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

Chao Jia [1] Yinfei Yang [1] Ye Xia [1] Yi-Ting Chen [1] Zarana Parekh [1] Hieu Pham [1] Quoc V. Le [1]
Yunhsuan Sung [1] Zhen Li [1] Tom Duerig [1]

Presented by Xinrui Song

08/11/2021

Challenges

- Limited dataset size for visual & visual-language representation

Challenges

- Limited dataset size for visual-language representation

Representation learning in NLP -> raw text without human annotation

Vision domain -> 300M sized dataset

Challenges

- Limited dataset size for visual-language representation pre-training

Representation learning in NLP -> raw text without human annotation

Vision domain -> 300M samples

Vision-language domain -> 10M samples

Challenges

- Limited dataset size for visual-language representation pre-training

Solution

- Make use of noisy labels

Challenges

- Limited dataset size for visual-language representation pre-training

Solution

- Make use of noisy labels

- Over 1B noisy image alt-text pairs
- Simple frequency-based filtering
- Dual encoder contrastive training method

- Over 1B noisy image alt-text pairs
- **Simple frequency-based filtering**
- Dual encoder contrastive training method

Figure 2. Example image-text pairs randomly sampled from the training dataset of ALIGN. One clearly noisy text annotation is marked in *italics*.

1. Remove pornographic images
2. Exclude alt-text shared by >10 images
3. Discard rare token
4. Discard alt-texts with <3 or >20 unigrams

- Over 1B noisy image alt-text pairs
- Simple frequency-based filtering
- **Dual encoder contrastive training method**

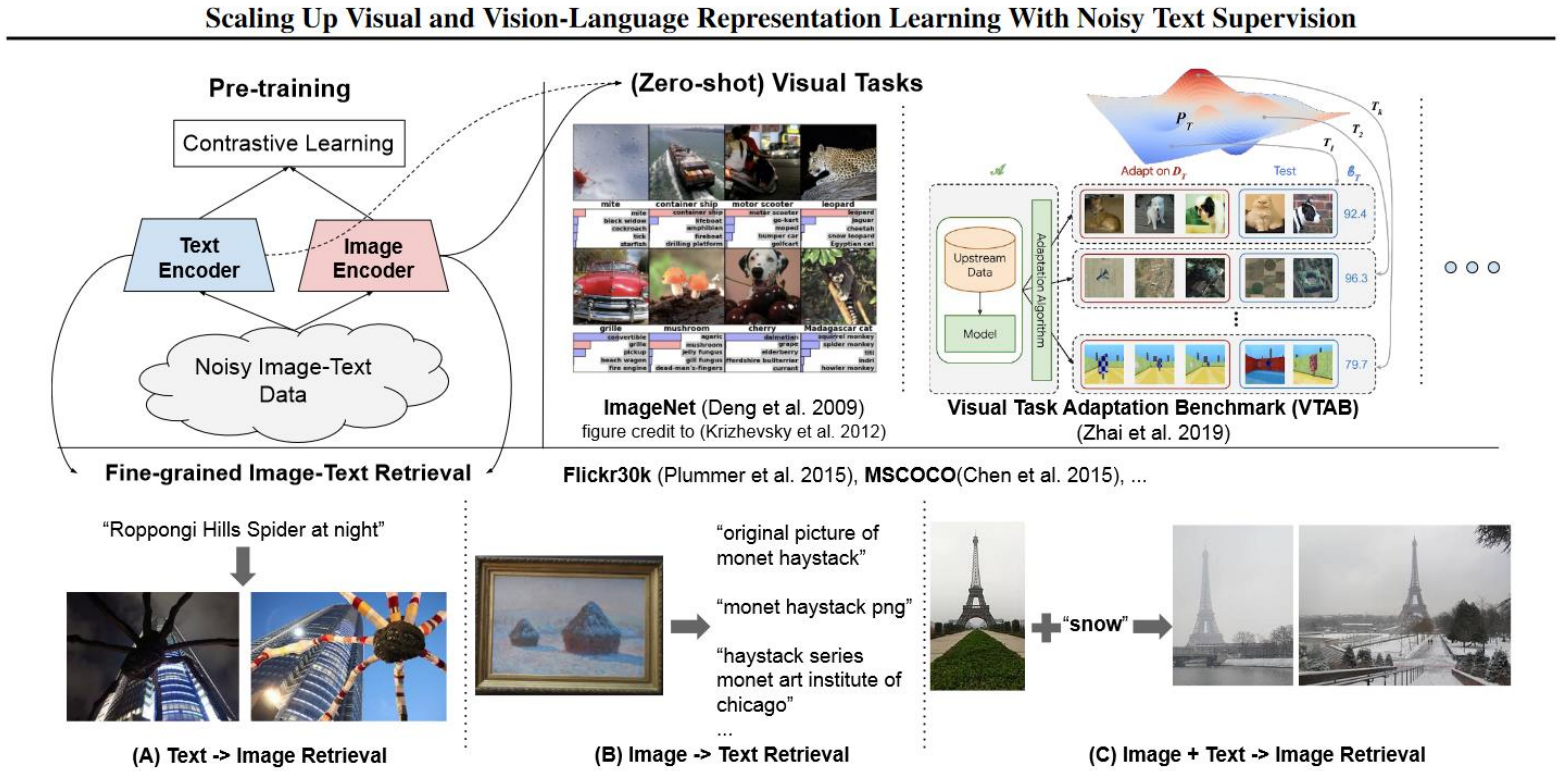
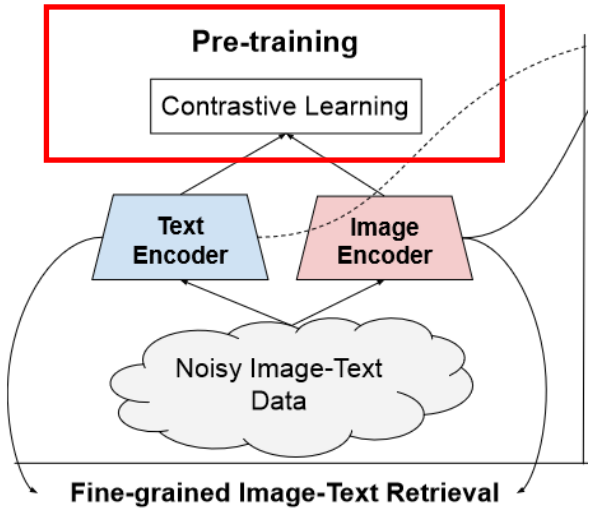


