# Asynchronous Perception Machine For Efficient Test Time Training

Rajat Modi, Yogesh Singh Rawat
CRCV, University Of Central Florida

NEURAL INFORMATION PROCESSING SYSTEMS

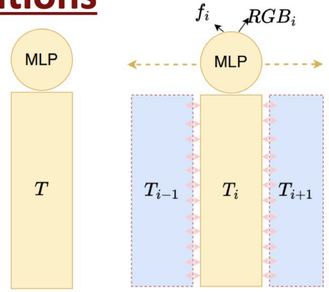
## Hinton's Islands of Agreement

*A Static Image is A Rather Boring Video*- Forward Forward, Some Preliminary Investigations.

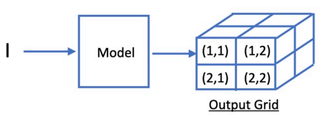
- Take a static image.
- Repeat it along time. It becomes a boring video.
- Give it to a video transformer (Mvitv2). Tsne on (H,W,D) -> (H,W,3)
- Backproject on RGB hypercube.
- Islands emerge using ONLY a classification objective.

## Contributions

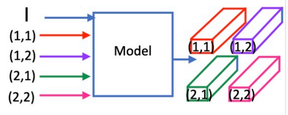
- A new mechanism: Folding-Unfolding.
- **First** implementation of GLOM.
- **Asynchronous Perception Machine/s**
- A new deep learning architecture that might help us move **BEYOND**:
  - Data Augmentation
  - Routing
  - Pretext Tasks
  - Parallel Perception
  - Softmax
  - Boxes
  - Encoder-Decoder




## Reformulating Feature Grids



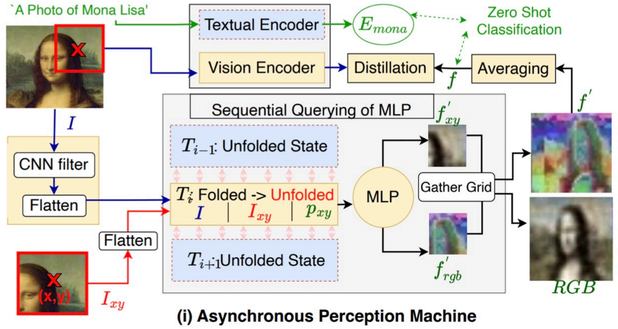

**Classical Perception**




**Location-Aware Queries**

- Layer Skipping: Take shortcut from first layer to last layer feature grid of a model.
- Location- Aware Queries: Querying a model with location-aware queries instead of predicting full grid. Memory Efficient.
- Independence of columns: Each location-query is independent of another. Allows parallelizing a single image across multiple GPUs.

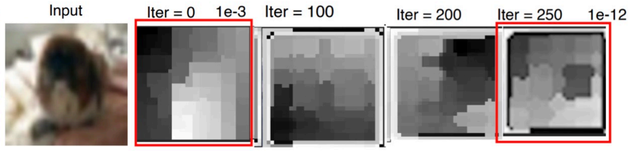
## Architecture For TTT




**(i) Asynchronous Perception Machine**

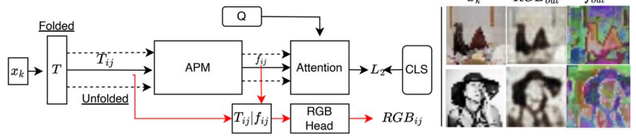
- Take an input Image I.
- Distill the representation from teacher's vision encoder ONCE.
- Overfit MLP for several iterations.
- Classify test sample via zero-shot textual encoder.
- Reset weights and move to next sample.

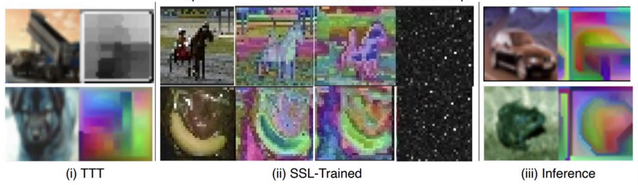
## Single-Sample Overfitting




- Overfitting on a single distilled token representation leads to islands of agreement.

