Contrastive Language–Image Pre-training for the Italian Language

Giuseppe Attanasio
Bocconi University (former DBDMG ❤️)

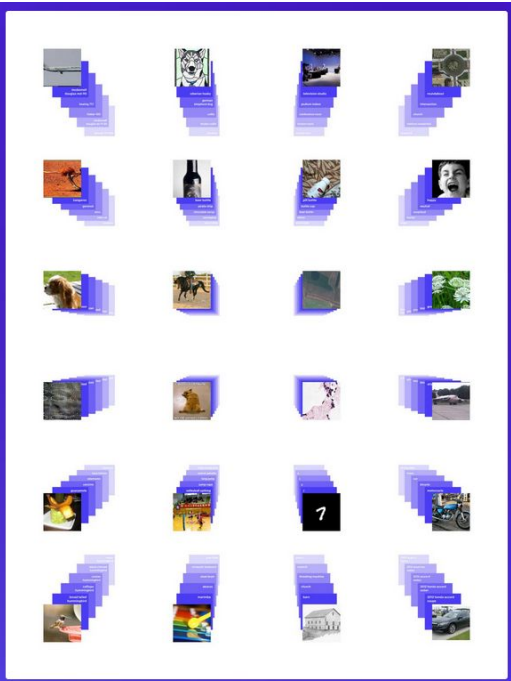
MilaNLP

CLIP: Contrastive Language–Image Pre-training

CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

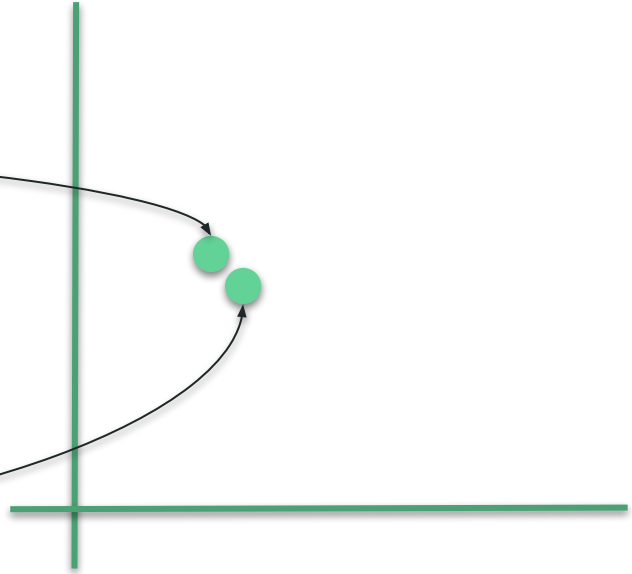
January 5, 2021
15 minute read



images from OpenAI's blog post

CLIP: Intuition

Contrastive Language–Image Pre-training: Intuition

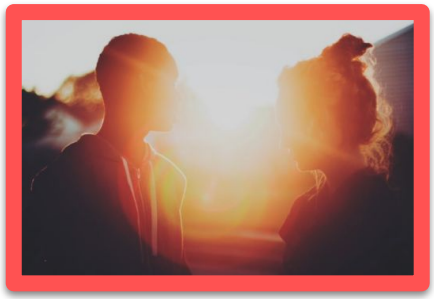


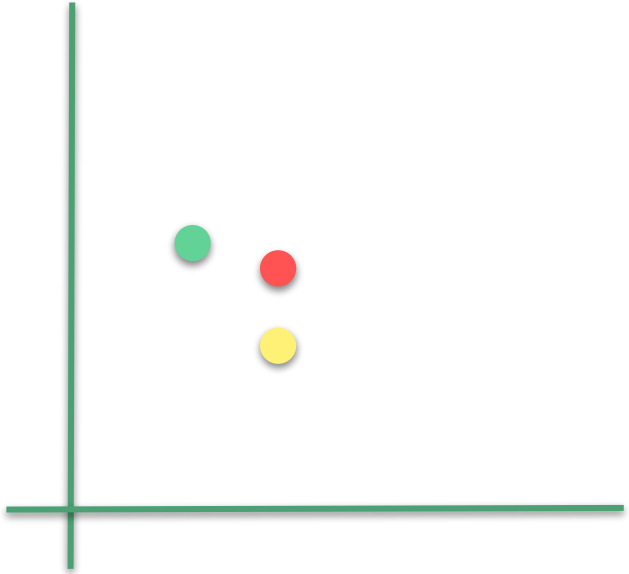
Two people on the mountain

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). **Learning Transferable Visual Models From Natural Language Supervision**. *ICML*.

