

# Deep Learning - MAI

Foundation Models

Dario Garcia Gasulla  
[dario.garcia@bsc.es](mailto:dario.garcia@bsc.es)

# Traditional DL

- ❖ Task oriented (aka supervised learning)
  - Optimize to solve one task
  - Use a set of labeled data <input,desired output> (cost!)
  - Struggles w/ generalization
  - Data selection bias

# Back to the essence of DL

- ❖ Representation Learning
  - Use unlabeled data <input>
  - Optimize to learn the distribution of data
  - No labeling cost, less biases, more reusable
- ❖ Two main learning objectives
  - Also Siamese, GANs, ...

# Masked learning

- ❖ Learn to reconstruct
  - Remove parts of the data
  - Add noise to the data

# Masked for text

- ❖ Next token prediction (GPT) aka “Autoregressive”



Context 

- ❖ Masked language modeling (BERT)



Prediction 

# Masked for images

❖ Patching



❖ Diffusion



# Contrastive Learning

- ❖ Contrastive
  - Learn to contextualize / Spread things properly
  - Transformations which are
    - Not disruptive for the modality
    - Highly disruptive for the data values

# Contrastive learning

❖ Color



❖ Rotation



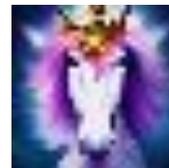
❖ Patching/  
Cropping



❖ Jigsaw



❖ Resolution



# Next step: Multimodality



❖ Image

❖ Text “A majestic unicorn. Front shot. Magic light.”

❖ Sound



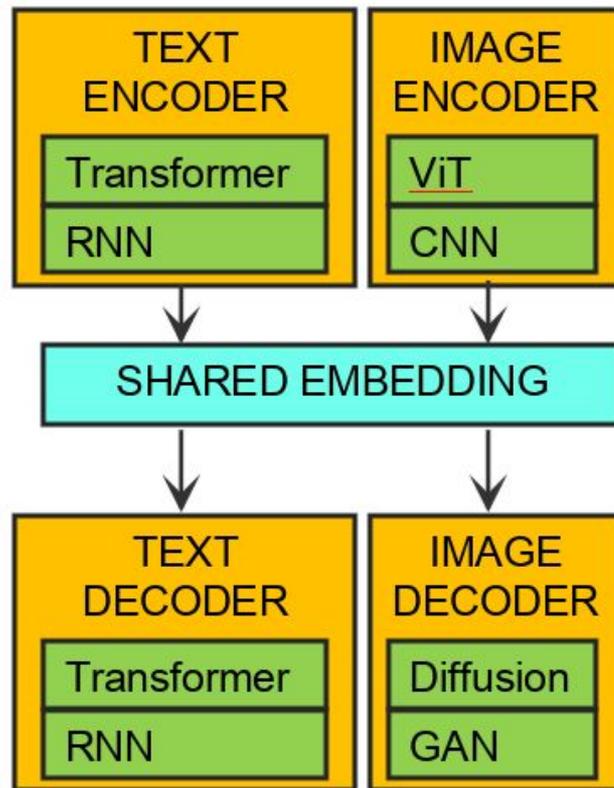
❖ Graph representation, 3D structure, ...

# Foundation models

- ❖ With SSL we can scale to billions of samples
  - Entire data modalities represented
  - Zero-shot generalization
  - No direct supervised application
  - Generative model fest

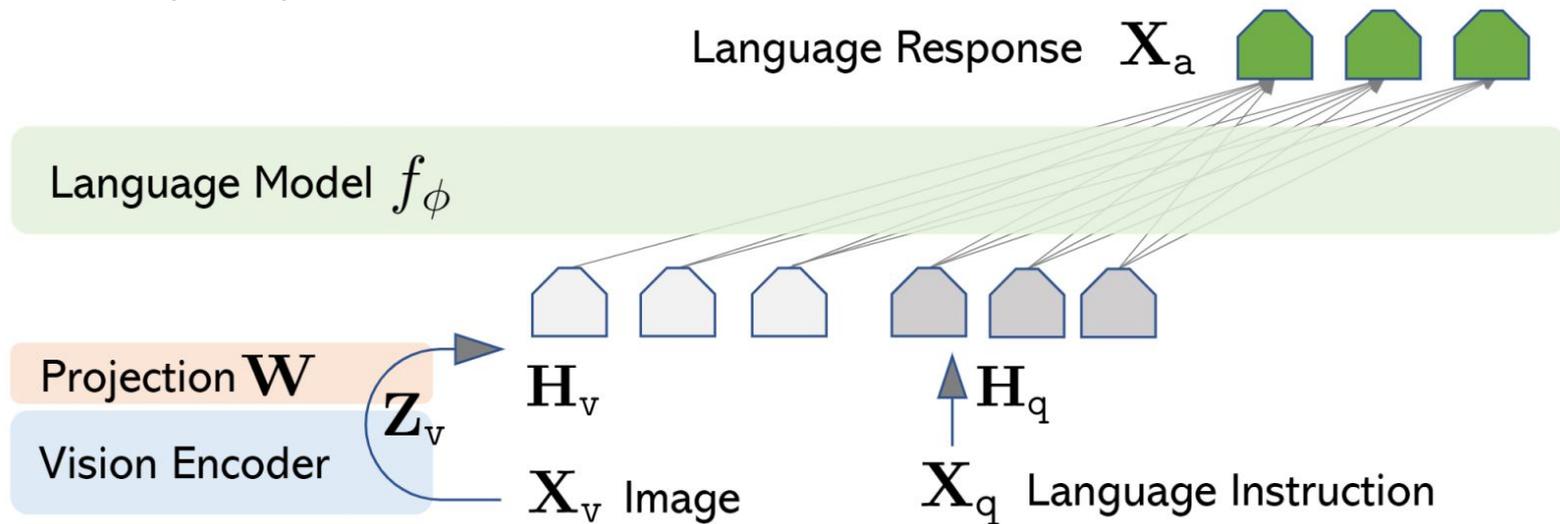
# A simple architecture

- ❖ Data oriented (storage & compute)
- ❖ Multimodal by design
- ❖ In inference, unplug modalities
- ❖ The *shared embedding*
- ❖ LLM as shared embedding
  - Everything as LLM tokens



# A specific example: LLaVA, Shikra

- ❖ Input: text (LLM), image (ViT), locations (LLM)
- ❖ Output: text (LLM)



# Big players

## ❖ **OpenAI** (Microsoft)

- GPT(text), codex (code), whisper (voice), CLIP(txt+img), DALL-E (txt+img)

## ❖ **DeepMind** (Google)

- Gato (agents), Chinchilla (txt), Sparrow (chat), FLAN (txt), LaMBDA, Bard

## ❖ **Meta AI** (Facebook)

- Galactica (papers, retracted), LLaMA (txt), LLaVA

## ❖ **Hugging Face** (USA company)

- Repo & cloud. *BigScience* initiative (Jean Zay - IDRIS/GENCI/CNRS), Bloom (txt)

# Emerging players

- ❖ **stability.ai** (UK company)
  - StableDiffusion (img+txt), open model movement
- ❖  **EleutherAI** (non-profit)
  - The Pile (txt dataset), GPT-J (txt), GPT-Neo (txt)
- ❖ **LAION**  (non-profit)
  - LAION-5B (txt/img dataset), OpenAssistant (ongoing, chat)
- ❖ **moz://a** + Mozilla.ai (non-profit?) Deepspeech, TTS
- ❖ Stanford (HAI, Alpaca), Mistral, DAIR (non-profit), CohereAI, Anthropic, China (Yi, Qwen), UAE (Falcon)...