Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

ALIGN (A Large-scale Image and Noisy-text embedding)

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, Tom Duerig

Google Research
Representative Datasets in training large-scale vision, language and vision+language models

- State-of-the-art models have large sizes and are trained on huge amounts of data!
- Unlike language models, the datasets for vision & multimodal learning are usually smaller, proprietary, or dependent on expensive models / annotators for cleaning

| Dataset Name       | Modality           | Size            | Generation                                      |
|--------------------|--------------------|-----------------|-------------------------------------------------|
| JFT                | vision             | 300M - 3B       | web + user + complex models + annotators        |
| Conceptual Captions| vision + language  | 3M - 12M        | web + complex models                            |
| C4                 | language           | 150B tokens     | web + heuristics                                |
| Ours (ALIGN)       | vision + language  | 1.8B            | web + heuristics                                |
Dataset

Raw image+alt-text data from web pages

Remove porn/too-small images

Minimal frequency-based text filtering

Remove Images

- pornographic images
- images that were too small (shorter dimension $\leq 200$ pixels)
- images with irregular shape (aspect ratio $\geq 3$)
- images associated with more than 1000 alt-texts
Dataset

- Raw image+alt-text data from web pages
- Remove porn/too-small images
- Minimal frequency-based text filtering

Remove Texts
- Associated with > 10 images (e.g., “1920x1080”, “alt_img”, “cristina”)
- Out-of-vocab (top 100M unigrams & bigrams)
- Too short (<3 unigrams) or too long (>20 unigrams)
Text-based filtering – overly frequent alt-texts

- Alt-texts shared by more than 10 images were discarded
- Examples: “1920x1080”, “alt img”, “cristina”, etc.
Text-based filtering – rare tokens

- Vocabulary: top 100M unigrams and bigrams
- Kept alt-texts with in-vocabulary rate == 1

| In-vocabulary rate | Example alt-text                                                      |
|--------------------|---------------------------------------------------------------------|
| == 1               | senior gentleman reading a newspaper and leaning against a wall stock image |
| [0.8, 1)           | special friends day always good for the soul being poolside is too kinder calmer racv summer breathe |
| [0.4, 0.8)         | shoulder bags travelcomputerbag star luggageampbag                   |
| [0.0, 0.4)         | image_tid 25&id mggqpuweqpd&cache 0&lan_code 0                      |
Text-based filtering – alt-text length

Exclude text that:
1. Too short (<3 unigrams)
2. too long (>20 unigrams)
Dataset

Final training data: 1.8B noisy image-text pairs
Contrastive Learning on Noisy Image-Text Data

- Data is noisy but can provide cross-modality supervision
- Contrastive learning is data-efficient and scales easily
- Text caption prediction has proven to be effective in learning vision models

Applications
- Visual classification
- Image-text matching/retrieval

Compared to concurrent work CLIP
- ALIGN data: minimal frequency-based filtering on raw web
- CLIP data: data balancing + controlled source blending (e.g. YFCC100M)
A Two-Tower Model

Positives: paired image-text data
Negatives: all others in the same batch

Contrastive loss higher similarity (dot product) for matched pairs and lower similarity for unmatched pairs

$$L_{t2i} = -\frac{1}{N} \sum_{i}^{N} \log \frac{\exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N} \exp(x_i^T y_j/\sigma)}$$

$$L_{i2t} = -\frac{1}{N} \sum_{i}^{N} \log \frac{\exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N} \exp(y_i^T x_j/\sigma)}$$
The Evolution of the Two-Tower Model

- **Text Tower**: BERT transformer
- **Image Tower**: EfficientNet

**Other Key technicals:**
- Optimizer
- Softmax temperature
- Embedding dims

Positives: paired image-text data
Negatives: all others in the same batch

Negatives beyond a single batch
The Evolution of the Two-Tower Model

Flickr30k image retrieval recall@1 and Flickr30k text retrieval recall@1

- ResNet101+BagOfWords + 50k text neg
- Transformer
- Cross-Replica negatives
- EfficientNet-B7
- LAMB
- English Data Only
- Initial Model
**Image-Text Tasks**