Contrastive Language–Image Pre-training: Intuition

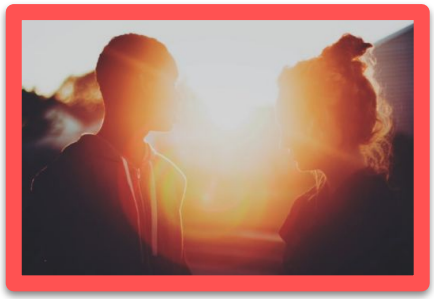
Two people on the mountain

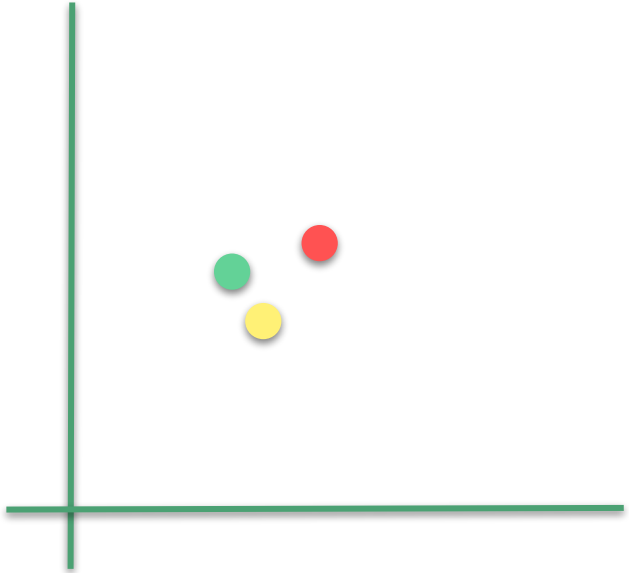




Contrastive Language–Image Pre-training: Intuition

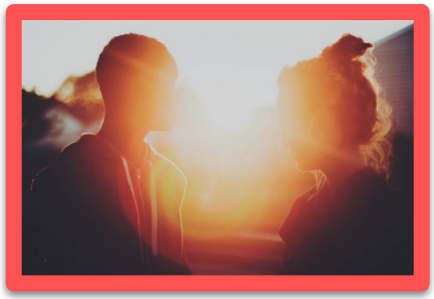
Two people on the mountain

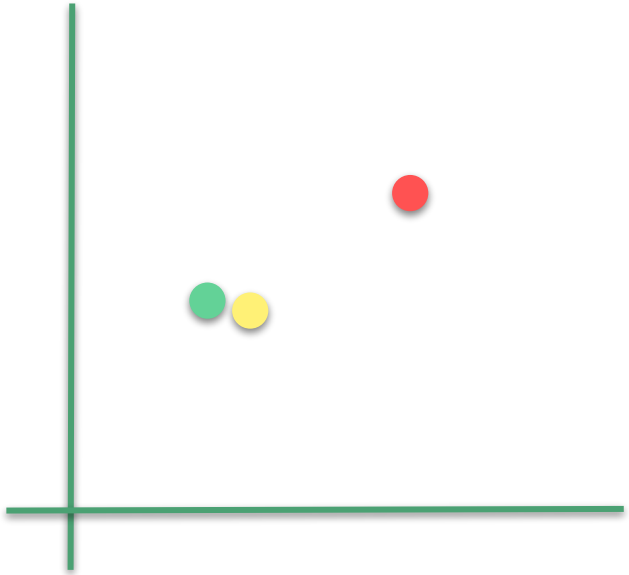




Contrastive Language–Image Pre-training: Intuition

Two people on the mountain





CLIP: Training

Contrastive Language–Image Pre-training: Training

Batches of:

Two people during sunset

Two people on the mountain

Contrastive Language–Image Pre-training: Training

Two people during sunset	1	0
Two people on the mountain	0	1

Ground truth

Contrastive Language–Image Pre-training: Training

similarity matrix

Image Encoder

I_1 I_2

Two people during sunset

Two people on the mountain

Text Encoder

T_1

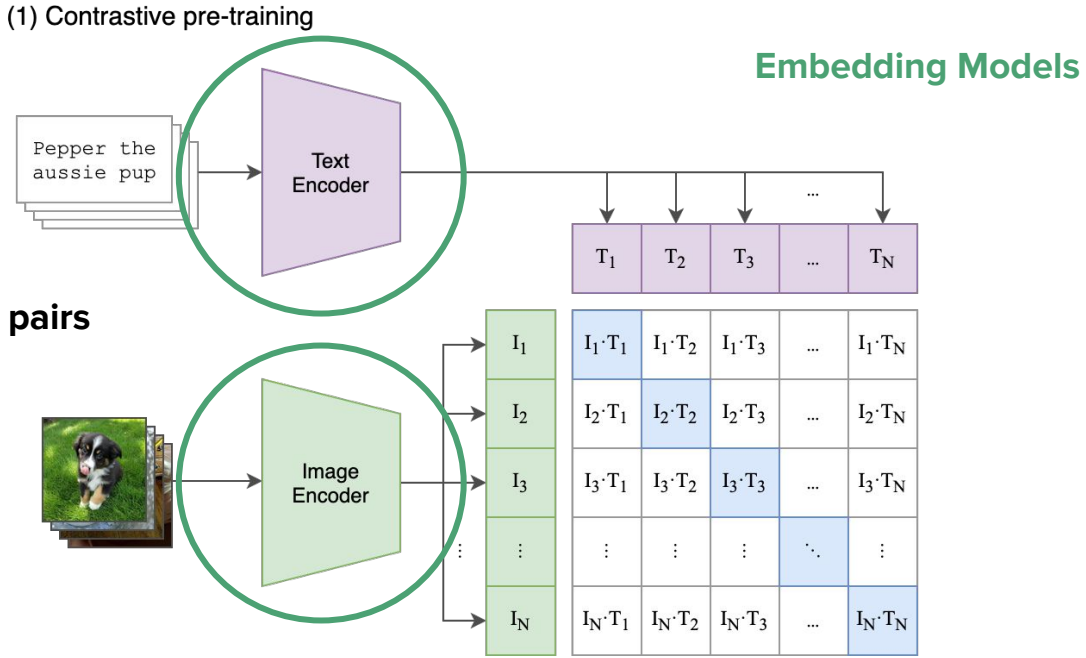
T_2

$I_1 \cdot T_1$ ↑	$I_2 \cdot T_1$ ↓
$I_1 \cdot T_2$ ↓	$I_2 \cdot T_2$ ↑

1	0
0	1

Ground truth

OpenAI’s CLIP

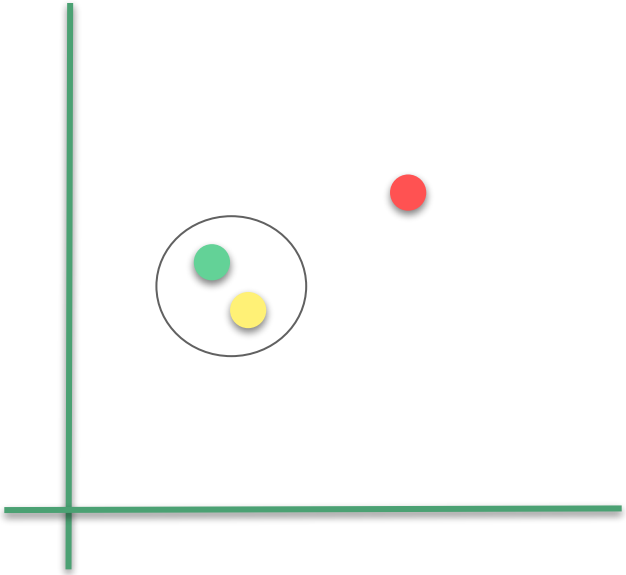


400 millions image-text pairs

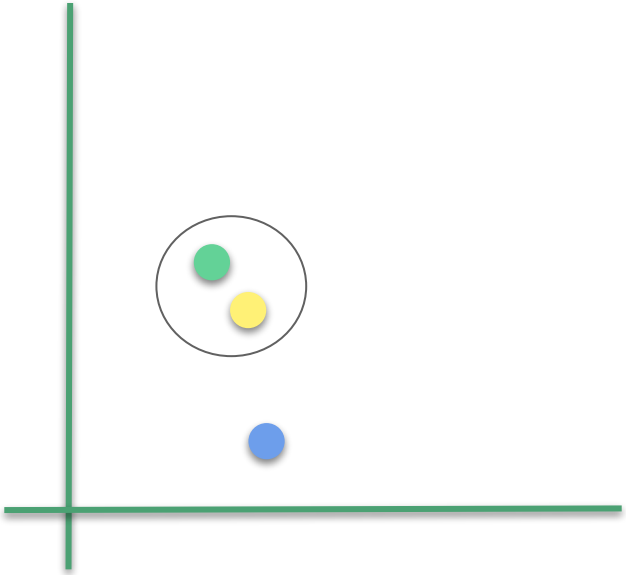
[...] This data is used to create the following proxy training task for CLIP: given an image, predict which out of a set of 32,768 randomly sampled text snippets, was actually paired with it in our dataset.

