Supervised Classification Heads as Semantic Prototypes

Unlocking Vision-Language Alignment via Weight Recycling

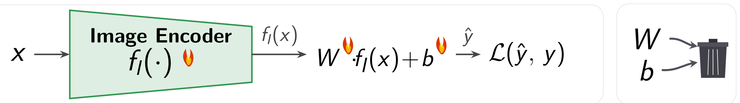
David Méndez[1] Roberto Confalonieri[2] Natalia Díaz-Rodríguez[1]
[1]University of Granada, Spain [2]University of Padova, Italy davidmendez@ugr.es

Main idea. Recycle supervised classifier weights as semantic prototypes to align vision and language with little or no paired image-text data.

Motivation

Vision-language models need a shared image/text space, but large-scale contrastive pretraining depends on massive paired datasets. Post-hoc alignment freezes existing image and text encoders, yet still often needs many image-caption pairs.

Supervised Pretraining



$(x, y) \sim \mathcal{D}$ *supervised pretraining dataset (e.g. ImageNet)* Trainable parameters Discarded

Question. Can supervised image models already contain reusable semantic anchors for language alignment?

Answer: use the rows of the discarded classifier head.

Classifier head

$$Wf_I(x) + b, \quad W \in \mathbb{R}^{C\times d}, \quad b \in \mathbb{R}^C, \quad W = \begin{pmatrix} w_1^\top \\ \vdots \\ w_C^\top \end{pmatrix}.$$

We treat each row w_i as the semantic prototype for class i.

Studying the rows of W as semantic prototypes

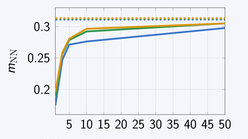
Cosine classifier

Compare acc. of $Wf_I(x) + b$ with a direction-only classifier $\cos(w_i, f_I(x))$ on ImageNet-1K. Cosine classifier on w_is retains almost all the acc.

Model	$Wf_I(x) + b$	$\cos(w_i, f_I(x))$
BEiT	82.71	82.10
CAFormer	80.81	80.28
ConvFormer	80.76	80.08
ConvNeXt	83.29	82.67
TinyViT	82.31	81.30

m_{NN} alignment

We use the m_{NN} metric to see if classifier rows w_i or averaged image embeddings preserve neighborhoods better with respect to text embeddings of class names.

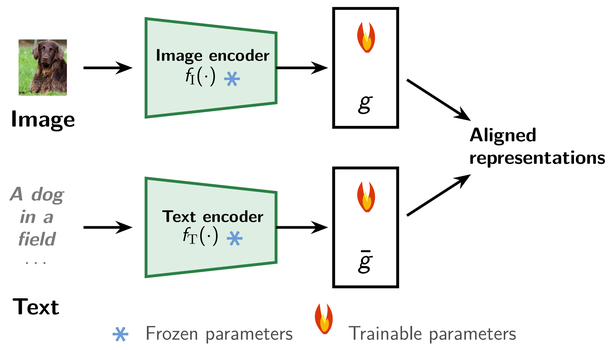


Intuition: neural collapse suggests that, under idealized conditions, classifier weights approach class means.

The classifier head is not just a disposable task head: its rows are reusable semantic anchors.

Post-hoc alignment

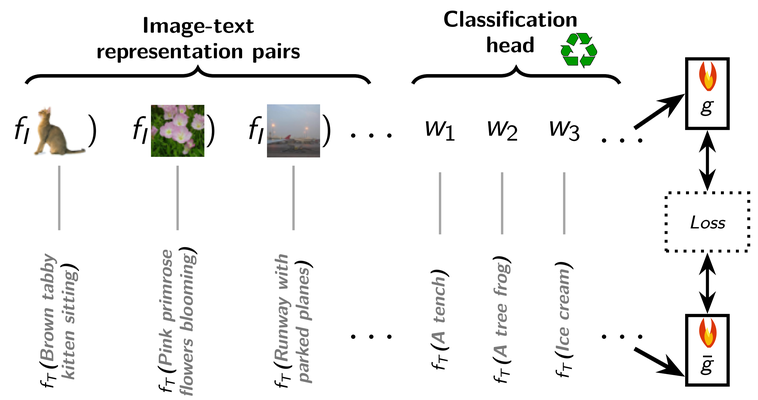
Learn **lightweight mappings** g and $\bar{g}$ between fixed, independently pretrained image and text encoders, so representations with the same semantic content are mapped into a shared space.



Weight Recycling

Classifier rows become alignment data.

- Each row w_i of the supervised classifier head is treated as a semantic prototype for class i.
- We pair w_i with the text embedding of its class name $f_T(t_i)$, creating $\mathcal{D}_{\text{weights}} = \{(w_i, f_T(t_i))\}$.
- The image and text encoders stay frozen; only the lightweight mappings g and $\bar{g}$ are trained.



No image-text pairs. We can endow a vision encoder with vision-language capabilities without any image-text pairs.

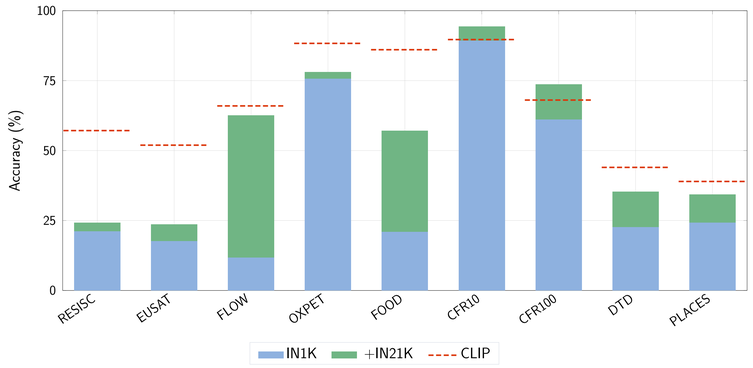
Data augmentation. Weights are compatible with images and can be used to augment img-txt pairs to boost alignment.