- **Image-Text Retrieval Tasks from Image Captioning Dataset**
  - MSCOCO
    - 5 captions per image
    - 112k training pairs, 5k test pairs
  - Flickr30k
    - 5 captions per image
    - 29k training pairs, 1k test pairs
  - **Metrics**
    - Recall@K

- **CrissCrossed Captions (CxC)**
  - Graded human judgments for MSCOCO caption-caption, image-caption and image-image pairs. Total ~270k pair
  - **Metrics**
    - Recall@K
    - Spearman’s Correlation Coefficient
Configurations

- **Pre-training**
  - BERT-Large + EfficientNet-L2
  - 16384 effective batch size (on 1024 cloud TPUv3 cores)
  - LAMB optimizer with weight decay ratio 1e-5,
  - Trained 1.2M steps with learning rate 1e-3, 10k warm up, linear decay to 0

- **Fine-tuning**
  - 2048 effective batch size.
  - 1e-5 initial learning rate with linear decay.
  - 3k/6k steps (MSCOCO, Flickr30k).
# Image-Text Retrieval Results

| Model  | Flickr30K (1K test set) R@1 | MS-COCO (5K test set) R@1 |
|--------|-----------------------------|---------------------------|
|        | image → text | text → image | image → text | text → image |
| Zero-shot |          |              |          |              |
| ImageBERT | 70.7         | 54.3         | 44.0     | 32.3         |
| UNITER  | 83.6         | 68.7         | -        | -            |
| CLIP    | 88.0         | 68.7         | 58.4     | 37.8         |
| ALIGN   | 88.6         | 75.7         | 58.6     | 45.6         |
| Fine-tuned |          |              |          |              |
| GPO     | 88.7         | 76.1         | 68.1     | 52.7         |
| UNITER  | 87.3         | 75.6         | 65.7     | 52.9         |
| ERNIE-Vil | 88.1        | 76.7         | -        | -            |
| VILLA   | 87.9         | 76.3         | -        | -            |
| Oscar   | -            | -            | 73.5     | 57.5         |
| ALIGN   | 95.3         | 84.9         | 77.0     | 59.9         |

Image-text retrieval results (recall@1) on Flickr30K and MS-COCO datasets (both zero-shot and fine-tuned).

ALIGN significantly outperforms existing methods including the cross-modality attention models that are too expensive for large-scale retrieval applications.
CxC Results

### Retrieval Eval R@1

|          | image->text | text->image | text->text | image->image |
|----------|-------------|-------------|------------|-------------|
| VSE++    | 43.1        | 32.5        | 38.7       | 36.4        |
| VSRN     | 52.4        | 40.1        | 41         | 44.2        |
| DE_{I2T} | 53.9        | 39.8        | 26         | 38.3        |
| DE_{I2T+T2T} | 55.9    | 41.7        | 42.4       | 38.5        |
| ALIGN    | 78.1        | 61.8        | 45.4       | 49.4        |

### Correlation Eval

|          | STS          | SIS          | SITS         | Mean Avg    |
|----------|--------------|--------------|--------------|-------------|
| VSE++    | **74.4±0.4** | 73.3±0.9     | 55.2±1.5     | **67.7**    |
| VSRN     | 73.0±0.4     | 70.1±1.0     | 60.4±1.3     | 67.8        |
| DE_{I2T} | 50.9±0.6     | **81.3±0.7** | 61.6±1.4     | 64.6        |
| DE_{I2T+T2T} | 74.2±0.2 | 74.5±0.9     | 61.9±1.3     | 70.2        |
| ALIGN    | 72.9±0.4     | 77.2±0.8     | **67.6±1.2** | **72.6**    |

**Retrieval:**
- ALIGN is better across the board. More than +20% R@1 on i->t and t->i.

**Correlation:**
- Training objective is aligned with the image-text correlation (SITS). Intermodal performance is stronger.
- Intramodal performance (STS, SIS) is relatively lower.
Multimodal Retrieval

Text -> Image

“Van Gogh Starry Night ...”
- “details”
- “in black and white”
- “on a canvas”
- “in dark wood frame”

“Lombard street ...”
- “view from bottom”
- “view from top”
- “bird’s eye view”
- “in heavy rain”

“seagull in front of ...”
- “Golden Gate Bridge”
- “London Tower Bridge”
- “Sydney Harbour Bridge”
- “Rialto Bridge”

