`T-MARS`: Improving Visual Representations by Circumventing Text Feature Learning

LAION dataset

LAION dataset consists of (image, caption) pairs. Authors first randomly sample 500 samples and analysis them:

1. Un-correlated image and caption: 3.7%
2. Correlated visual feature and caption: 46.7%
3. Correlated visual feature and caption, random OCR text: 9.8%
4. Both visual feature and OCR text correlated with caption: 19.1%
5. Correlated OCR text and caption: 20.7%

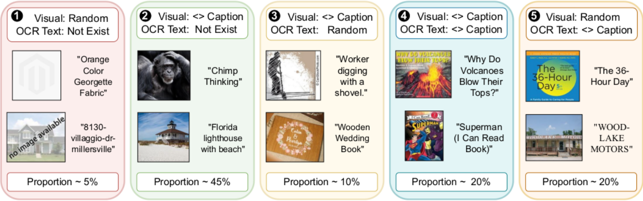

Figure 1: Categorizing LAION dataset

Objective

Properly curating the dataset is necessary!

Formally

Given an image-caption dataset $\mathcal{S} = (i, t)^n$ for contrastive training for CLIP like models, the goal is to curate a subset $\hat{\mathcal{S}} \subset \mathcal{S}$, such that under a fixed compuation budget, model trained on $\hat{\mathcal{S}}$ performs better on zero-shot classification than model trained on $\mathcal{S}$.

T-MARS: Text-Masking and Rescoring

1. **Text Detection:** Apply text detection algorithm to identify the bounding boxes of text regions in the image.
2. **Text Masking:** Mask the text region by replacing it with average RGB value of surrounding pixels.
3. **Rescoring:** Evaluate the cosine similarity between original image and masked image.
4. **Filtering:** Filter out half of the lowest cosine similarity.

Other Contributed Baselines

- **C-SSFT**
  - Second-Split Forgetting Time (SSFT) identify mislabeled examples by finetuning the converged model on validation dataset, and check which examples change its labels the earliest.
  - Similarly, C-SSFT proposes to finetune a pretrained CLIP model on Conceptual-Captions dataset, and keeps only the highest cosine similarity score between pretrained and finetuned models.
- **C-RHO**
  - RHO proposes a loss to select samples that are worth learning, but not yet learned.
  - Similarly, C-RHO proposes to 1) train model for one epoch on entire dataset 2) use model trained on CC3M dataset as a validation model 3) compare cosine similarity scores between two models.

Existing Baselines

- **LAION filtering (LAION-400M):** Evaluate CLIP score using OpenAI's ViT-B/32 and filter samples with score lower than 0.281.
- **CLIP Score:** More stronger CLIP score threshold.
- **Text Match:** Removing all images with text that overlaps with the caption.

Experiment: Setting

- Dataset: six different data pools from LAION-400M with 2M to 64M samples
- Fixed computation budget: total 32M samples
- Evaluation: Zeroshot evaluation on 1) ImageNet 2) ImageNet-ODD 3) VTAB 4) Retrieval
- Backbone: ResNet50 and ViT-B-32

Experiment: Results

Scale	Filtering	Dataset size	ResNet-50 ImageNet	ResNet-50 ImageNet dist. shifts	ResNet-50 VTAB	ResNet-50 Retrieval	ViT-B-32 ImageNet	ViT-B-32 ImageNet dist. shifts	ViT-B-32 VTAB	ViT-B-32 Retrieval
16M	LAION	100%	16.63	15.04	24.20	16.79	09.39	08.46	19.83	12.58
	CLIP Score (@ 50%)	50.0%	15.58	14.28	23.67	16.28	09.02	08.42	20.13	12.60
	Text-Match	86.4%	17.83	15.83	24.63	17.11	10.16	08.89	20.63	12.84
	C-SSFT	90.0%	17.49	15.61	24.90	17.31	10.10	08.94	19.67	13.26
	C-RHO	50.0%	19.46	17.39	26.45	18.60	10.87	09.34	21.22	13.93
	T-MARS	50.0%	20.25	17.71	26.50	18.45	12.09	10.35	22.64	14.15
	T-MARS ∩ C-SSFT	45.2%	20.81	18.28	26.49	18.96	12.56	10.60	21.96	14.36
	T-MARS ∩ C-RHO	27.5%	**21.63**	**18.62**	**26.70**	**19.53**	**12.61**	**10.94**	**23.48**	**14.58**
32M	LAION	100%	21.90	18.90	27.30	20.18	14.98	12.38	23.21	16.03
	CLIP Score (@ 50%)	50.0%	20.84	18.79	25.71	19.54	14.69	12.86	22.81	15.32
	Text-Match	86.4%	23.80	20.70	28.74	21.41	15.96	13.26	24.45	16.44
	C-SSFT	90.0%	22.87	19.85	26.10	21.00	15.55	13.34	22.95	16.40
	C-RHO	50.0%	25.44	21.81	27.65	22.61	16.76	13.98	25.60	17.48
	T-MARS	50.0%	26.73	22.79	29.88	22.62	18.75	15.30	26.71	16.82
	T-MARS ∩ C-SSFT	45.2%	26.89	22.83	28.81	**22.99**	**19.18**	**15.86**	**27.13**	17.82
	T-MARS ∩ C-RHO	27.5%	**27.20**	**23.30**	**30.30**	22.77	19.15	**15.86**	26.93	**18.04**
64M	LAION	100%	26.34	23.24	29.09	23.91	20.37	17.97	27.85	18.83
	CLIP Score (@ 50%)	50.0%	25.66	22.83	29.05	23.36	20.07	17.27	27.55	18.33
	Text-Match	86.4%	29.11	24.94	30.35	25.75	23.11	19.04	28.82	19.37
	C-SSFT	90.0%	28.15	24.13	29.73	25.58	21.80	18.20	27.69	19.54
	C-RHO	50.0%	28.66	24.83	30.13	19.79	23.27	19.23	27.94	21.10
	T-MARS	50.0%	32.47	27.52	33.05	24.99	**25.78**	**21.05**	**31.69**	20.52
	T-MARS ∩ C-SSFT	45.2%	**32.77**	**27.68**	**33.13**	**26.35**	25.63	21.01	30.02	**21.27**
	T-MARS ∩ C-RHO	27.5%	32.63	27.23	32.77	25.57	25.62	20.73	31.57	20.63

Figure 2: Zeroshot accuracy

Experiment: Results

- Taking intersection of curated subsets gained addtional benefits
- Near linear gain on the size of dataset
- Filtering out bad examples are more important than adding new samples

Thank You

Q & A