FICUS: Few-shot Image Classification with Unsupervised Segmentation

Jonathan Lys, Frédéric Lin, Clément Béliveau, Jules Decaestecker, Yassir Bendou, Aymane Abdali, Bastien Pasdeloup

IMT Atlantique, Lab-STICC

August 27, 2024

The problem we consider

Disambiguating image embeddings

- Image classification: embedding of reference images
- Embedding quality matters
- FSL: Only ~ 5 images per class
- Challenge: ambiguous images
- Using segmentation to improve embeddings

A problem with ambiguous images

- Non-ambiguous classification datasets (one instance per image)
- Nontrivial embedding of several concepts
- Overfitting in few-shot

Figure. 1: Non ambiguous vs ambiguous image

Contributions

Approach: FICUS

- 2 pre-trained foundation models for segmentation/embedding
- Disambiguation of visual concepts

Results

- No model training
- Boost few-shot accuracy (even on non-ambiguous images)

Few-shot image classification

Definition

- Classify new images with minimal supervision and limited data
- Most common setting: 5-ways, $1 - 10$ shots

Terminology

- Number of classes: *ways*
- K images per class: *shots* in the support set
- Query/Support
- Inductive/Transductive

Setting

- Inductive, 5 ways, $1 - 10$ shots

Image segmentation

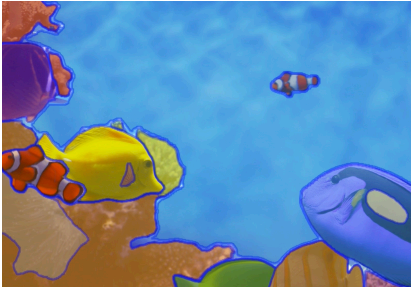


Figure. 2: Segment Anything [1]: a promptable foundation segmentation model

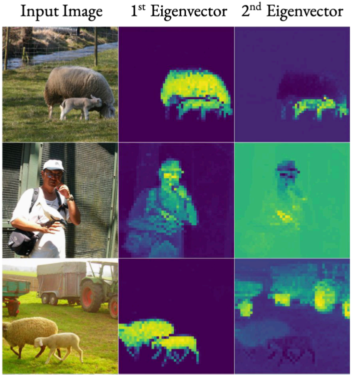


Figure. 3: Deep Spectral Method [2]: Unsupervised Semantic Segmentation and Localization

State-of-the-art

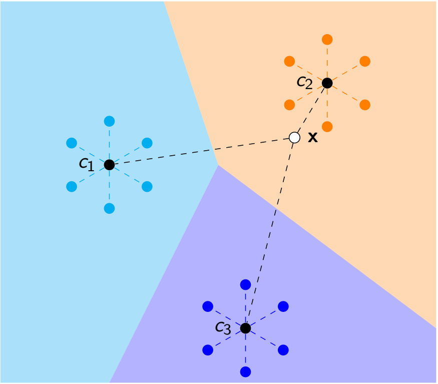


Figure. 4: Nearest Class Mean, [3]

Segment Anything Automatic Mask Generator (AMG)

- Regular grid of points
- Deduplicate with NMS

Rationale and approach

One instance $\rightarrow$ one embedding
Multiple instances $\rightarrow$ multiple embeddings

- Use DSM and SAM to localize objects
- Embed the crops of each instance
- Modified NCM with multiple embeddings

Method outline (1/3)

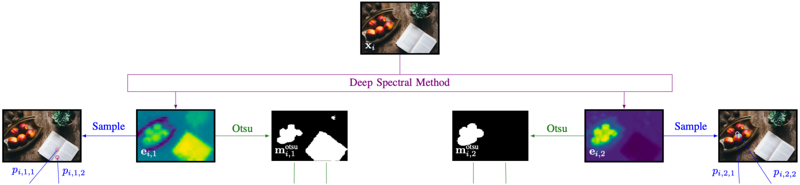


Figure. 5: Eigenmap extraction, Otsu thresholding, point sampling

Method outline (2/3)

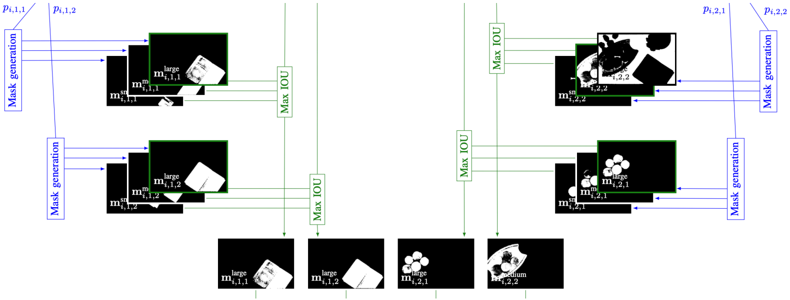


Figure. 6: Mask generation and selection

Method outline (3/3)

