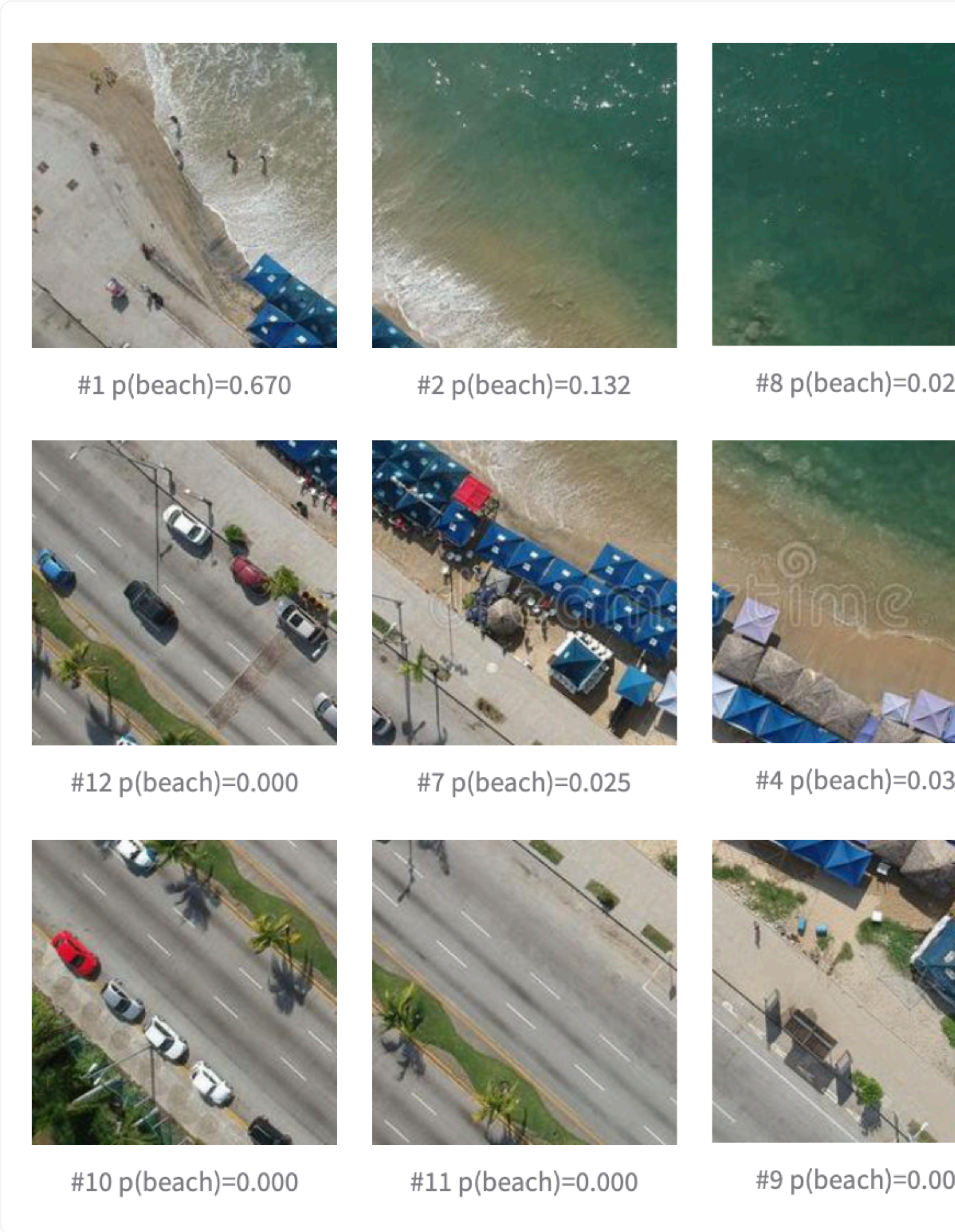
cataluna84
Mayank Bhaskar guest
ghosh-r
Ritobrata Ghosh guest
Sujit

