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

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

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”

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

Zero-Shot Classification

Pick the “closest” text given a query image
- “bandiera dell’Italia”
- “bandiera della Spagna”
- “un gatto”
- “bandiera della Francia”
- "uno squalo"
- "un cavallo"

Images: Unsplash 25K dataset