Figure 1. A summary of our method, ALIGN. Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers zero-shot visual classification and cross-modal search including image-to-text search, text-to-image search and even search with joint image+text queries.

- Over 1B noisy image alt-text pairs
- Simple frequency-based filtering
- **Dual encoder contrastive training method**



We minimize the sum of two losses: one for image-to-text classification

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)} \quad (1)$$

and the other for text-to-image classification

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)} \quad (2)$$

Testing and application

Image-Text Matching & Retrieval

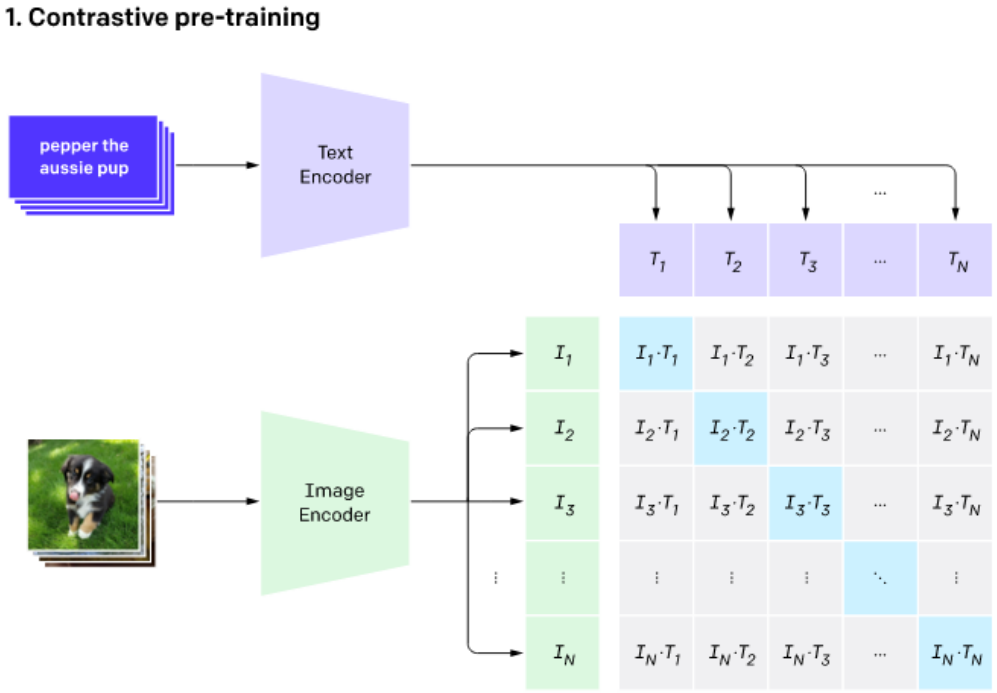


Illustration from CLIP

Testing and application

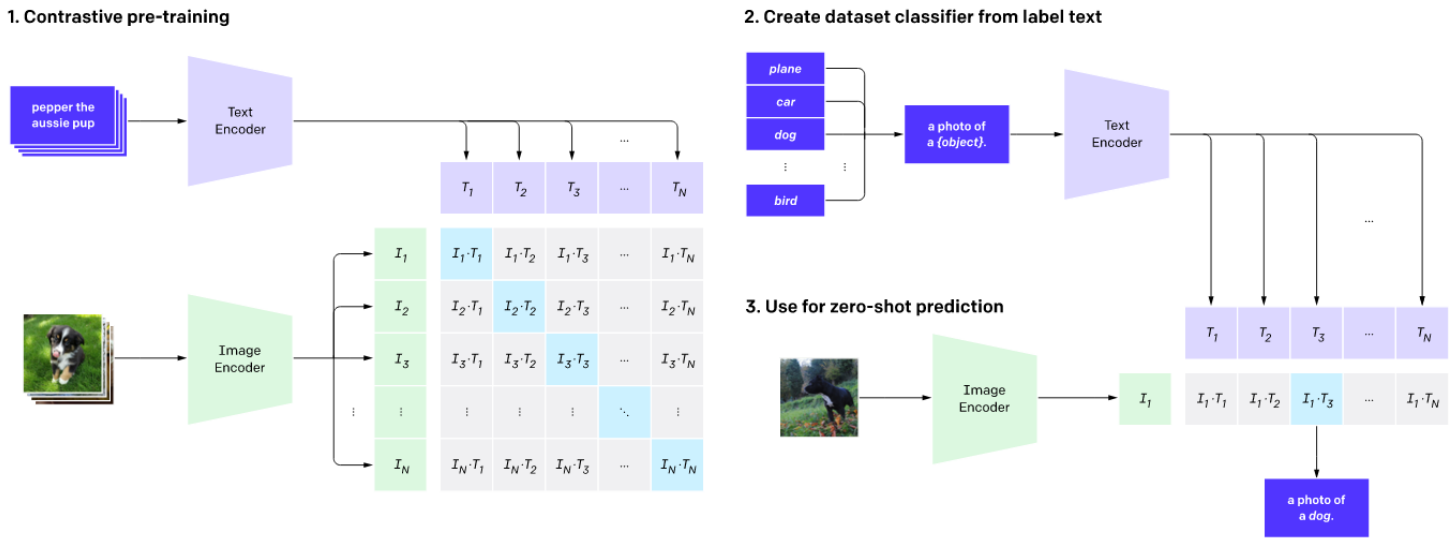
Image-Text Matching & Retrieval

Table 1. Image-text retrieval results on Flickr30K and MSCOCO datasets (zero-shot and fine-tuned). ALIGN is compared with ImageBERT (Qi et al., 2020), UNITER (Chen et al., 2020c), CLIP (Radford et al., 2021), GPO (Chen et al., 2020a), ERNIE-ViL (Yu et al., 2020), VILLA (Gan et al., 2020), and Oscar (Li et al., 2020).

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**	**58.6**	**83.0**	**89.7**	**45.6**	**69.8**	**78.6**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	**89.8**
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**77.0**	**93.5**	**96.9**	**59.9**	**83.3**	**89.8**

Testing and application

Extending to Image Classification



CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a *dog*" and predict the class of the caption CLIP estimates best pairs with a given image.

Illustration from CLIP

Testing and application

Extending to Image Classification

Table 4. Top-1 Accuracy of zero-shot transfer of ALIGN to image classification on ImageNet and its variants.

