (např. GPT-4o, Gemini)
Vstupy	Pouze text	Text, Foto, Audio, Video
Pochopení	Pouze sémantika slov	Kontext, vizuální detaily, emoce v hlase
Výstup	Text	Text, Obrázek, Mluvené slovo
Příklad	Chatbot na webu	Asistent, kterému ukážete rozbitý motor a on vám řekne, co opravit

Hlavní modality a jejich využití

- **Text + Obrázek:** Analýza rentgenových snímků s popisem diagnózy, nebo generování obrázků z textu (DALL-E, Midjourney).
- **Text + Audio:** Přepis řeči s pochopením sarkasmu nebo emocí, okamžitý překlad mluveného slova.
- **Video + Text:** Automatické vytváření titulků nebo vyhledávání konkrétních momentů ve videu („Najdi část, kde pes skáče do bazénu“).

Současní lídři na trhu

```
1. **OpenAI (GPT-4o):** "Omni" model, který reaguje v reálném čase na hlas i video.
2. **Google (Gemini 1.5 Pro):** Model s obrovským kontextovým oknem, schopný
```

```
analyzovat hodinová videa najednou.
3. **Anthropic (Claude 3.5 Sonnet):** Špičkový model v analýze grafů,
schémat a vizuálního programování.
```

—

Proč je to důležité?

> Multimodalita je klíčem k dosažení **AGI** (obecné umělé inteligence). Aby AI mohla skutečně pomáhat v reálném světě (např. v robotech), musí být schopna vnímat prostor a zvuk stejně přirozeně jako textové instrukce.

Viz také: LLM, Computer Vision, Neural Networks