Using CLIP: Image-Retrieval

Two people on the mountain



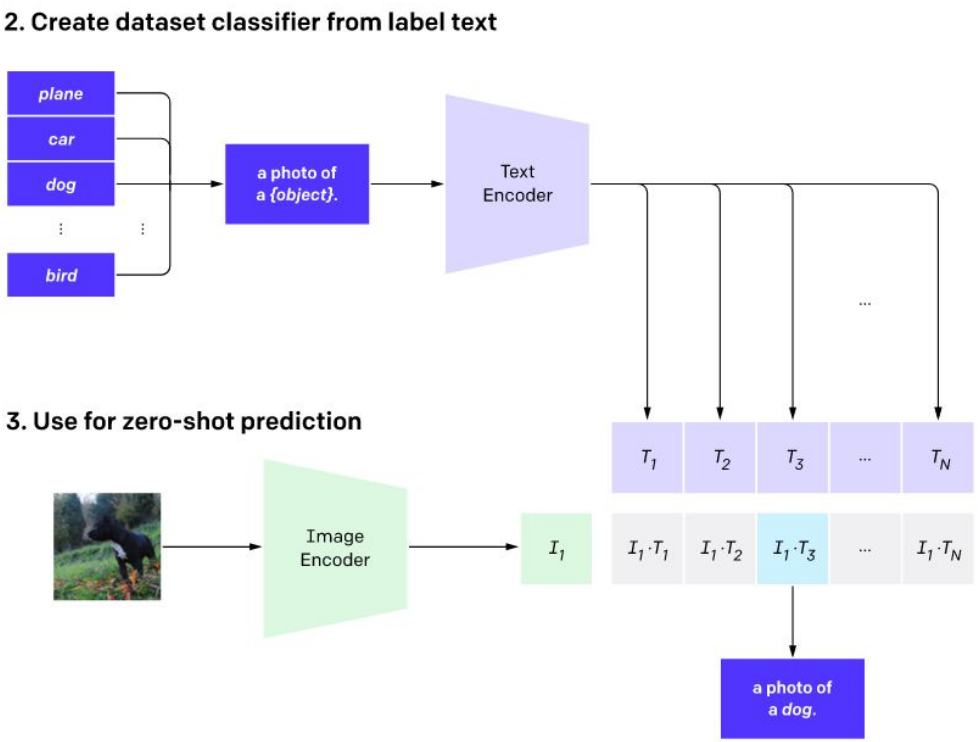
Using CLIP: Zero-Shot Classification



Two people on the mountain

Two cats on a chair

OpenAI’s CLIP



OpenAI’s CLIP: Results

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- × a photo of **ceviche**, a type of food.
- × a photo of **edamame**, a type of food.
- × a photo of **tuna tartare**, a type of food.
- × a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

- ✓ a photo of a **television studio**.
- × a photo of a **podium indoor**.
- × a photo of a **conference room**.
- × a photo of a **lecture room**.
- × a photo of a **control room**.

Contrastive Language–Image Pre-training

```
encoded_images = image_encoder(images)
encoded_texts = text_encoder(texts)
logit_scale = 20

# using the projections
embedded_images = l2_normalization(image_projection(encoded_images))
embedded_texts = l2_normalization(text_projection(encoded_texts))

logits = np.dot(embedded_images, embedded_text.T) * logit_scale

labels = np.arange(n) # correct image-text match is on the diagonal
loss_images = cross_entropy_loss(logits, labels, axis=0)
loss_texts = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_images + loss_texts)/2
```

CLIP-Italian

Context: HuggingFace Community Week

[Open-to-the-community] Community week using JAX/Flax for NLP & CV

Flax/JAX Projects

patrickvonplaten 5 Jun 23

Learn how to use JAX/Flax with Transformers 🤗 + 🧬

We partnered-up with Google's Flax, JAX, and Cloud teams to organize a new community week from **July 7th to July 14th**. We want to teach you how to effectively use JAX/Flax for Natural Language Processing (NLP) and Computer Vision (CV).

Free access to a TPUv3-8 VM will kindly be provided by the Google Cloud team 🤯!

We can guarantee TPU access for the first 400 participants, so it might be worth to sign-up quickly 😉.

Jun 23

1 / 57

Jun 23

Back

We have been given 2 TPUv3-8 VM to run this project

CLIP-Italian Squad

Federico Bianchi

NLP

Raphael Pisoni

CV

Silvia Terragni

NLP

Gabriele Sarti

NLP

Sri Lakshimi

AI

Context: HuggingFace CLIP in JAX

huggingface / **transformers** Public

Watch 772 | Star 51.3k | Fork 12.2k

Code | Issues 342 | Pull requests 85 | Actions | Projects 23 | Wiki | Security | Insights

master | **transformers** / examples / research_projects / jax-projects / **hybrid_clip** /

Go to file | Add file | ···

edugp [Flax/run_hybrid_clip] Fix duplicating images when captions_per_image… ··· 0a22335 14 days ago **History**

..

README.md	Point to the right file for hybrid CLIP (#12599)	2 months ago
configuration_hybrid_clip.py	fix loading clip vision model (#12566)	2 months ago
modeling_hybrid_clip.py	[FlaxCLIP] allow passing params to image and text feature methods (#1…	last month
requirements.txt	[examples/flax] clip style image-text training example (#12491)	2 months ago
run_hybrid_clip.py	[Flax/run_hybrid_clip] Fix duplicating images when captions_per_image…	14 days ago

README.md

Vision-Text dual encoder model training examples

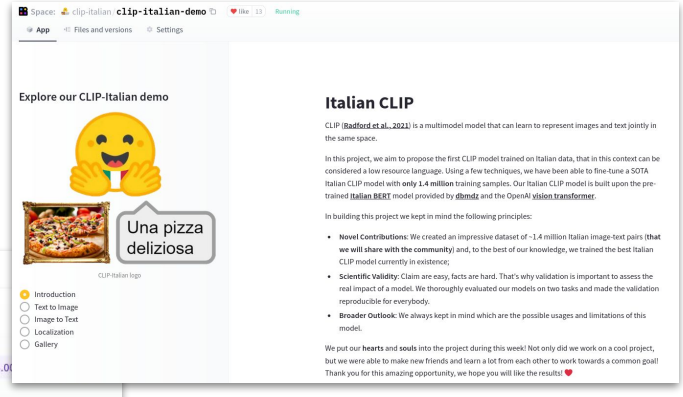
> Note: This example is experimental and might not give the best possible results

The following example showcases how to train a CLIP like vision-text dual encoder model using a pre-trained vision and text encoder using the JAX/Flax backend.