## RGB Decoding

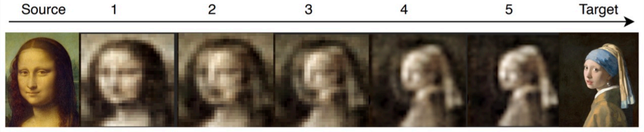



(i) TTT (ii) SSL-Trained (iii) Inference

- APM can reconstruct RGB for any image in-the-wild.
- TTT on a single test-sample with random weights leads to semantically-aware features.
- Scales up to COCO, and yields similar features as Dinov2 in a single forward pass.

## Perceptual Interpolation




- Towards validating GLOM's insight: input percept is a field.

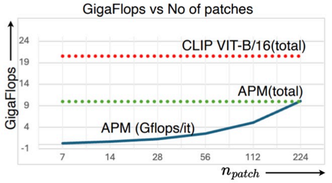
## Quantitative Results

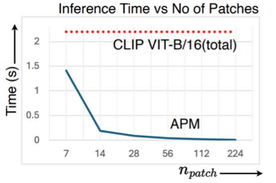
| Method | P | ImageNet Top1 acc. ↑ | ImageNet-A Top1 acc. ↑ | ImageNet-V2 Top1 acc. ↑ | ImageNet-R Top1 acc. ↑ | ImageNet-Sketch Top1 acc. ↑ | Average | OOD Average |
|---|---|---|---|---|---|---|---|---|
| CLIP-ViT-B/16 | ✗ | 66.7 | 47.8 | 60.8 | 73.9 | 46.0 | 59.1 | 57.2 |
| Ensemble | ✗ | 68.3 | 49.8 | 61.8 | **77.6** | 48.2 | 61.2 | 59.4 |
| TPT | ✗ | **68.9** | **54.7** | 63.4 | 77.0 | 47.9 | 62.4 | 60.8 |
| APM (Ours) | ✗ | 68.1 | 52.1 | **67.2** | 76.5 | **49.3** | **62.6** | **61.2** |
| CoOp | ✓ | 71.5 | 49.7 | 64.2 | 75.2 | 47.9 | 61.7 | 59.2 |
| CoCoOp | ✓ | 71.0 | 50.6 | 64.0 | 76.1 | 48.7 | 62.1 | 59.9 |
| TPT + CoOp | ✓ | 73.6 | 57.9 | 66.8 | 77.2 | 49.2 | 64.9 | 62.8 |
| TPT + CoCoOp | ✓ | 71.0 | 58.4 | 64.8 | 78.6 | 48.4 | 64.3 | 62.6 |
| CLIP VIT-L/14 | ✗ | 76.2 | 69.6 | 72.1 | 85.9 | 58.8 | 72.5 | 71.6 |
| APM (Ours) | ✗ | **77.3** | **71.8** | **72.8** | **87.1** | **62.2** | **74.2** | **73.4** |
| OpenCLIP-VIT-H/14 | ✗ | 81.6 | 79.1 | 80.7 | 92.9 | 72.8 | 81.4 | 81.3 |
| APM (Ours) | ✗ | **84.6** | **84.2** | **83.9** | **94.9** | **77.1** | **84.9** | **85.0** |