Model	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**

Table 5. ImageNet classification results. ALIGN is compared with WSL (Mahajan et al., 2018), CLIP (Radford et al., 2021), BiT (Kolesnikov et al., 2020), ViT (Dosovitskiy et al., 2021), NoisyStudent (Xie et al., 2020), and Meta-Pseudo-Labels (Pham et al., 2020).

Model (backbone)	Acc@1 w/ frozen features	Acc@1	Acc@5
WSL (ResNeXt-101 32x48d)	83.6	85.4	97.6
CLIP (ViT-L/14)	85.4	-	-
BiT (ResNet152 x 4)	-	87.54	98.46
NoisyStudent (EfficientNet-L2)	-	88.4	98.7
ViT (ViT-H/14)	-	88.55	-
Meta-Pseudo-Labels (EfficientNet-L2)	-	**90.2**	**98.8**
ALIGN (EfficientNet-L2)	**85.5**	88.64	98.67

Novel Applications

Multilingual ALIGN

Table 11. Multimodal retrieval performance on Multi30K dataset. The metric is the mean Recall (mR).

Model	en	de	fr	cs
zero-shot				
M^3P	57.9	36.8	27.1	20.4
$ALIGN_{EN}$	**92.2**	-	-	-
$ALIGN_{mling}$	90.2	84.1	**84.9**	63.2
w/ fine-tuning				
M^3P	87.7	82.7	73.9	72.2
UC2	88.2	**84.5**	83.9	**81.2**

Novel Applications

IMG + text retrieval

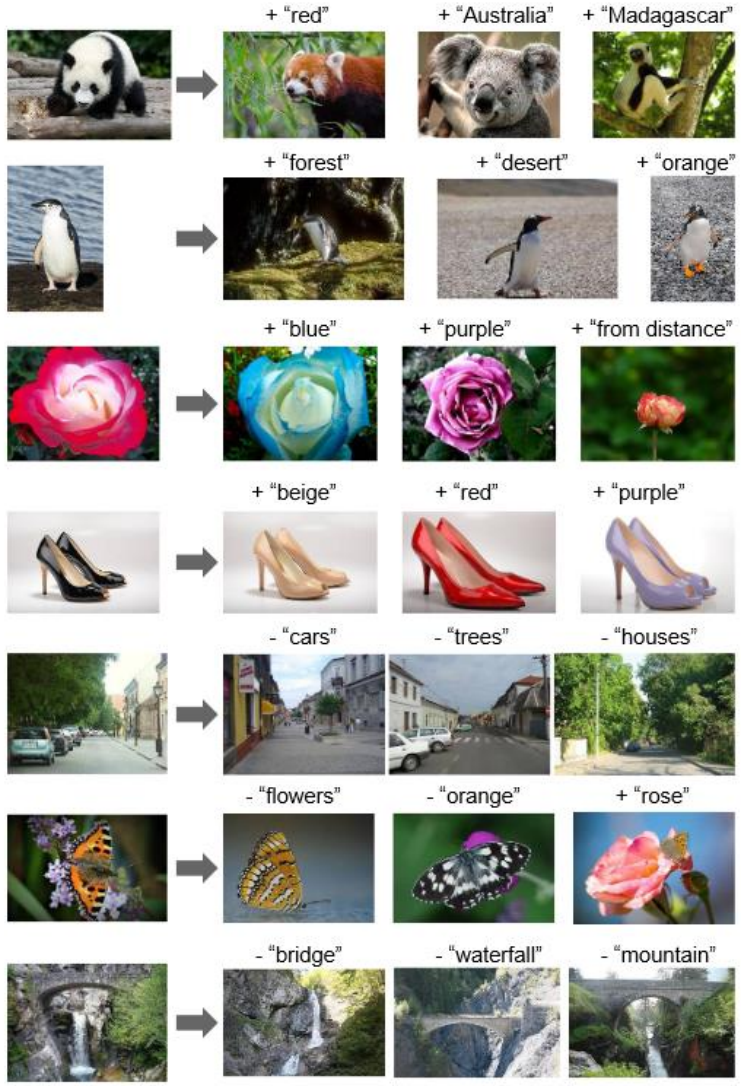


Figure 5. Image retrieval with image±text queries. We add (or subtract) text query embedding to (or from) the image query embedding, and then use the resulting embedding to retrieve relevant images using cosine similarity.