Topic 22: Vision-Language Models (VLMs)

1. IBM's introduction to vision-language models, training techniques and use cases: https://www.ibm.com/think/topics/vision-language-models
2. Overview of how large AI models combine vision and language: https://www.twelvelabs.io/blog/foundation-models-are-going-multimodal
3. An Introduction to Vision-Language Modeling: Tutorial-style paper explaining VLMs: how they work, how to evaluate them: https://arxiv.org/html/2405.17247v1
4. Generalized Visual Language Models: Blog post exploring how pretrained LMs are extended with vision: https://lilianweng.github.io/posts/2022-06-09-vlm/
5. VLM Prompt Engineering Guide: https://www.edge-ai-vision.com/2025/03/vision-language-model-prompt-engineering-guide-for-image-and-video-understanding/
6. Hugging Face blog introducing VLMs, open-source models, and how to use them: https://huggingface.co/blog/vlms
7. Hugging Face blog covering recent trends in multimodal models: any-to-any input, object detection, and counting: https://huggingface.co/blog/vlms-2025
8. Rohit Bandaru's blog covering core components of VLMs (vision encoders, architectures, training recipes): https://rohitbandaru.github.io/blog/Vision-Language-Models/
9. OpenAI blog introducing CLIP, a model that learns visual concepts from natural language supervision: https://openai.com/index/clip/
10. DALL·E: OpenAI blog announcing DALL-E, showing early multimodal generation of images from text. https://openai.com/blog/dall-e/
11. DeepMind blog introducing Flamingo, a few-shot vision-language model handling many tasks: https://deepmind.google/discover/blog/tackling-multiple-tasks-with-a-single-visual-language-model/
12. Roboflow blog discussing the GPT-4 multimodal version (image input) for tasks like OCR, visual puzzles, object detection: https://blog.roboflow.com/gpt-4-vision/
13. Hugging Face community tutorial on building a simple VLM in PyTorch: https://huggingface.co/blog/AviSoori1x/seemore-vision-language-model
14. OpenAI blog showing that CLIP develops neurons that respond to both image and text concepts: https://openai.com/index/multimodal-neurons/
15. OpenCV Live Webinar (Intro to VLMs): https://www.youtube.com/watch?v=trYjGmI_ouk
16. Understanding Multimodal LLMs: how LLMs are extended to handle images: https://magazine.sebastianraschka.com/p/understanding-multimodal-llms