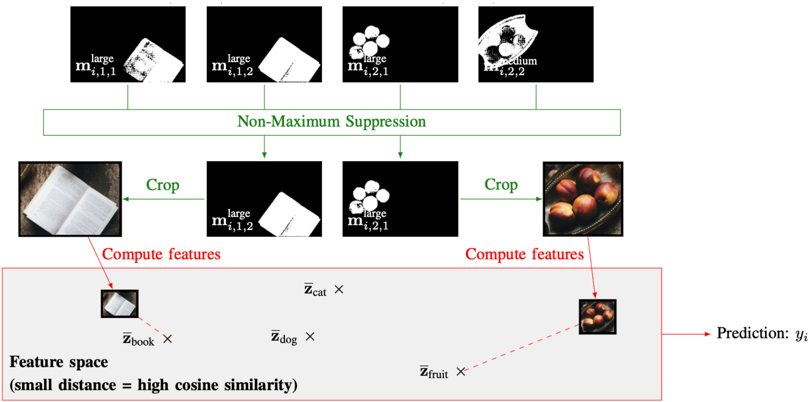


Figure. 7: Deduplication (NMS), cropping and embedding, prediction

Experimental settings

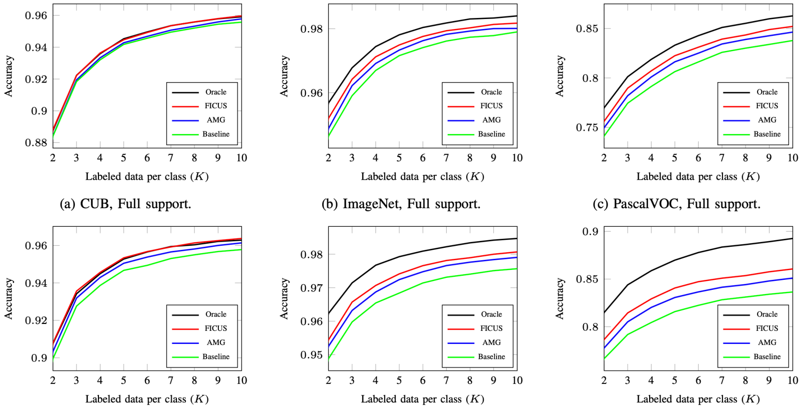
- Datasets: ImageNet, CUB, PascalVoc [4]–[6]
- Support (Full/Human)
- Query (Baseline, AMG, FICUS, Oracle)
- Metric: 5 ways 1 – 10 shots accuracy
- Image encoder: DiNO [7]

Results (1/2)

Support		Full				Human			
Query		Baseline	AMG	FICUS	Oracle	Baseline	AMG	FICUS	Oracle
ImageNet	$K=1$	90.22 ± 0.67	90.53 ± 0.67	$\mathbf{90.83} \pm \mathbf{0.68}$	91.50 ± 0.66	91.20 ± 0.61	91.71 ± 0.61	$\mathbf{92.06} \pm \mathbf{0.61}$	92.90 ± 0.58
	$K=5$	97.16 ± 0.29	97.33 ± 0.30	$\mathbf{97.49} \pm \mathbf{0.28}$	97.81 ± 0.27	96.84 ± 0.31	97.24 ± 0.29	$\mathbf{97.41} \pm \mathbf{0.29}$	97.93 ± 0.26
CUB	$K=1$	80.43 ± 1.00	80.18 ± 1.02	$\mathbf{80.51} \pm \mathbf{1.05}$	80.44 ± 1.06	82.53 ± 0.94	83.12 ± 0.97	$\mathbf{83.18} \pm \mathbf{0.98}$	83.12 ± 1.00
	$K=5$	94.17 ± 0.47	94.38 ± 0.46	$\mathbf{94.40} \pm \mathbf{0.49}$	94.45 ± 0.49	94.73 ± 0.43	94.98 ± 0.46	$\mathbf{95.26} \pm \mathbf{0.46}$	95.27 ± 0.45
PascalVOC	$K=1$	66.43 ± 1.07	67.22 ± 1.07	$\mathbf{67.86} \pm \mathbf{1.08}$	69.39 ± 1.11	69.95 ± 0.98	70.95 ± 0.99	$\mathbf{71.64} \pm \mathbf{1.00}$	74.36 ± 0.98
	$K=5$	80.63 ± 0.79	81.63 ± 0.80	$\mathbf{82.25} \pm \mathbf{0.80}$	83.31 ± 0.80	81.58 ± 0.77	83.06 ± 0.75	$\mathbf{84.06} \pm \mathbf{0.73}$	86.98 ± 0.66

Table. 1: Accuracies for $K \in \{1, 5\}$ per dataset per support setting

Results (2/2)



(a) CUB, Full support. (b) ImageNet, Full support. (c) PascalVOC, Full support.

(d) CUB, Human support. (e) ImageNet, Human support. (f) PascalVOC, Human support.

Figure. 8: Few-shot accuracy

Examples

Figure. 9: DSM/AMG points

Figure. 10: Interpretable masks

Conclusion

Advantages

- Efficient method
- No need for model training

Limitations

- Confidence intervals to improve
- Improvable settings (model size, local encoding)

Future work

- Transductive setting
- Other ambiguous datasets

References

[1] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 4015–4026.

[2] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization," en, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 8354–8365, ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.00818.

[3] Y. Bendou, Y. Hu, R. Lafargue, *et al.*, "EasyEnsemble Augmented-Shot-Y-Shaped Learning: State-of-the-Art Few-Shot Classification with Simple Components," en, *Journal of Imaging*, vol. 8, no. 7, p. 179, Jun. 2022, ISSN: 2313-433X. DOI: 10.3390/jimaging8070179.

[4] O. Russakovsky, J. Deng, H. Su, *et al.*, *ImageNet Large Scale Visual Recognition Challenge*, en, arXiv:1409.0575 [cs], Jan. 2015.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "Caltech-UCSD Birds-200-2011 (CUB-200-2011)," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[7] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," en, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9630–9640, ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00951.

https://github.com/lin-frederic/FICUS

Thank you!