Context-Aware Multimodal Pretraining

Supplementary Material

A. Experimental Details

In particular, we study various data scales on WebLI [6], ranging from least 1.5 billion up to 15 billion training examples. Our batchsize is set to 32768 by default following optimal suggestions in Zhai et al. [99]. Images for training are resized to 256×256. We use Adam-W optimizer [47] with learning rate of 10^{-3}, weight decay of 10^{-4}, gradient clipping to norm 1, and $\beta_2 = 0.95$ following recommendations in [16, 99]. The full pipeline is implemented in `jax` [3]. Our vision encoder is parameterized as a vision transformer [13]. The corresponding text-encoder (a standard transformer [86]) tokenizes input text using the sentencepiece tokenizer [43] pretrained on the English C4 dataset [66]. If not noted otherwise, **LIxP**-training utilizes $\alpha = 0.9$, $\tau_{\text{ctx}} = 1$, $\tau_1 = 10$ following [99], and $\tau_2 = \tau_1$. While more detailed hyperparameter grid searches would likely provide even better results, we opt for a simple and transferable parameter grid for easiest reuse and replication.

To evaluate both the zero-shot transfer capabilities as well as the few-shot adaptation performance, we measure performance on 21 diverse datasets commonly used for few-shot and domain adaptation works: CUB200-2011 [91], Stanford Cars [41], Cassave [53], CIFAR100 [42], Colorectal Histology [35], DomainNet-{ClipArt, Infograph, Quickdraw, Sketch} [61], DTD [9], EuroSAT [29], Food101 [2], ImageNet2012 [71], ImageNet-Sketch [88], Oxford IIIT Pets [60], Places365 [104], Plant-Village [31], RESISC45 [8], Stanford Dogs [11], SUN397 [94] and UC Merced [97]. Datasets are selected to allow for shot counts of at least up to 28-32, and were queried through the `tensorflow datasets` interface, see `tensorflow.org/datasets/catalog`. For datasets where only a single split was available (such as only train or test), we create a support/test split to allow for sufficient adaptation examples, but ensuring that the number of classes are maintained. The exact splits are provided in Tab. 6. Ablation runs are evaluated on a subset (eleven, ≈half) of our evaluation benchmarks, and cover: CUB200-2011, Stanford Cars, Colorectal Histology, DTD, EuroSAT, Food101, ImageNet2012, ImageNet-Sketch, Oxford IIIT Pets, Places365 and UC Merced — reporting average 16-shot performance.

B. Nearest-neighbor voting classifiers

As described in the main part of this paper, we study multiple different instantiations of nearest-neighbor classifiers based on varying neighbor sample weights w_i. These are:

Plurality-Voted Nearest-Neighbor Classifier, e.g. [54]. We compute k nearest neighbors $\mathbf{X}^k_{\text{spt}}$ to x_{test} and the label for x_{test} is computed as the majority label from the corresponding labels L^k_{spt}. For all our experiments with plurality voting, we fix $k = 32$, but capped to the maximum number of shots for a given few-shot classification task.

Softmax-Voted Nearest-Neighbor Classifier, e.g. [4, 22] For each of the k nearest neighbors $\mathbf{X}^k_{\text{spt}}$ with respect to x_{test}, we assign a softmax sample weight for the i-th neighbor in $\mathbf{X}^k_{\text{spt}}$ (with temperature τ_s) as

$$w_i = \frac{\exp(x_q \mathbf{X}^k_{\text{spt},i} / \tau_s)}{\sum_{j=1}^k \exp(x_q \mathbf{X}^k_{\text{spt},j} / \tau_s)}. \tag{10}$$

We follow existing literature [4, 22, 58, 93] and keep $\tau_s = 0.07$, while setting $k = 32$. The final output logits are then simply computed as the softmax-weighted aggregation of the one-hot labels $\mathbf{L}^k_{\text{spt}}$ of the neighbors.

Rank-Voted Nearest-Neighbor Classifier [22]. This nearest-neighbor classifier computes the weights of k-neighbors following a simple power-function $w_i = 1/(\gamma + \text{rank}_i)$ with offset $\gamma = 2.0$ [22], and rank of support image index i $\mathbf{X}^k_{\text{spt},i}$ within the k neighborhood.

C. Additional Results

In Tab. 7 we provide a finer resolved numerical overview over gains across different shot counts utilized in Fig. 1, as well as the application of additional post-hoc classifiers over exemplary four and sixteen shots in Tab. 8 utilizing training-free variants of [108] and [102].