Setting I: No image-text pairs

Train g and $\bar{g}$ using classifier weights and class names only. We experiment with different **frozen vision encoders**.

Downstream zero-shot classification evaluation.

- Here we present results for the supervised pretrained BEiT-B/16.
- Weight-only alignment achieves competitive performance with CLIP on some benchmarks despite not using image-text pairs.
- Even though ImageNet-1K classes are cleaner (balanced, with no class hypo/hypernym relations), ImageNet-21K weights provide stronger semantic coverage.



Red dashed lines show CLIP reference performance; the comparison concerns only the additional post-hoc alignment step.

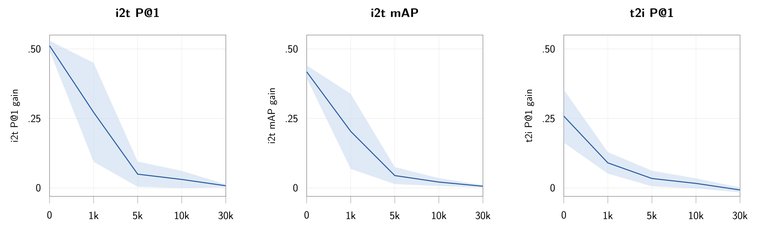
Downstream retrieval evaluation on Flickr30K and COCO.

- Weight-only alignment yields non-trivial retrieval results, outperforming the same budget of image-text pairs.

Discarded classifier heads can endow a vision encoder with vision-language capabilities without paired image-text data.

Setting II: Data Augmentation

ImageNet-21K classifier weights improve downstream retrieval most in the low-pair regime on Flickr30K.



Flickr30K retrieval; the x-axis is the number of image-caption pairs used for alignment.

For zero-shot classification, we add one image-caption representation pair per class to the ImageNet-21K weights.

Zero-shot classification gains from adding 1 image-caption pair/class to ImageNet-21K weights.

	RESISC	EUSAT	FLOW	OXPET	FOOD	CFR10	CFR100	DTD	PLACES
Δ acc.	+8.16	+7.18	+11.11	+2.92	+7.64	−0.29	+0.38	+3.57	+0.44

Classifier weights are complementary to real image-text pairs and are most useful exactly where paired data is hardest to collect.

Further Analyses

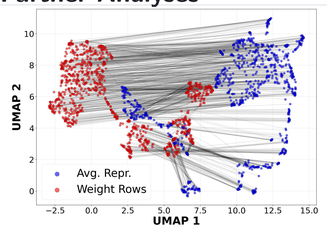


Image-Weights modality gap. Classifier weights and image embeddings occupy distinct regions of the frozen image feature space.

- Centroid-distance permutation shows a significant gap ($p < 0.001$).
- Single-layer MLP separates the two representation types with 99.75% accuracy.
- Images and weights occupy two different regions of space in the UMAP visualization.

Compatibility. Yet the weights remain compatible with image embeddings to support alignment augmentation.

Source matters. ImageNet-21K provides the strongest general-purpose anchors; narrower supervised heads still help, but transfer less broadly.

Weight vs Img. on equal budget. With the same number of alignment pairs, classifier weights provide stronger alignment data than averaged image representations for downstream retrieval and classification.

Limitations and Future Work

- Weights and image embeddings occupy different regions, so better methods are needed to combine them.
- Combining heads from multiple supervised datasets is promising, but only when they share the same frozen feature space; developing methods to combine weights from different backbones is an interesting direction.

Conclusion

- **Supervised classifier heads** are not just disposable task heads: they are **reusable semantic prototypes** for resource-efficient vision-language alignment.
- **These semantic anchors** allow endowing a vision encoder with **text capabilities** without any **image-text data**.
- **They are compatible with image representations** and can serve to **augment them in a low-data regime**.
- This compatibility happens **despite the modality gap** between image and weight representations. In addition, weights are **more aligned to text representations** and better for alignment to text under equal budget.

Selected references.
Huh et al. *Position: The Platonic Representation Hypothesis.* ICML, 2024.
Papyan et al. *Prevalence of Neural Collapse during the terminal phase of deep learning training.* PNAS, 2020.
Liang et al. *Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning.* NeurIPS, 2022.
Li et al. *CSA: Data-efficient Mapping of Unimodal Features to Multimodal Features.* ICLR, 2025.

Cofinanciado por la Unión Europea · Junta de Andalucía · Financiado por la Unión Europea NextGenerationEU · Gobierno de España, Ministerio para la Transformación Digital y de la Función Pública · Plan de Recuperación, Transformación y Resiliencia · España | digital 2026 · repsol · Universidad de Granada · UNIMORE

Funding and acknowledgement. This publication is part of the grant DGP_PRED_2024_00569, funded by Junta de Andalucía/CUII and FSE+. This work was also supported by the Arqus Talent Scholarship, the 2022 Leonardo Grant from the BBVA Foundation, and TSI-100927-2023-1 under the Recovery, Transformation and Resilience Plan funded by the European Union NextGenerationEU.