Fine tuning CLIP with Remote Sensing (Satellite



We fine-tuned the CLIP Network from OpenAI with satellite images and capti
CLIP network learns visual concepts by being trained with image and caption
by using text paired with images found across the Internet. During inference,
relevant image given a text description or the most relevant text description g
enough to be used in zero-shot manner on everyday images. However, we felt
sufficiently different from everyday images that it would be useful to fine-tun
turned out to be correct, as the evaluation results (described below) shows. I
our training and evaluation process, and our plans for future work on this pro

The goal of our project was to provide a useful service and demonstrate how
Our model can be used by applications to search through large collections of
queries. Such queries could describe the image in totality (for example, beach
etc) or search or mention specific geographic or man-made features within th
fine-tuned for other domains as well, as shown by the medclip-demo team for

The ability to search through large collections of images using text queries is
and can be used as much for social good as for malign purposes. Possible app
and anti-terrorism activities, the ability to spot and address effects of climate
unmanageable, etc. Unfortunately, this power can also be misused, such as fo
by authoritarian nation-states, so it does raise some ethical questions as well.

You can read about the project on our project page, download our trained mo
own data, or see it in action on our demo.

Training

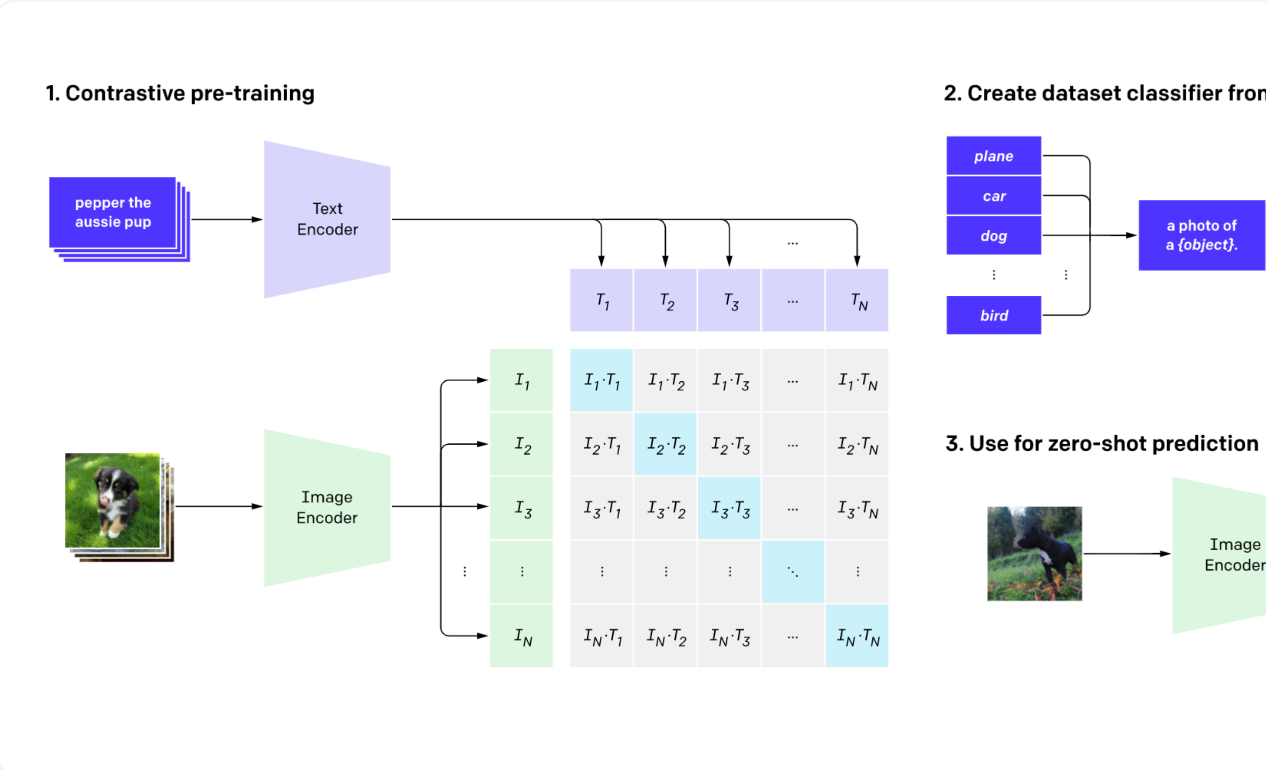
Dataset

We fine-tuned the CLIP model primarily with the RSICD dataset. This dataset

image has 5 captions. The Sydney dataset contains images of Sydney, Austral
613 images belonging to 7 classes. Images are (500, 500) RGB and provides 5
these additional datasets because we were not sure if the RSICD dataset woul
CLIP.

