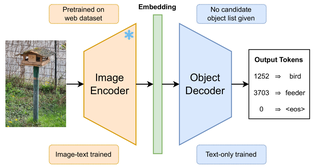
Unconstrained Open Vocabulary Image Classification: Zero-Shot Transfer from Text to Image via CLIP Inversion

Philipp Allgeuer, Kyra Ahrens, and Stefan Wermter
{philipp.allgeuer, kyra.ahrens, stefan.wermter}@uni-hamburg.de
Knowledge Technology Group, Dept. of Informatics, University of Hamburg, Germany

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

NOVIC: uNconstrained Open Vocabulary Image Classification

TL;DR: Given an image and **nothing else** (i.e. no prompts or candidate labels), NOVIC can generate an **accurate fine-grained** textual classification label in **real-time**, with **equal coverage** of the vast majority of the English language.



NOVIC is:

- **Generative:** Image classifications are produced directly as free-form text
- **Unconstrained:** Does not require any predefined list of candidate classifications
- **Open vocabulary:** Trained using a full English dict of nouns (no core/priority vocab)
- **Text-only trained:** No images are used during training, only synthetic text!
- **Efficient:** Trains in 1-3 days on a single RTX A6000 (FAST as no images required)
- **Real-time:** 26ms inference time per single image, or 7ms per image if batched
- **Reliable:** Achieves fine-grained prediction scores up to 87.5% on arbitrary images
- **Scalable:** Performance scales well with performance of underlying CLIP model
- **Open source:** Detailed GitHub repository with commands to reproduce results

Contributions: (a) Novel open vocabulary object decoder model and training scheme, (b) Object noun dictionary, (c) Two new large-scale synthetic caption-object datasets, (d) Three new image sets for evaluation of open vocabulary image classification

Motivation

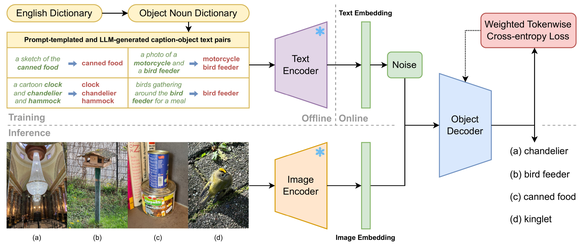
- **Existing methods** for open vocabulary image classification generally **require either**:
  - An explicit prompt or category specification(s) from the user
  - A predefined or inherent list of candidate classification labels
  - Training or fine-tuning on a dataset of limited cover (e.g. LVIS, CC-3M, CC-12M)
- **How can we determine what is in an image without any context or restriction?**
  - **Example:** Arbitrary web images, Vision of household robot, Anomaly detection
- **Image classification** should be **FAST** and **energy efficient**:
  - Required for real-time vision applications, mobile robotics, inferencing on consumer-grade GPUs, applications requiring large amounts of data, and more
  - Multimodal LLMs can perform image classification, but have large practical limitations: Compute, response time, throughput, GPU hardware, energy use, …

References & Acknowledgements

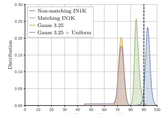
[1] X. Huang, Y. Zhang, J. Ma, et al. Tag2Text: Guiding Vision-Language Model via Image Tagging. In *12th International Conference on Learning Representations*, 2024.
[2] Y. Zhang, X. Huang, J. Ma, et al. Recognize Anything: A Strong Image Tagging Model. In *Proceedings of the IEEE/CVF CVPR Workshops*, pp. 1724–1732, 2024.

The authors gratefully acknowledge the support from the DFG (TRR 169 – CML) and the European Commission (TRAIL).

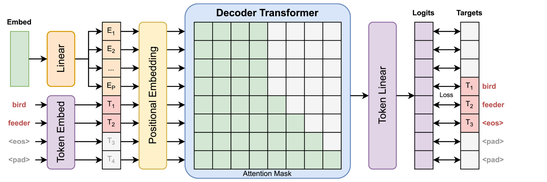
Approach & Architecture



- **Training:** Given a caption-object text pair, the **caption** is embedded, augmented with noise, and used to train an **object decoder** to predict the **object noun**
  - Multiple target object nouns are trained per caption in parallel
  - Noise overcomes **huge** image-text **modality gap**
- **Inference:** Object decoder gets CLIP-embedded images as input instead of caption text embeddings
  - Non-trivial zero-shot transfer from text to images
  - Object decoder inverts the frozen CLIP text encoder for the specific task of object prediction



Object Decoder:



- **Autoregressive** text transformer with learned positional and token embeddings
- Input CLIP embedding vector is linearly projected to form P **prefix tokens**

