DATI, AI E ROBOTICA @POLITO

RICERCA, TRASFERIMENTO TECNOLOGICO E SUPPORTO ALLE AZIENDE SUI TEMI FONDAMENTALI DEI BIG DATA, INTELLIGENZA ARTIFICIALE, ROBOTICA E RIVOLUZIONE DIGITALE

CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING FOR THE ITALIAN LANGUAGE

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, Sri Lakshmi

CLIP: CONNECTING IMAGE AND TEXT

Contrastive Language-Image Pre-Training is OpenAI's latest multi-modal model:

- Learning visual concepts with natural language supervision
- **400M** image, text pairs from the web
- Impressive zero-shot capabilities, but...
- ... English only

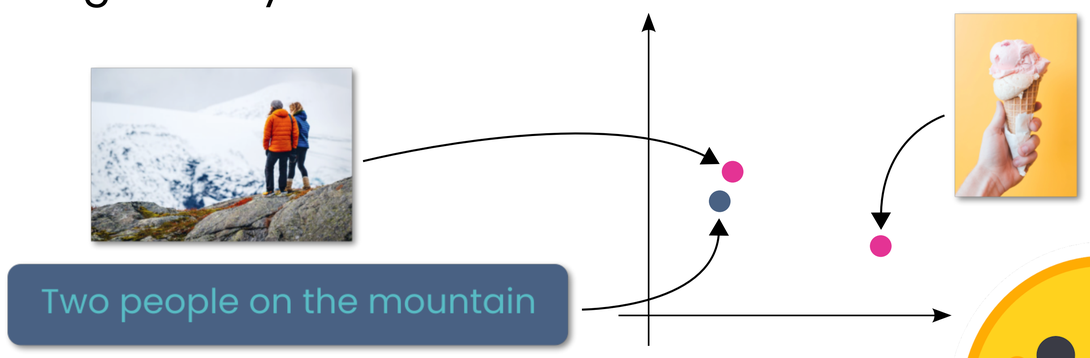


IMAGE RETRIEVAL

Find the "closest" image given a text query

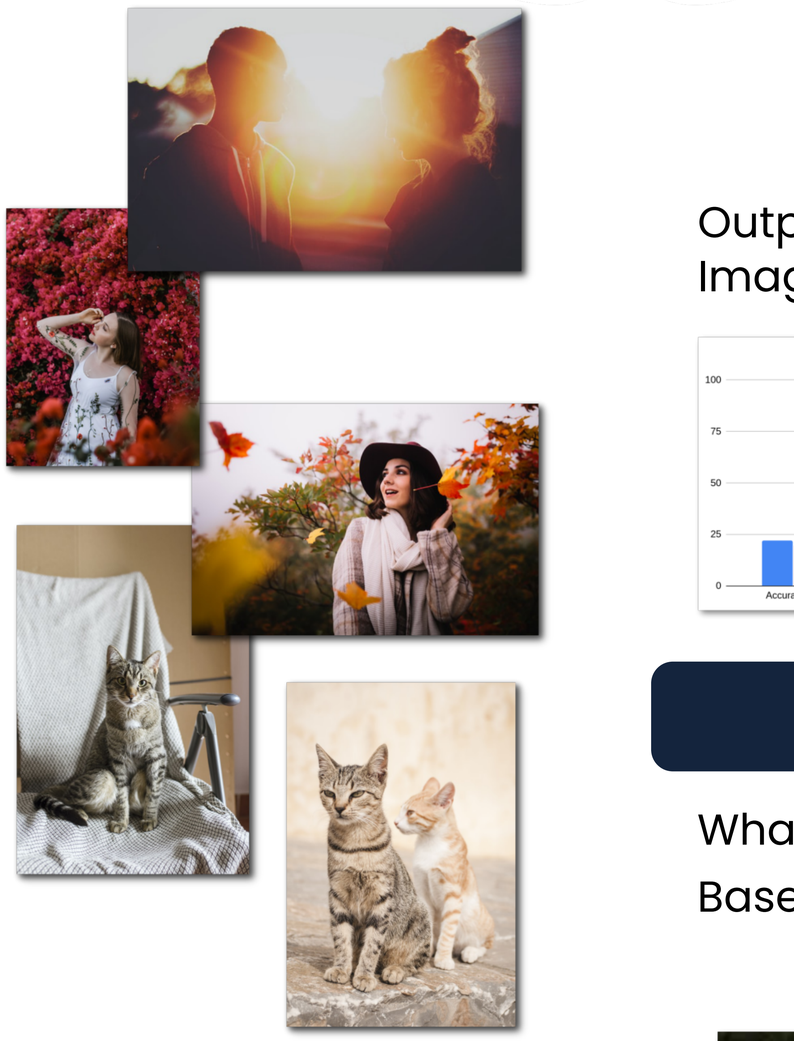
"una coppia al tramonto"