C.1. Further Buffer Studies

We include an additional buffer design ablation, within which we study the option to populate the key and value contextualization buffer with augmented variants ("*Augmented Entries*") of the input batch $\mathcal{B}_I$ (and consequently removing the self-attention mask $\mathbf{M}$). In this scenario, we distinguish between only populating the buffer with augmented examples ("*Buffer Only*"), as well as jointly training on them with and without the addition of a separate InfoNCE objective. Our results show that gains are only visible if augmented examples are treated as independent entries, effectively mimicking our main contextualization objective in Eq. (8).

Dataset	Type	Support/Test Split	Support Examples	Test Set Size	#Classes
CUB200-2011 [91]	Finegrained, Birds	train, test	5994	5794	200
Stanford Cars [41]	Finegrained, Cars	train, test	8144	8041	196
Cassava [53]	Cassava Leafs	train, test	5656	1885	5
CIFAR100 [42]	Visual Recognition	train, test	50000	10000	100
Col. Histology [35]	Colorectal Cancer Histology	train[:2000], train[:2000]	2000	3000	8
DomainNet - ClipArt [61]	Visual Recognition, ClipArt	train[:30K], test[:20K]	30000	20000	345
DomainNet - Infograph [61]	Visual Recognition, Infographics	train[:30K], test[:20K]	30000	20000	345
DomainNet - Quickdraw [61]	Visual Recognition, Quickdraws	train[:30K], test[:20K]	30000	20000	345
DomainNet - Sketch [61]	Visual Recognition, Sketches	train[:30K], test[:20K]	30000	20000	345
DTD [9]	Textures	train, test	1880	1880	47
EuroSAT [29]	Remote Sensing	train[:22K], train[22K:]	22000	5000	10
Food101 [2]	Finegrained, Food	train[:30K], validation	30000	25250	101
ImageNet2012 [71]	Visual Recognition	train[:100K], validation	100000	50000	1000
ImageNet-Sketch [88]	Visual Recognition, Sketch	test[:30K], test[35K:]	30000	15889	1000
Oxford IIIT Pets [60]	Finegrained, Pets	train, test	3680	3669	37
Places365 (small) [104]	Finegrained, Places	train[:20K], validation[:15K]	20000	15000	365
Plant-Village [31]	Finegrained, Plant leaves	train[:30K], train[30K:]	30000	24303	38
RESISC45 [8]	Remote Sensing	train[:20K], train[20K:]	20000	11500	45
Stanford Dogs [11]	Finegrained, Dogs	train, test	12000	8580	120
SUN397 [94]	Scene Understanding	train[:30K], validation	30000	10875	397
UC Merced [97]	Remote Sensing	train[:1K], train[1K:]	1000	1100	21

Table 6. **Exact default support and test configurations** for all benchmark datasets studied. For most datasets with a clearly defined and available train and test split, we utilize these to define the pool of support examples to sample from for $K - shot$ few-shot studies, and the number of test examples evaluated on. For datasets (such as "Col. Histology" or "Imagenet-Sketch") where only one split was available through `tensorflow.datasets`, we split accordingly into support and test pool. For the remaining datasets (primarily DomainNet), we randomly subsample to maintain comparable support and test pools, though we note no relevant changes in relative performances across methods with either full or subsampled pools.

ViT S/16 (1.5B), Shots →	0	1	2	4	8	16	32
Gain (Perc. Points)	+0.4%	+2.3%	+4.2%	+5.5%	+5.5%	+5.3%	+5.4%

Table 7. Gains for TiP-Adapter using Sig**LIxP** versus SigLIP.

Method →	Proto	TiP	CV-TiP	Plurality	Rank	Softmax	APE	DMN-TF
Gain (4-shot)	+4.6%	+5.6%	+2.4%	+2.2%	+2.0%	+2.9%	+2.6%	+3.1%
Gain (16-shot)	+4.2%	+5.3%	+2.6%	+2.1%	+1.5%	+2.7%	+2.3%	+3.3%

Table 8. Gains for metric-based classifiers for Sig**LIxP** versus SigLIP including more recent training-free variants of APE [108] and DMN-TF [102] on exemplary four and sixteen shots.

Method	Avg. Zero-Shot	Avg. 16-Shot
No Augmentations	50.5	64.1 ± 0.5
Augmented (Buffer only)	48.3	60.0 ± 0.3
Augmented (All)	49.5	63.8 ± 0.3
Augmented (All) + InfoNCE	51.0	63.9 ± 0.3

Table 9. **Additional Buffer Ablations:** Inclusion of augmented entries, with and without additional InfoNCE-style training augmenting the base image-text contrastive training.