Training & Evaluation Datasets

- **Object noun dictionary:** Dictionary of simple/compound English object nouns based on WordNet, GCIDE, image datasets, and GWTWC (word freqs), with 42919 entries and 96526 variants
- **Prompt-templated caption-object pairs:** Uses 85+68 prompt templates to construct captions from the object noun dictionary
- **LLM-generated caption-object pairs:** Leverages the contextual knowledge of LLMs to place object nouns in relevant contexts
- **Evaluation image sets** for open vocabulary image classification:
  - **World:** 272 original images from 12 people in 10 countries, including diverse, atypical, deceptive and indirect objects
  - **Wiki:** 1000 Wikipedia lead images (sampled from 18K)
  - **Val3K:** 3000 ImageNet-1K validation images (3 per class)

Zero-Shot Classification Performance

NOVIC training results:

- The object decoder with ≈12M params is trained for 1-3 days using a **single** RTX A6000 on 92M noun targets, with up to 1B samples seen over 18 epochs
  - **Extremely efficient:** Exactly the same training applied to ImageNet-1K gives 88.21% in only 19 minutes!
- **Inference** requires only 5GB GPU memory, and 7-26ms per image depending on whether the batch size is 1 or larger

Category-constrained scores:

- Can evaluate NOVIC on **classification datasets** by restricting the beam search to only the available class object nouns
  - Knowing which classes are available is a **strong** prior for traditional models (e.g. if *kinglet* is the only bird class), and partially helps the model focus on the right object
- Restricting NOVIC to a fixed vocabulary however **forces** it to choose objects nouns it does not even consider to be best
  - A significant **underestimation** of NOVIC's capabilities

Open vocabulary scores:

- Given model predictions on an image set, human (-H) and LLM (-L) annotations are performed, e.g. **World-H**, **Wiki-L**
  - Cannot pre-annotate every possible correct English noun for every visible entity in every part of every image
- 87.5% is a **strong** result considering the model must work for **any** conceivable image and without **any** contextual clues
  - Regularly makes correct predictions as detailed as *brazil nut tree*, *electron microscope*, and *egretta caerulea*
- For testing purposes, training dataset can be filtered to only include words above a certain frequency threshold (FT)

Tag	Annotation	Points	-H	-L	Primary Score	Prediction Score	Overall Score
(P)	Correct primary	1.0	✓	✓	✓		✓
(S)	Correct secondary	0.8	✓	✗	✗	✓	✓
(CP)	Close primary	0.5	✓	✓	✓	✓	✓
(CS)	Close secondary	0.4	✓	✗	✗	✓	✓
(I)	Incorrect	0.0	✓	✓	✓	✓	✗
	Coarse specificity	×0.5	✓	✗	✗	✗	✓

Table 1. **Open vocabulary scoring system:** The top model prediction for each image is annotated as one of five categories, depending on correctness and whether the prediction refers to a main element of the image, and assigned corresponding points. The *Prediction Score* is the mean points per image as a percent. For the *Overall Score*, the points for an image are halved if the prediction is overly coarse. The *Primary Score* counts secondary predictions as incorrect. The LLM annotations (-L) only support the primary score.

Annotations	FT9	FT6	FT2	FT0
ImageNet-1K	46.52	69.47	67.92	60.72
Open (correct)	**70.33**	**75.04**	**72.90**	**66.19**
Open (correct + close)	*71.90*	*76.50*	*74.35*	*67.86*

Table 2. **Performance on Val3K dataset (% mean of 3),** showing that category-constrained scores are a significant underestimation of the true open vocabulary performance of NOVIC.

Configuration	World-H	Wiki-H	Wiki-L	IN1K
FT9	76.13	71.20	63.57	46.76
FT6	78.42	**72.33**	62.62	**69.50**
FT2*	**78.92**	72.03	**63.97**	68.11
FT0	73.98	71.48	61.27	60.56
$\mathcal{M}_1$ dataset	70.11	-	50.05	63.88
+ LLM captions	72.16	-	55.37	66.06
+ $\mathcal{M}_2$ dataset	78.37	-	61.60	67.97
+ $\mathcal{M}_3$ dataset*	**78.92**	**72.03**	**63.97**	**68.11**
$\mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3$	77.56	71.60	62.08	65.36
$\mathcal{M}_3$ + LLM captions	77.88	71.28	62.93	61.12
Gauss 1.00	19.44	-	15.62	13.36
Gauss 1.75	70.37	-	52.07	49.23
Gauss 2.50	76.26	-	62.10	67.94
Gauss 3.25	77.57	70.95	62.38	66.46
+ 15% Uniform 45–75*	**78.92**	**72.03**	**63.97**	**68.11**
Uniform 45–75	76.46	-	63.27	64.75
DataComp-1B B/16	76.85	-	63.55	64.17
SigLIP B/16*	78.92	72.03	63.97	68.11
DataComp-1B L/14	84.27	-	70.35	70.05
SigLIP SO/14	**87.49**	-	70.28	74.85
DFN-5B H/14-378	87.08	-	**74.88**	**75.82**
→ FT0	86.41	-	74.48	75.68

Table 3. **Frequency threshold, dataset, training noise, and CLIP model ablations (% mean of 3).** The prediction scores on World-H and Wiki-H, the primary score on Wiki-L, and the ImageNet-1K classification score (IN1K) for NOVIC in various configurations. The asterisks show the nominal ablation baseline. Some Wiki-H scores are unavailable due to the high cost of manual annotation.

