Multimodality definition is blurry, but we care because humans synthesise information. Multimodal is one of the main frontiers for the new foundation modal revolution because ‘we’re running out of text data’. Images are high bandwidth than language - Yan LeCunn. Object tracking is so good now that it ‘feels’ like its real time. Audio can be represented as vision by first representing the audio as Log spectrogram. **Whisper** by Radford. Can extend to video – just need to subsample the key frames **MERLOT** (video, text, audio). Excitement waned for simulated environments (robotics) because it’s too hard to get high quality data – longer term might make a comeback – learn about the world. textTo3D **POINT-E** (2022). Olfactory embeddings (Kiela et al. 2015).

Features and fusion

Why isn’t every system multimodal? Text dominates vision. Don’t have full coverage over modalities – this means that’s it’s not robust to missing data from one modality. It’s hard. Featurizing text is easy but featurizing images is harder. In 2014, CNNs were discovered to replace traditional computer vision techniques. Since 2020, now images are encoded via transformers, where the inputs aren’t tokens as in the case with NLP, but ‘patches’ of an image. There are five broad classes of multimodal fusion: similarity, linear/sum, attention, multiplicative, bilinear. “most the literature on multimodality is about this question: what is the best way to do fusion?” The next question is when in the pipeline to do fusion: early fusion (mix inputs), middle fusion (concatenate features) or late fusion (combine final scores).

The best paper is “IMHO one of the best papers ever written in our field” **CLIP** by Radford et al. 2021 because it generalises better across many of the canonical data sets. It employs a transformer encoder for text, another one for image, and is trained on 300 million image-text pairs. The Radford guy is known for creating high quality datasets. Google implemented the same idea but with 1.8bn image-text pairs (Jia et al. 2021). What the hell. There’s now an open-source group called **Laion** 5bn! that created an image-text pair dataset, which was used to train Stable Diffusion.

There has been lots of experiments with BERT and images. The recommended paper is Bugliarello et al. (2021), which conducted a meta-analysis on all these experiments and concluded that they are all the same ‘pound-for-pound’.

The end goal is one foundational model to ‘rule them all’ **FLAVA** (Singh et al. 2021), which can do many downstream tasks like visual recognition (ImageNet), language understanding (GLUE), multimodal reasoning (VQA). This is where he predicts the field will go in the ‘near future’.

Generative models. Move away from contrastive and discriminative to more rich representation via generative. SOTA is **CoCA** Contrastive Captioner. **Frozen**. **Flamingo** (2022) takes **Chinchilla** which has 80bn – paper has diagram and code next to each other! **BLIP/BLIP2** (2023) this is the future, like ChatGPT but it can see.

**Evaluation**. If you want to make a difference you should work on improve accuracy and reducing bias in deep learning models. Careful evaluation design is a hard skill. Hateful Memes (Kiela 2020) showed that multimodal pretraining doesn’t work, which confirmed the intuition that the everyone had (Bugliarello 2021). **Winoground** stress tests contrastive learning and word order showed that SOTA models performed worse than random chance on this data set.

**Where to next?** One foundation model to rule them all, multimodal scaling laws, retrieval augmented generative multimodal models.

**Language is not all you need.**