"un vestito primaverile"

"un vestito autunnale"

"un gatto su una sedia"

"due gatti"



ZERO-SHOT CLASSIFICATION

Pick the "closest" text given a query image

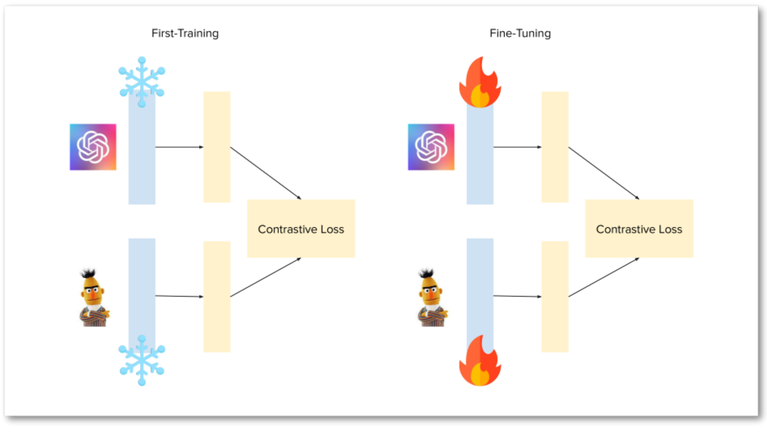
"bandiera dell'Italia"
"bandiera della Spagna"
"un gatto"
"bandiera della Francia"

CLIP ITALIAN

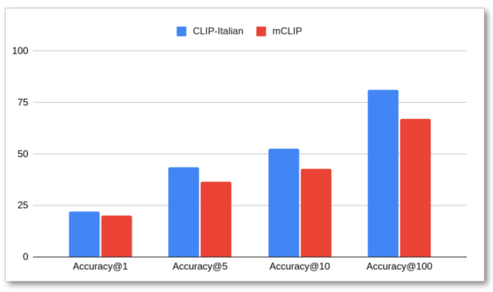
Challenge: bring CLIP to the Italian language
- "low resource" setting

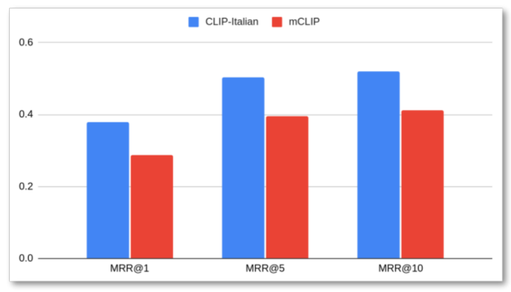
HuggingFace Flax/JAX Community Week
- CLIP JAX implementation
- efficient TPU v3-8 training

Curated datasets and training
- MSCOCO-IT, Google CC, WIT, Il Post
- high quality translation (when needed)
- **1.4M** image-text pairs
- OpenAI's ViT and dbmdz's Italian BERT
- backbone freezing, then unfreeze & fine-tune



Outperforms multilingual CLIP on zero-shot ImageNet classification & IR on MSCOCO-IT Val





LOCALIZATION

What part of the image makes it "close" to text?
Based on image occlusion

PAPER

DEMO

Images: Unsplash 25K dataset