Comparison with Related Work

Model	Prompt-free Classification	Text-only Training	Seen Objects	Used Objects	RAM Vocab	Primary Score	Prediction Score		Overall Score
Tag2Text	✗	✗	3 429	176	91.9	62.32	←15.29	77.61	−13.49, 64.12
RAM	✗	✗	4 585	160	100.0	57.35	←17.65	75.00	−15.86, 59.14
RAM (FT2)	✗	✗	11 897	188	76.8	64.71	←11.61	76.32	−13.84, 62.48
RAM (FT0)	✗	✗	**42 919**	188	74.3	63.97	←11.03	75.00	−13.84, 61.16
NOVIC (FT9)	✓	✓	2 919	235	66.9	74.82	←2.72	77.61	−2.79, 74.39
NOVIC (FT6)	✓	✓	5 899	252	58.5	76.47	←2.71	78.68	−1.01, 77.67
NOVIC (FT2)	✓	✓	11 897	253	52.6	77.94	←1.62	79.56	−0.64, 78.92
NOVIC (FT0)	✓	✓	**42 919**	251	49.6	71.14	←2.53	74.67	−1.10, 73.57
NOVIC† (FT2)	✓	✓	11 897	**263**	55.5	86.77	←1.17	87.94	−0.81, 87.13
NOVIC† (FT0)	✓	✓	**42 919**	260	**48.5**	**86.95**	←1.92	**88.27**	−0.37, **87.90**

Table 4. **Open vocabulary World-H results in comparison to baselines (best of 3).** †DFN-5B H/14-378 CLIP model.

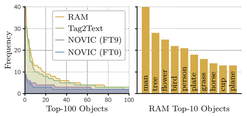
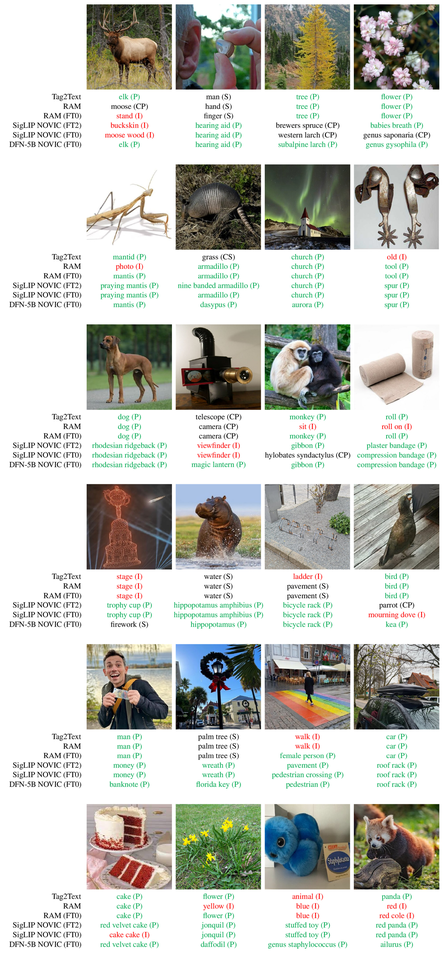


Figure 1. **Diversity comparison of predicted object nouns.** Plots showing the sorted top frequency counts of the object nouns predicted for Wiki-H. Even when trained on FT9, NOVIC shows much greater diversity and has a less peaked frequency distribution. **Right:** The top-10 nouns for RAM are generic and overused.

- **Tag2Text** [1] and the **Recognize Anything Model (RAM)** [2] have:
  - Low *Primary/Overall Scores* relative to *Prediction Score* ⇒ **Many secondary or coarse**
  - Limited tendency to use object nouns beyond the core set, and even less so correctly
  - Tendentially **simple** and **repetitive** predictions, with a severe lack of diversity and specificity
  - **Example:** 88% of Imagewoof validation samples predict *dog* (NOVIC predicts 90+ dog breeds)
- **NOVIC is fine-grained, prompt-free, and outperforms Tag2Text/RAM** (especially with DFN-5B†)

Qualitative Results



Language generalization abilities:

- Photo of willow tree taken in New Zealand ⇒ *New Zealand willow*
- **Pseudo-reading:** *staphylococcus* ⇒ *stuffed toy* when text is blurred
- NOVIC can achieve high performance in **any non-English language** without requiring **ANY** images or captions in that language!

⇐ arXiv

GitHub ⇒

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4