Candidates pool:
160M CC-BY licensed images that are separate from our training set.
Multimodal Retrieval

Image + Text -> Image
Multimodal Retrieval

Image - Text -> Image

- “cars”  - “trees”  - “houses”

- “flowers”  - “orange”  + “rose”

- “bridge”  - “waterfall”  - “mountain”

- “tree”  - “red”  - “snow”
Visual Classification

Classification layers trained with a set of training data (could be expensive to obtain)

Supervised visual classification with an image representation model
Visual Classification -- Zero-shot

A image-to-text retrieval problem with the pre-trained image & text encoders

ALIGN data covers almost all visual concepts -- **No additional training data is needed.**
Visual Classification -- Zero-shot

Same text prompt ensembling as in CLIP: averaging embedding of templates like “A photo of a {classname}”. +2.9% ImageNet top-1 accuracy

|       | ImageNet | ImageNet-R | ImageNet-A | ImageNet-V2 |
|-------|----------|------------|------------|-------------|
| CLIP  | 76.2     | 88.9       | 77.2       | 70.1        |
| ALIGN | 76.4     | 92.2       | 75.8       | 70.1        |

Classnames → ALIGN training images retrieval: large amount of non-natural training images
# Visual Classification -- Supervised learning

- Train classification head → Fine-tune all variables
- InceptionNet cropping + FixRes
- SGD w/ momentum = 0.9

| Methods                               | Frozen feature | Top-1 | Top-5 |
|----------------------------------------|----------------|-------|-------|
| Instagram (ResNext-101 32x4d)          | 83.6           | 85.4  | 97.6  |
| CLIP (ViT-L/14)                        | 85.4           | n/a   | n/a   |
| BiT (ResNet152x4)                      | n/a            | 87.54 | 98.46 |
| Vision Transformer (ViT-H/14)          | n/a            | 88.4  | 98.7  |
| Noisy Student (EfficientNet-L2)        | n/a            | 88.44 | 98.7  |
| Meta Pseudo Labels (EfficientNet-L2)   | n/a            | 90.2  | 98.8  |
| ALIGN (EfficientNet-L2)                | 85.5           | 88.64 | 98.67 |

## Fine-grained tasks

| Methods                               | Oxford Flowers | Oxford Pets | Stanford Cars | Food 101 |
|----------------------------------------|----------------|-------------|---------------|----------|
| BiT-L (ResNet-152 x4)                  | 99.63          | 96.62       | n/a           | n/a      |
| SAM-baseline (EfficientNet-L2)         | 99.60          | 96.92       | 95.07         | 96.03    |
| SAM-final (EfficientNet-L2)            | 99.65          | 97.10       | 95.96         | 96.18    |
| ALIGN (EfficientNet-L2)                | 99.65          | 96.19       | **96.13**     | 95.88    |
Visual Classification -- VTAB

- 19 Tasks w/ three groups
  - Natural: Caltech101, CIFAR-100, etc.
  - Specialized: Resisc45, Diabetic Retinopathy, etc.
  - Structured: Clevr, dSprites, etc.
- Few shot: Fine-tuned on 1000 samples
- Hyper-param sweep for each task for 50 trials (800 train + 200 val)

|        | All tasks | Natural | Specialized | Structured |
|--------|-----------|---------|-------------|------------|
| BiT-L  | 78.72     | -       | -           | -          |
| ALIGN  | **79.99±0.15** | 83.38   | 87.56       | 73.25      |
Ablation Study -- Modal Capacity

- Image encoder quality relies more on image encoder capacity (not surprising)
- Vision Transformer backbone? (on-going work)
  - ViT-H outperforms EfficientNet-L2 (model quality is not saturated yet)
  - More robust to optimizer choices: Adam / Adafactor works well
# Ablation Study -- Dataset Size