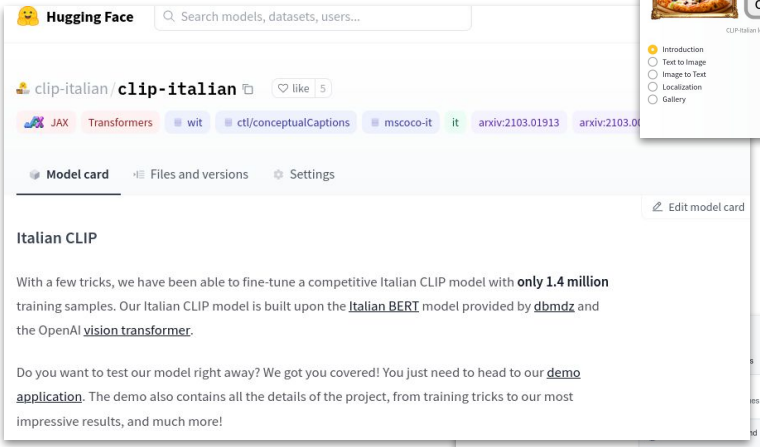
Such a model can be used for natural language image search and potentially zero-shot image classification. The model is inspired by the CLIP approach, introduced by Alec Radford et al. The idea is to train a vision encoder and a text encoder jointly to project the representation of images and their captions into the same embedding space, such that the caption embeddings are located near the embeddings of the images they describe.

CLIP-Italian: Open-Source

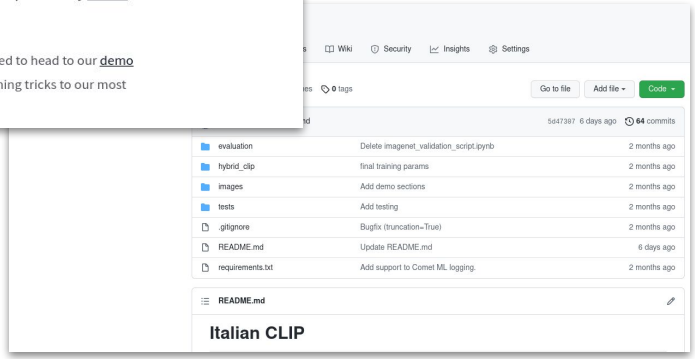
Demo



Pre-Trained Model



Code



Problem

We do not have 400 millions text-image pairs for Italian

Curated Datasets

Improved Training

CLIP-Italian: Datasets

CLIP-Italian Datasets: MSCOCO-IT

README.md

MSCOCO-it Dataset

A large scale dataset for Image Captioning in Italian

MSCOCO is a large scale dataset for training of image captioning systems. It contains(2014 version) more than 600,000 image-caption pairs. It contains training and validation subsets, made respectively of 82, 783 and 40, 504 images, where every image has 5 human-written annotations in English.

~100K images with captions, translated in Italian

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

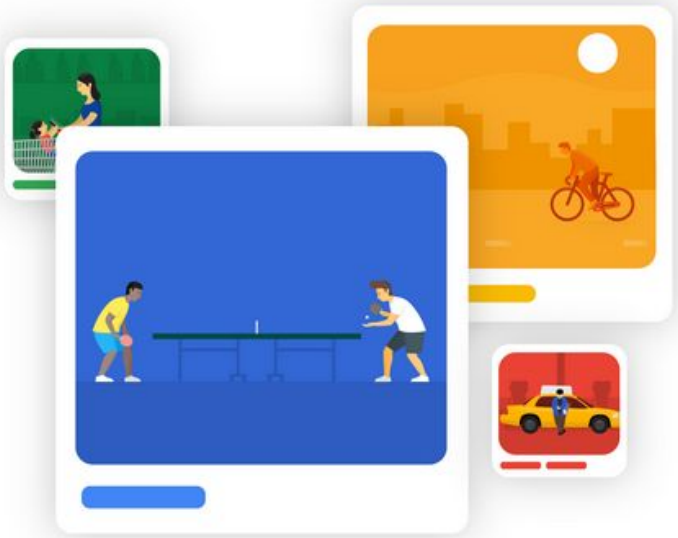
Scaiella, A., Croce, D., & Basili, R. (2019). Large scale datasets for Image and Video Captioning in Italian. IJCoL. Italian Journal of Computational Linguistics, 5(5-2), 49-60.

CLIP-Italian Datasets: 3CC

LANGUAGE TEAM

Google's Conceptual Captions

- We translated ~700K with DeepL out of 3.3M



Piyush Sharma, Nan Ding, Sebastian Goodman and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. Proceedings of ACL.

CLIP-Italian Datasets: WIT

WIT : Wikipedia-based Image Text Dataset

Wikipedia-based Image Text (WIT) Dataset is a large **multimodal multilingual** dataset. WIT is composed of a curated set of 37.6 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. Its size enables WIT to be used as a pretraining dataset for multimodal machine learning models.

Half Dome

Lo Half Dome visto da Washburn Point, da cui si nota il suo profilo caratteristico

Stato	Stati Uniti
Stato federato	California
Contea	Contea di Mariposa
Altezza	2 694 m s.l.m.
Prominenza	414,528 m
Catena	Sierra Nevada
Coordinate	37°44′45.73″N 119°31′58.58″W

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021). WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

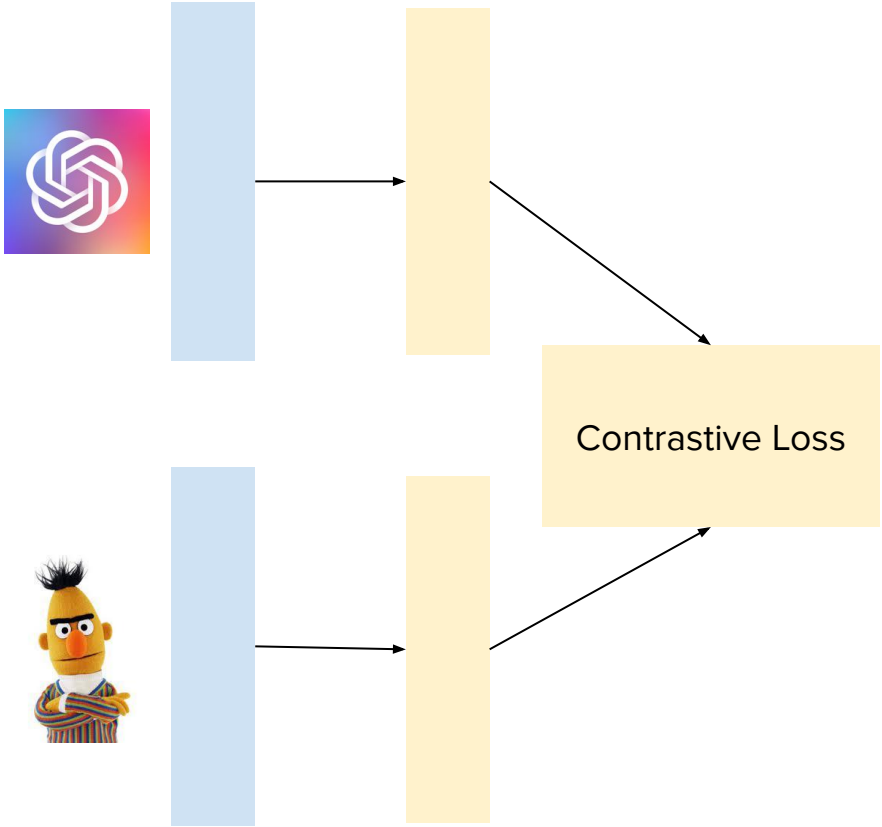
CLIP-Italian Final Dataset

- **MSCOCO-IT**: 100K
- **WIT**: 600K
  - (we applied some preprocessing, removing non meaningful captions)
- **3CC**: 700K
  - (we manually evaluated translation quality and computed inter-rater agreement)