| | P | brigh | cont | defoc | elast | fog | frost | gauss | glass | impul | jpeg | motn | pixel | shot | snow | zoom | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Joint Train | ✓ | 62.3 | 4.5 | 26.7 | 39.9 | 25.7 | 30.0 | 5.8 | 16.3 | 5.8 | 45.3 | 30.9 | 45.9 | 7.1 | 25.1 | 31.8 | 24.8 |
| Fine-Tune | ✓ | 67.5 | 7.8 | 33.9 | 32.4 | 36.4 | 38.2 | 22.0 | 15.7 | 23.9 | 51.2 | 37.4 | 51.9 | 23.7 | 37.6 | 37.1 | 33.7 |
| ViT Probe | ✓ | 68.3 | 6.4 | 24.2 | 31.6 | 38.6 | 38.4 | 17.4 | 18.4 | 18.2 | 51.2 | 32.2 | 49.7 | 18.2 | 35.9 | 32.2 | 29.2 |
| TTT-MAE | ✓ | 69.1 | 9.8 | 34.4 | 50.7 | 44.7 | 50.7 | 30.5 | 36.9 | 32.4 | 63.0 | 41.9 | 63.0 | 33.0 | 42.8 | 45.9 | 44.4 |
| OpenCLIP VIT-L/14 | ✗ | 71.9 | 47.0 | 50.3 | 32.7 | 58.3 | 46.9 | 26.0 | 26.5 | 28.1 | 62.7 | 37.7 | 58.3 | 28.2 | 50.4 | 37.9 | 42.1 |
| APM (Ours) | ✗ | **77.4** | **51.9** | **56.6** | **37.9** | **64.8** | **53.2** | **28.7** | **31.4** | **33.0** | **68.4** | **44.1** | **64.5** | **33.1** | **56.9** | **43.9** | **50.3** |

| Method | P | Flower102 | DTD | Pets | UCF101 | Caltech101 | Food101 | SUN397 | Aircraft | EuroSAT | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CoOp | ✓ | 68.7 | 41.9 | 89.1 | 66.5 | 93.7 | 85.3 | 64.2 | 18.5 | 46.4 | 63.9 |
| CoCoOp | ✓ | 70.9 | 45.5 | 90.5 | 68.4 | 93.8 | 84.0 | 66.9 | 22.3 | 39.2 | 64.6 |
| CLIP-ViT-B/16 | ✗ | 67.4 | 44.3 | **88.3** | 65.1 | 93.4 | 83.7 | 62.6 | 23.7 | 42.0 | 63.6 |
| Ensemble | ✗ | 67.0 | 45.0 | 86.9 | 65.2 | 93.6 | 82.9 | 65.6 | 23.2 | 50.4 | 64.6 |
| TPT | ✗ | **69.0** | 47.8 | 87.8 | 68.0 | **94.2** | **84.7** | 65.5 | 24.8 | 42.4 | 65.1 |
| APM (Ours) | ✗ | 62.0 | **48.9** | 81.6 | **72.6** | 89.6 | 84.2 | **65.7** | **29.7** | **55.7** | **65.5** |

## Computational Analysis




| | $t$ | Params(M)↓ | $M_{meas}(GB)$ ↓ | $M_i(GB)$ ↓ | $GFlops_{meas}$ ↓ | $GFlops_i$ ↓ |
|---|---|---|---|---|---|---|
| Swin[57] | 1-20 | 87 | 1.5 | 1.4 | 353 | 308 |
| TPT[90] | 1-20 | 151.3 | 3.1 | 2.7 | 529 | 476 |
| CLIP VIT-B/16 | 1-20 | 149.2 | 2.3 | 1.8 | 462 | 410 |
| CLIP VIT-B/16(u)[90] | 1 | 149.2 | 1.8 | 1.8 | 20.5 | 20.5 |
| APM(s) | 1 | 174.2(s+u) | 2.7 (s+u) | 1.8(u) + 0.6(s) | 20.5(u) | 20.5(u) |
| APM(s) | 2 | 174.2(s+u) | 2.7 (s+u) | 1.8(u) + 0.6(s) | 10(s) | 10(s) |
| Peak Use | 1-20 | 174.2(s+u) | 2.7 (s+u) | 1.8(u) + 0.6(s) | **241.7 (s+u)** | **210.5 (s+u)** |