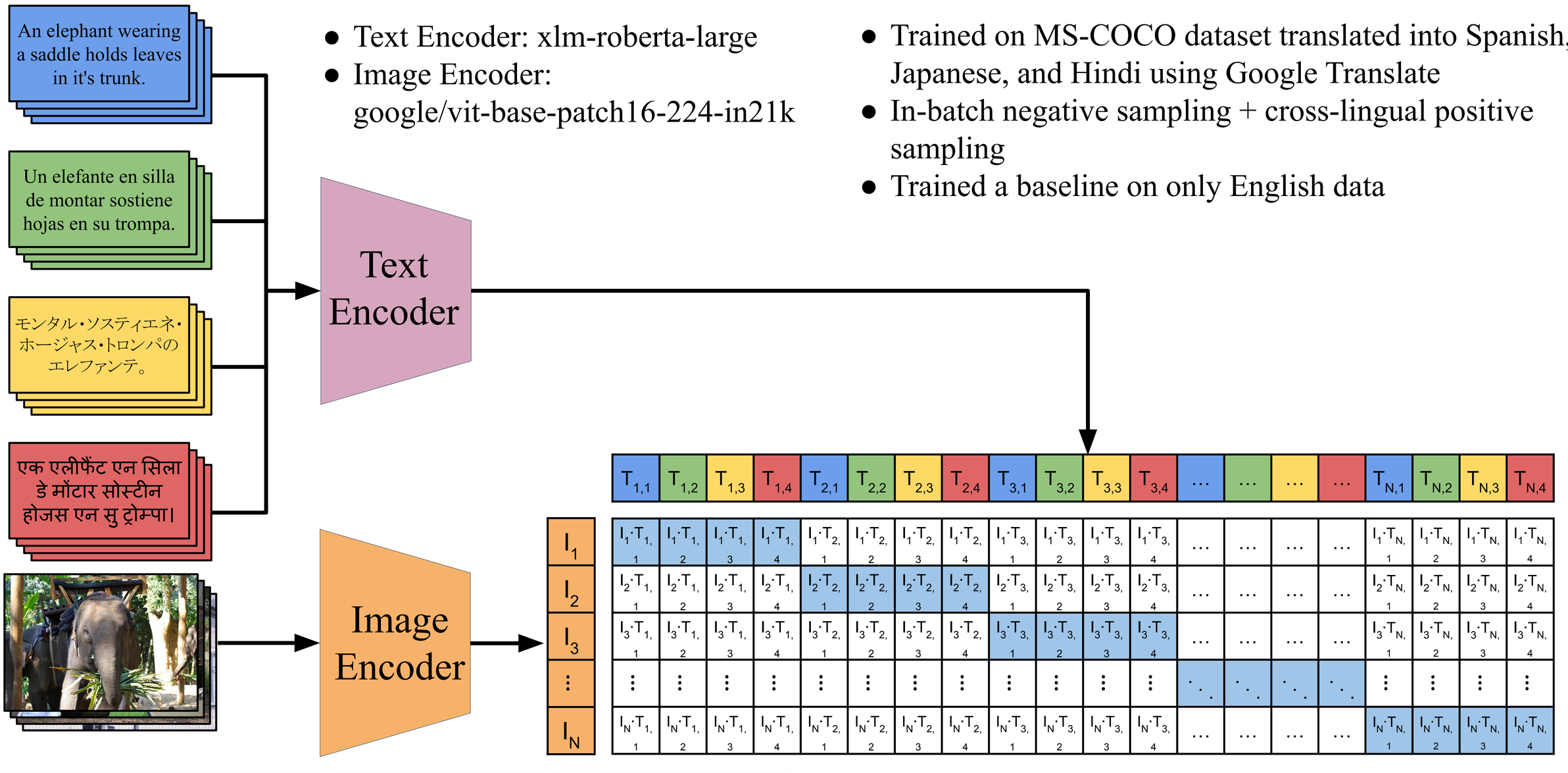
Parallel Multilingual Pre-Training for Multi-Modal Representations