Final Dataset: 1.4 million image-captions pairs (95% training, 5% validation)

CLIP-Italian: Training

CLIP-Italian Training: Architecture



- OpenAI's CLIP Vision Checkpoint
- Italian BERT

CLIP Training: Optimization

Before:

AdamW optimizer.

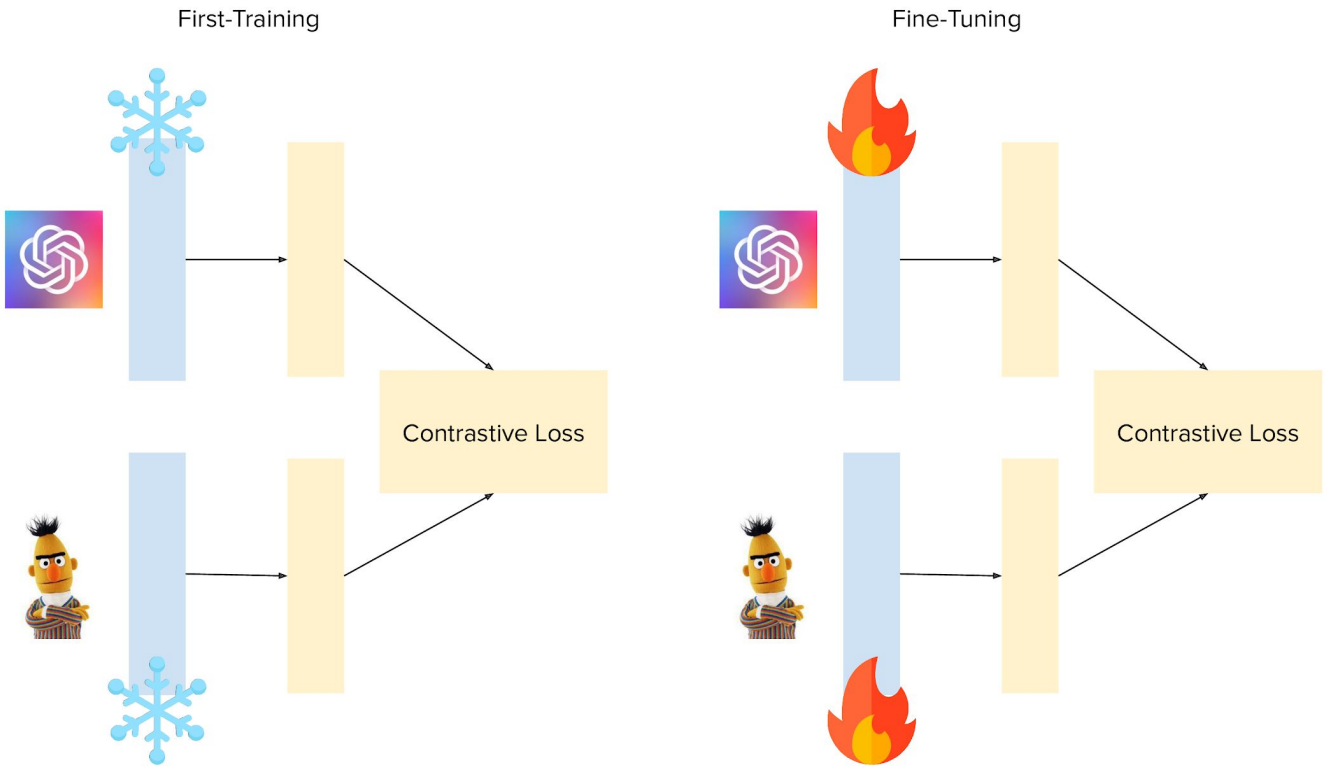
- Overfits quickly and the weight decay made this effect worse.

After:

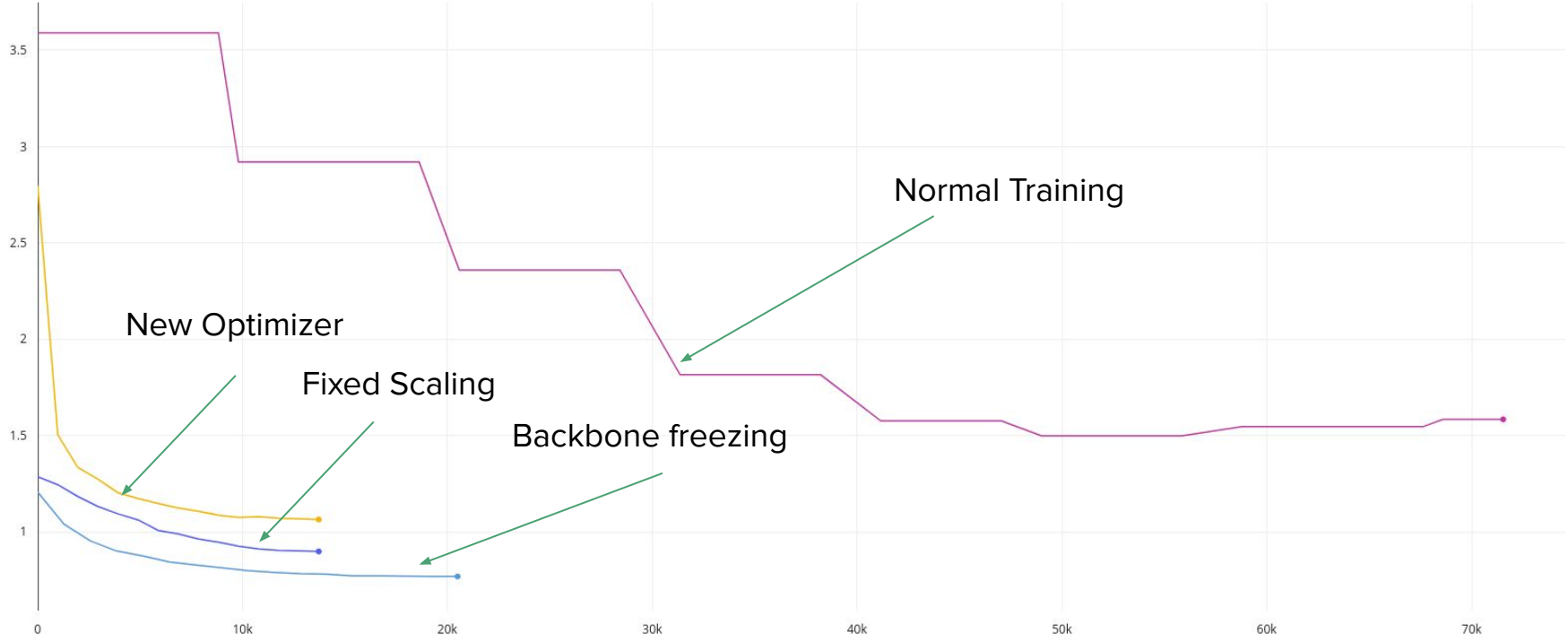
AdaBelief with Adaptive Gradient Clipping (AGC) and a Cosine Annealing Schedule.