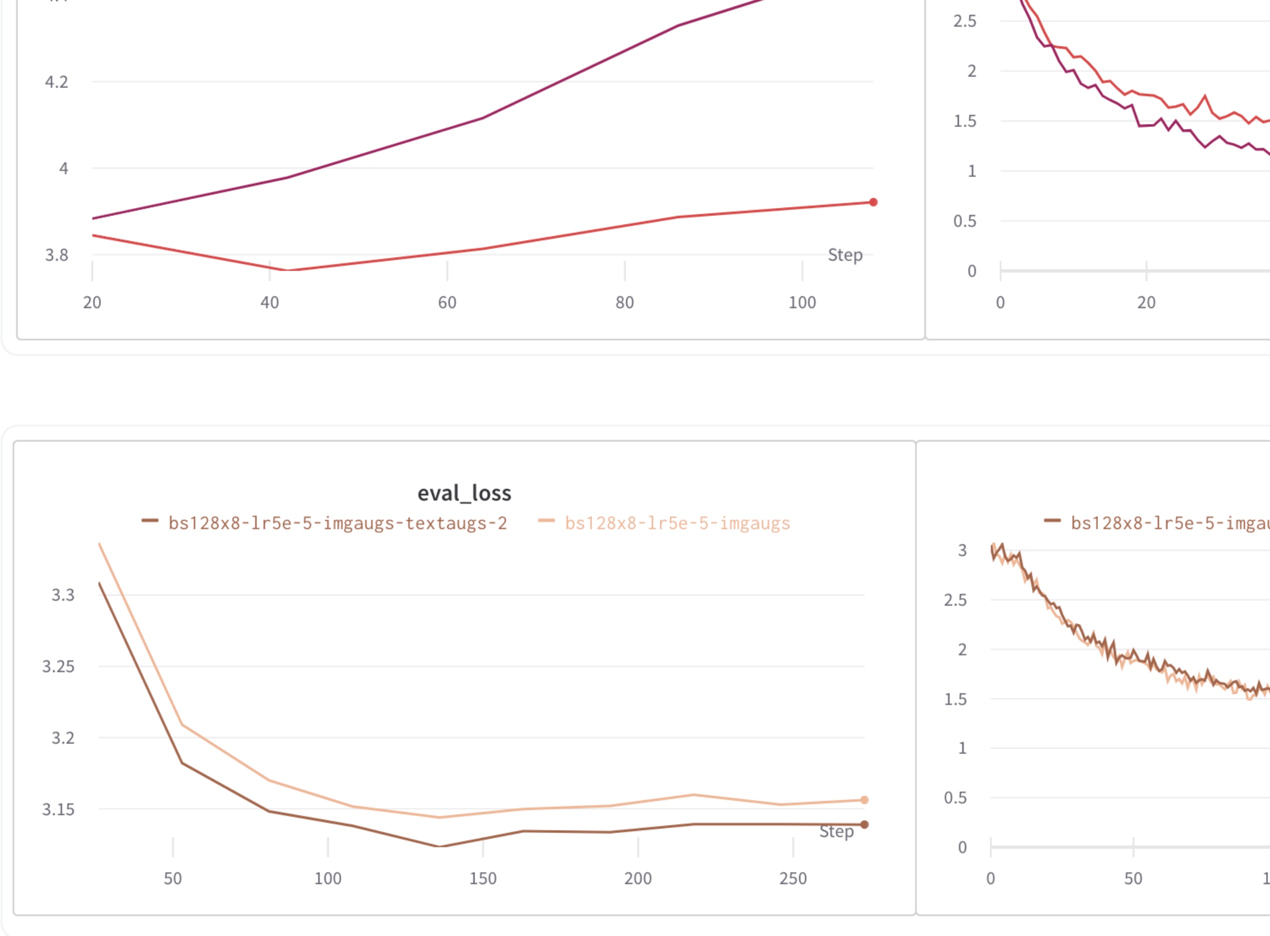
Model

Our model is just the fine-tuned version of the original CLIP model shown bel
batch of captions and a batch of images passed through the CLIP text encoder
The training process uses contrastive learning to learn a joint embedding rep
In this embedding space, images and their respective captions are pushed clo
and similar captions. Conversely, images and captions for different images, or
are likely to be pushed further apart.



CLIP Training and Inference (Image Credit: CLIP: Connecting Text and Ima

Data Augmentation



Evaluation and Training loss plots comparing (top) no augmentation vs image

augmentation vs text+image augmentatior

Evaluation

Metrics

A subset of the RSICD test set was used for evaluation. We found 30 categorie

evaluation was done by comparing each image with a set of 30 caption senter

`photograph of {category}"`. The model produced a ranked list of the 30 c

least relevant. Categories corresponding to captions with the top k scores (fo

compared with the category provided via the image file name. The scores are

images used for evaluation and reported for various values of k, as shown bel