Nathaniel Krasner nkrasner@gmu.edu

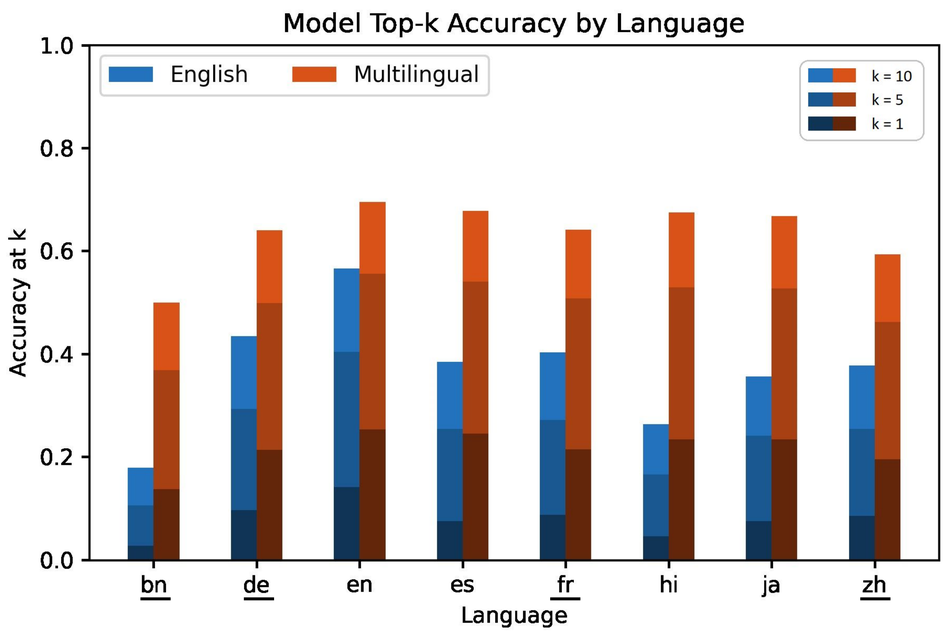
Nicholas Lanuzo nlanuzo@gmu.edu

Parallel Contrastive Pre-Training

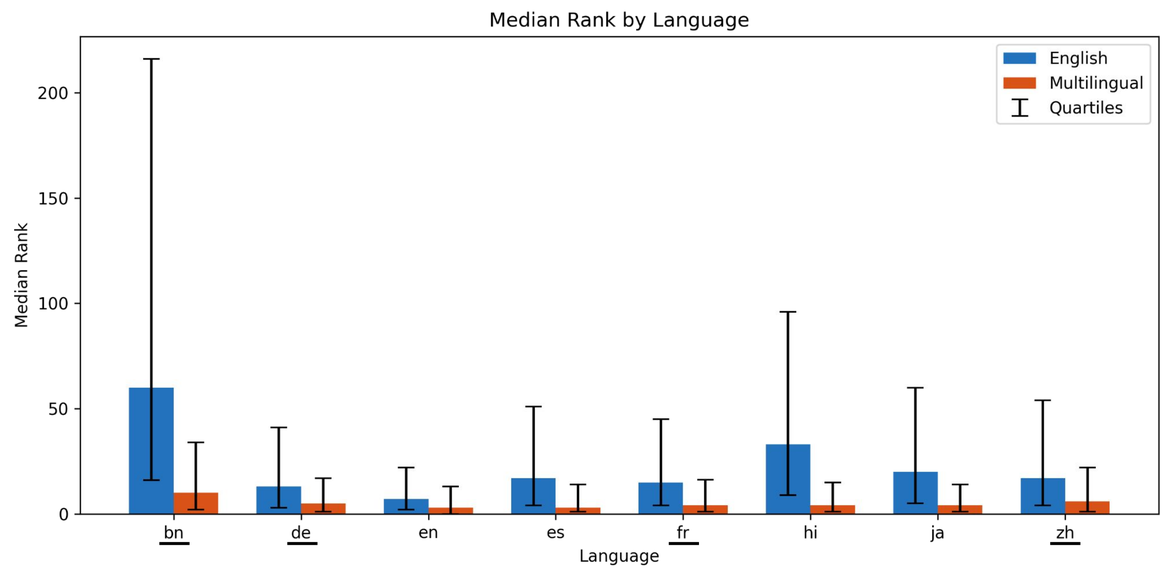


- Text Encoder: xlm-roberta-large
- Image Encoder: google/vit-base-patch16-224-in21k
- Trained on MS-COCO dataset translated into Spanish, Japanese, and Hindi using Google Translate
- In-batch negative sampling + cross-lingual positive sampling
- Trained a baseline on only English data

Cross-Modal Retrieval



- Accuracy for retrieving nearest neighbor image based on encoding distance to sentence
- Higher is better
- Multilingual training even improves English performance
- Improves zero-shot performance greatly (underlined)



- Median and quartiles for where correct image ranks by distance from sentence encoding
- Lower is better
- Multilingual training improves consistency as well as overall performance

Code

https://github.com/nkrasner/ML-CLIP

Bitext Retrieval

- Tested bitext retrieval from English encodings to others
- Multilingual training improves alignment between languages
- Over the COCO validation set, so from a pool of 5000 candidate sentences