- 25% improvement on the validation loss

CLIP Training: BackBone Freezing



CLIP Training: Validation Loss



https://www.comet.ml/g8a9/clip-italian/view/zhLk2alJaOe5wuPMn86HL4KeE

Super simple to use

Super simple to use

```python
from transformers import VisionTextDualEncoderModel, AutoProcessor

model = VisionTextDualEncoderModel.from_pretrained("clip-italian/clip-italian")
processor = AutoProcessor.from_pretrained("clip-italian/clip-italian")

inputs = processor(text="Un unicorno in tangenziale", return_tensors="pt")
embed = model.get_text_features(**inputs)
```

CLIP-Italian: Evaluation

Experiments

Task 1: Zero-Shot Classification

Dataset: ImageNet

Models:

- CLIP-Italian
- Multilingual CLIP (mCLIP)

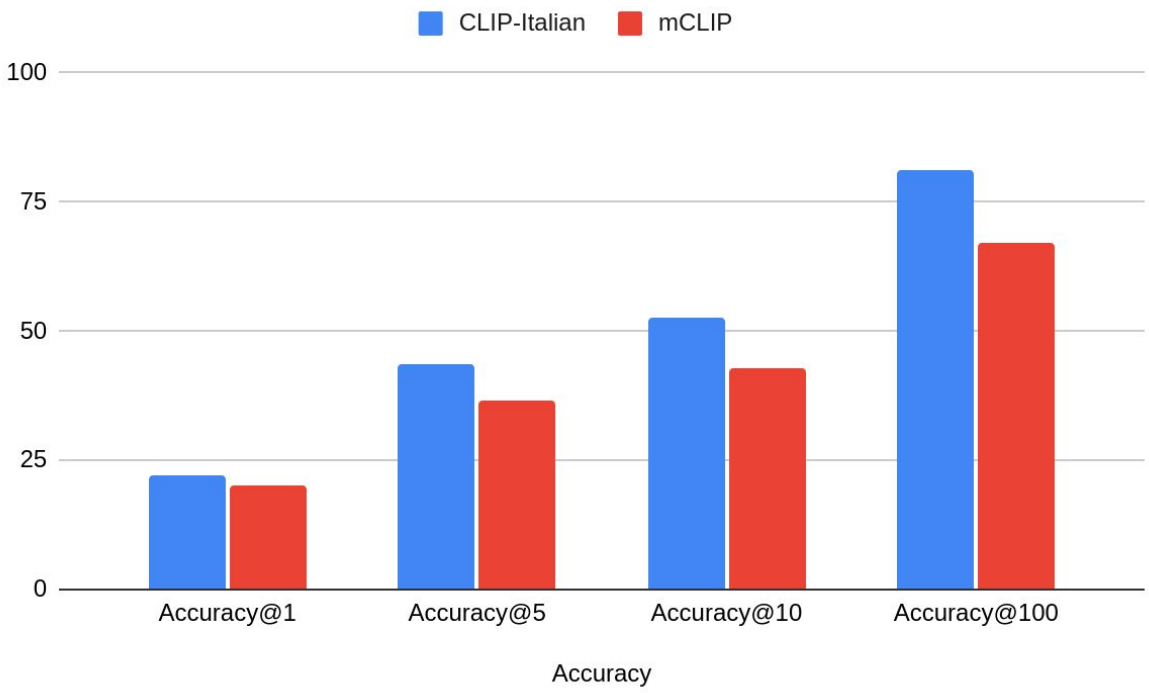
Task 2: Image Retrieval

Dataset: MSCOCO-IT Validation Set

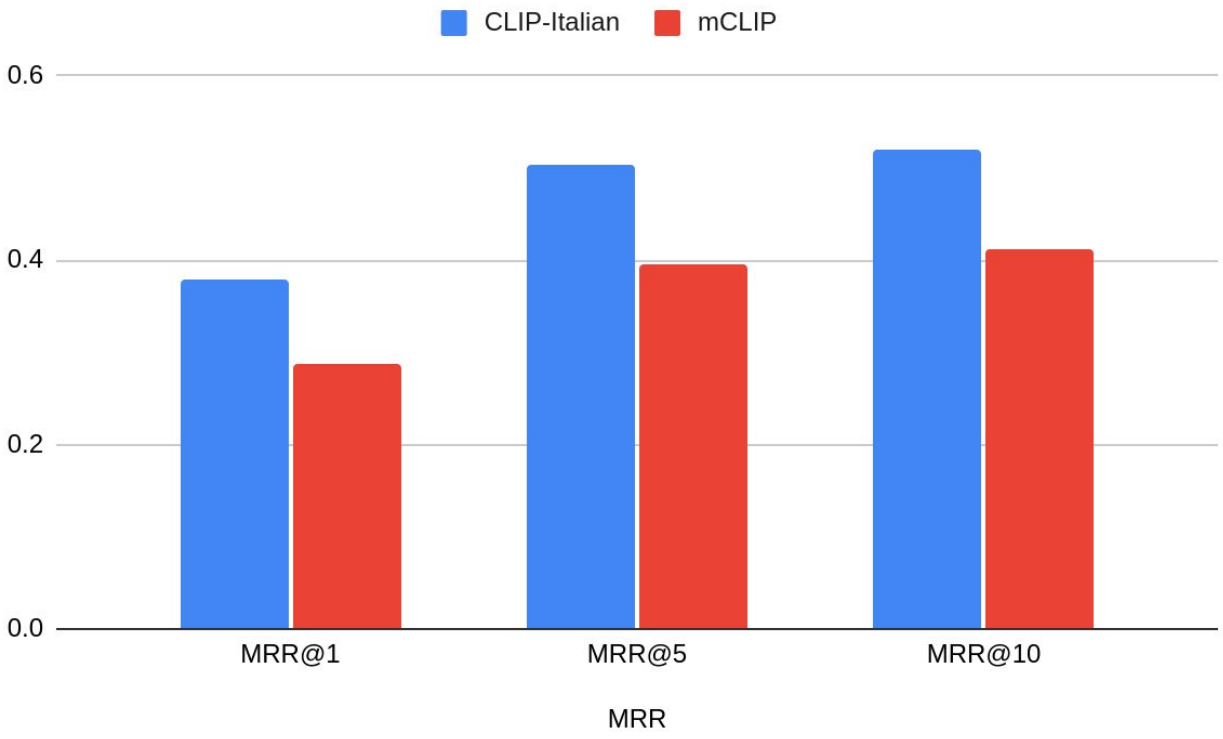
Models:

- CLIP-Italian
- Multilingual CLIP (mCLIP)

Task 1: Zero-Shot Classification



Task 2: Image Retrieval



CLIP-Italian: Examples

Image Retrieval

- We use the Unsplash dataset, 25K images
- CLIP-Italian has to find matching images in this big dataset of images

Query: A Couple (Una coppia)

Query: A Couple on the Mountain (Una coppia in montagna)

Query: A dress for the spring (un vestito primaverile)

Query: A dress for the autumn (un vestito autunnale)

Query: A Cat (un gatto)

Query: Two Cats (due gatti)

CLIP-Italian: Demo

Demo: ZeroShot Image Classification

Insert label 1

bandiera della spagna

Insert label 2

bandiera dell'italia

Insert label 3

un gatto

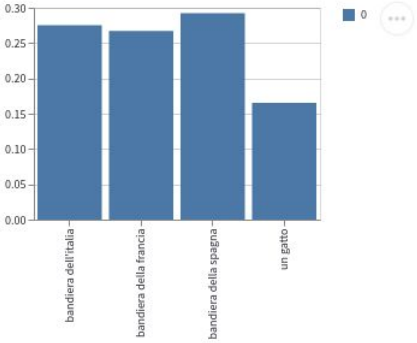
Insert label 4

bandiera della francia

Number of labels

4

CLASSIFY

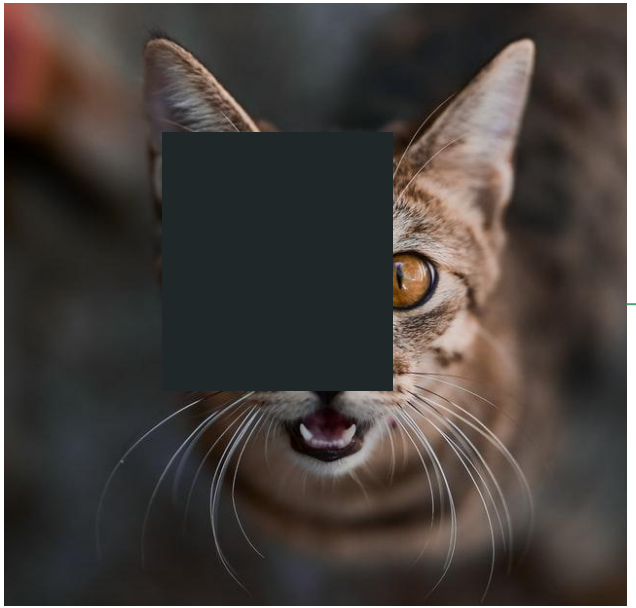


Demo: Localization

embedding

sim(image, “a cat”) = 0.8

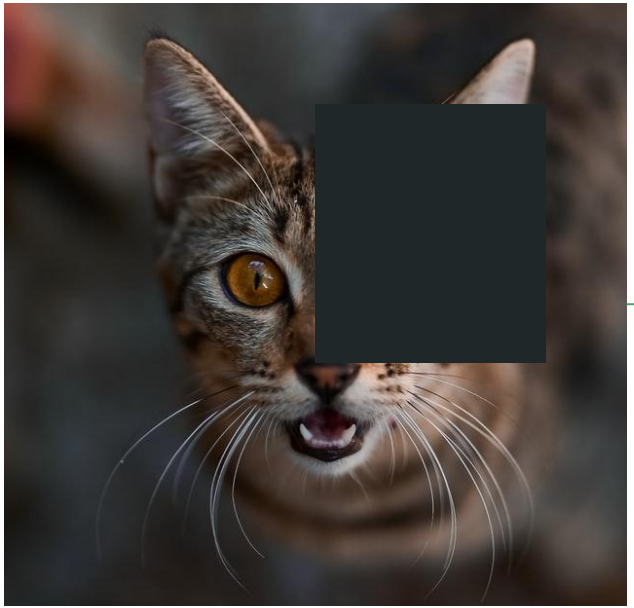
Demo: Localization



embedding

sim(image, “two cats”) = 0.5

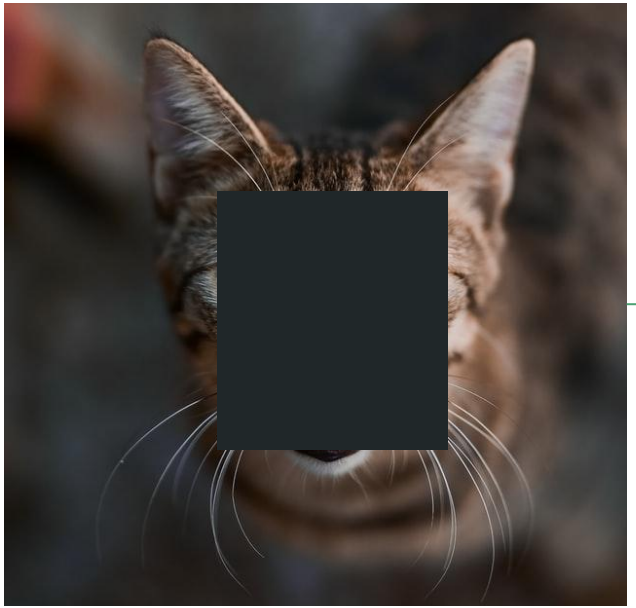
Demo: Localization



embedding