The `baseline` model represents the pre-trained `openai/clip-vit-base-pa`

bs128x8-lr5e-5-imgaugs-textaugs/ckpt-8	0.831	(
bs128x8-lr5e-5-imgaugs/ckpt-4	0.746	(
bs128x8-lr5e-5-imgaugs-textaugs-2/ckpt-4	0.811	(
bs128x8-lr5e-5-imgaugs-textaugs-3/ckpt-5	0.823	(
bs128x8-lr5e-5-wd02/ckpt-4	0.820	(
bs128x8-lr5e-6-adam/ckpt-1[1]	**0.883**	(

1 - our best model, 2 - our second best model

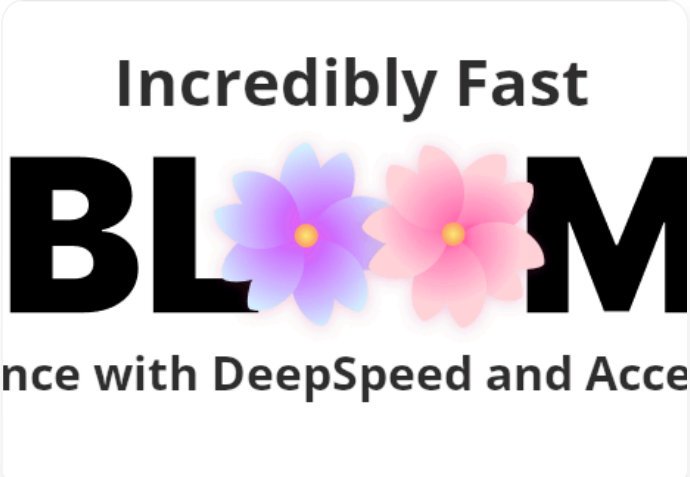
Demo

You can access the CLIP-RSICD Demo here. It uses our fine-tuned CLIP model functionality:

- Text to Image search
- Image to Image search
- Find text feature in image

The first two functionalities use the RSICD test set as its image corpus. They a tuned CLIP model and stored in a NMSLib index which allows Approximate N For text-to-image and image-to-image search respectively, the query text or in and matched against the image vectors in the corpus. For the third functional into patches and encode them, encode the queried text feature, match the tex vector, and return the probability of finding the feature in each patch.

More articles from our Blog



Optimization story: Bloom inference

By Narsil   Oct 12, 2022

Introducing D
Object Identifi
and Models

By sylvestre   Oct

TOS   Privacy   Team   Jobs   Models