Large models can overfit on small datasets

When dataset is sufficiently large, large models scale better

| Model + Data          | MSCOCO I2T R@1 | MSCOCO T2I R@1 | ImangeNet KNN R@1 |
|-----------------------|----------------|----------------|------------------|
| **B7 + BERT-base**    |                |                |                  |
| + ALIGN full data     | 55.4           | 41.7           | 69.3             |
| + ALIGN 10% data      | 52.0           | 39.2           | 68.8             |
| + CC-3M data          | 18.9           | 15.5           | 48.7             |
| **B7 + BERT-mini**    |                |                |                  |
| + ALIGN full data     | 37.4           | 24.5           | 56.5             |
| + ALIGN 10% data      | 36.7           | 24.4           | 55.8             |
| + CC-3M data          | 22.1           | 17.3           | 48.9             |

| Model + Data          | MSCOCO I2T R@1 | MSCOCO T2I R@1 | ImangeNet KNN R@1 |
|-----------------------|----------------|----------------|------------------|
| **B7 + BERT-base**    |                |                |                  |
| + ALIGN 12M data      | 23.8           | 17.5           | 51.4             |
| + ALIGN 6M data       | 15.8           | 11.9           | 47.9             |
| + ALIGN 3M data       | 8.1            | 6.3            | 41.3             |
| + CC-3M data          | 18.9           | 15.5           | 48.7             |

4x noisy data (12M ALIGN samples) outperforms clean data (3M Conceptual Captions)
Multilingual

Adding another text "cute cat" in Swahili

Adding text in Chinese

Adding another picture of a cute cat
Multilingual Dataset

- Flickr30K
  - One of the earliest; Images from Flickr
  - Multi30K (Translations/Human Generations in cs, de, fr)

- MS-COCO
  - STAIR (Human generations in ja)

- XTD
  - Test set for 7 well-resourced languages in es, it, ko, pl, ru, tr, zh

- WIT: Wikipedia-based Image-Text Dataset
  - Combines scale of Wikipedia with Human annotations
  - First time ever in 108 languages
# Multilingual Dataset

| Name    | Train-I | Train-T | Dev-I | Dev-T | Test-I | Test-T | #Langs |
|---------|---------|---------|-------|-------|--------|--------|--------|
| Multi30k| 29k     | 145k    | 1k    | 5k    | 1k     | 5k     | 4      |
| MS-COCO | 82k     | 410k    | 5k    | 25k   | 5k     | 25k    | 1      |
| STAIR   | 82k     | 410k    | 5k    | 25k   | 5k     | 25k    | 1      |
| WIT     | 11.4m   | 16m     | 5/3/1k| 5/3/1k| 5/3/1k | 5/3/1k | 108    |
| XTD     | -       | -       | -     | -     | -      | 1k     | 1k     | 7      |
## Multilingual Model

- Lift language constraint and match the size of English training data
- Same architecture and training params as EN model (Vocab size: 100K → 250K)
- Evaluated on Multi30K dataset
  - mean Recall: avg R@1, R@5, R@10 on img-to-txt & txt-to-img retrieval

|          | EN  | DE  | FR  | CS  |
|----------|-----|-----|-----|-----|
| zero-shot|     |     |     |     |
| M³P      | 57.9| 36.8| 27.1| 20.4|
| ALIGN\textsubscript{EN} | 92.2| --  | --  | --  |
| ALIGN\textsubscript{mling} | 90.2| 84.1| 84.9| 63.2|
| with fine-tuning |     |     |     |     |
| M³P      | 87.7| 82.7| 73.9| 72.2|
| UC2      | 88.2| 84.5| 83.9| 81.2|
Multilingual ALIGN Data

~52% data is from English
LaBSE – Learning Multilingual From Text

LaBSE: Language-agnostic BERT Sentence Embedding: [https://arxiv.org/abs/2007.01852](https://arxiv.org/abs/2007.01852)
Combining the best of two worlds

- Added balanced text-text paired data.
- Shared Text encoder between two tasks (image-text and text-text)
- Task-specific Projection layer on Text encoder

Data:
- 1.8 Billion ALIGN data
- 6 Billion translation pairs
Summary and Future Work

Summary

- Large-scale image-text data from the web with minimal frequency-based filtering.
- Scale up model sizes with simple dual-encoder & contrastive learning
- SOTA results on visual classification and image-text retrieval

Future work

- Responsible AI: harmful data and unfair bias in multimodal models
- Quality improvement on low-resourced languages
- Limitations on model scaling with contrastive learning (negative sample size)