sim(image, “two cats”) = 0.7

Demo: Localization



embedding

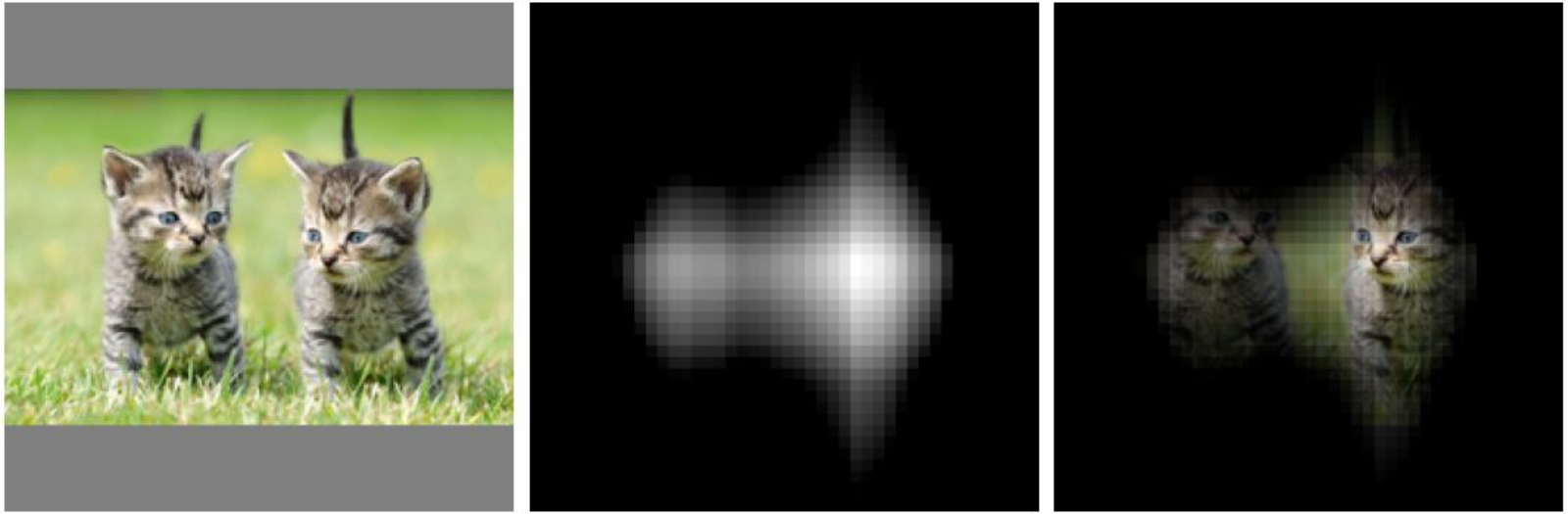
sim(image, “two cats”) = 0.2

Demo: Localization

embedding

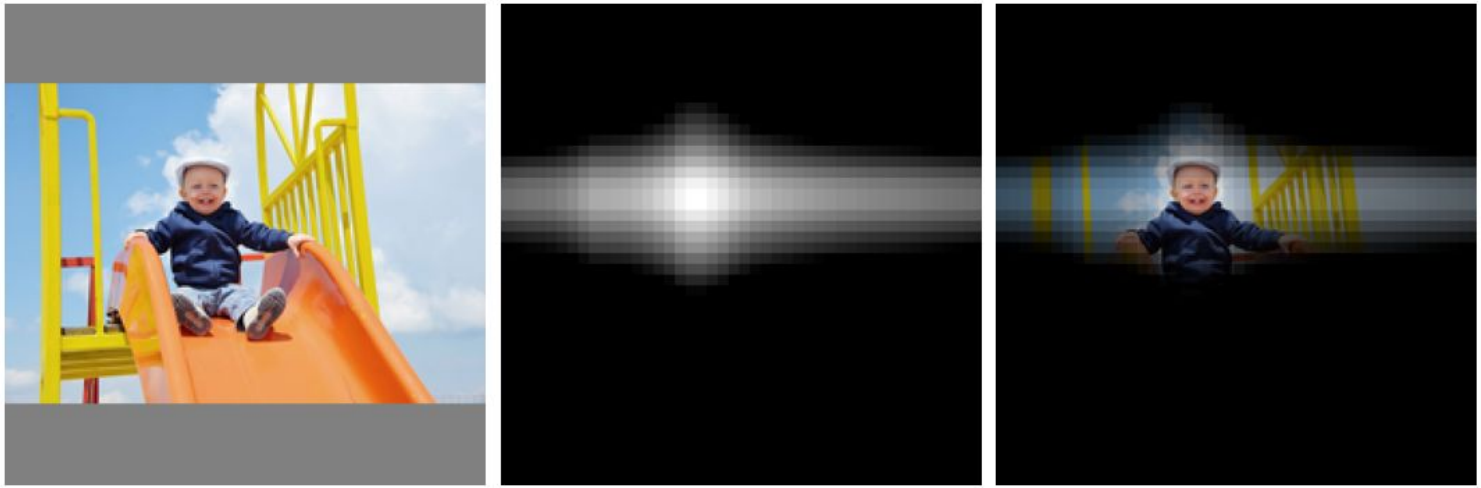
sim(image, “two cats”) = 0.1

Demo: Localization



“un gatto (a cat)”

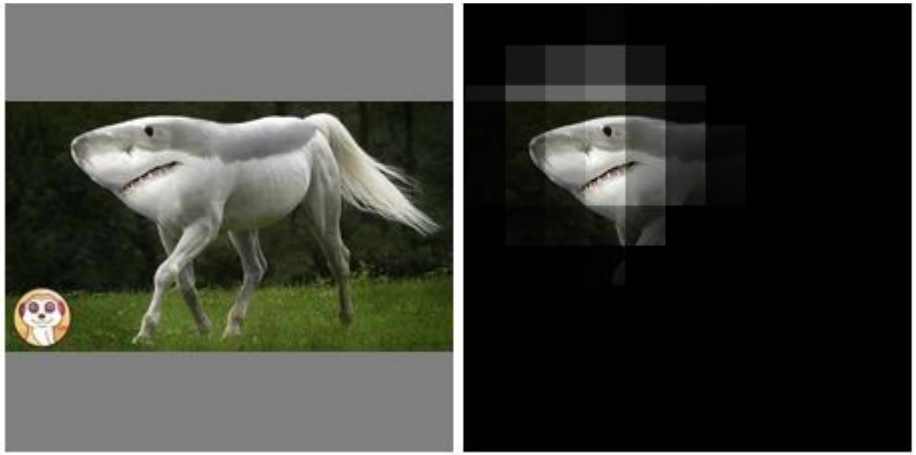
Demo: Localization



“un bambino (a baby)”

Demo: Localization

Demo: Localization



“uno squalo (a shark)”

Demo: Localization

“uno squalo (a shark)” “un cavallo (a horse)”

Thank you :)

Resources

- **Model**: https://huggingface.co/clip-italian/clip-italian
- **Colab** to play with: https://colab.research.google.com/drive/1SSddpjohAqRS_XxJvwz5HN1YFPevVJmy?usp=sharing
- **Blog**: https://towardsdatascience.com/how-to-train-your-clip-45a451dcd303
- **Demo**: https://huggingface.co/spaces/clip-italian/clip-italian-demo
- **Code**: https://github.com/clip-italian/clip-italian

- “mum, I’m on the press!”: https://www.html.it/24/07/2021/clip-italian/
- Yannic Kilcher’s: https://www.youtube.com/watch?v=SPOqoI0zOPQ&t=1407s

Thank you! :)