Trading robust representations for sample complexity through self-supervised visual experience

Andrea Tacchetti [1 2] Stephen Voinea [1] Georgios Evangelopoulos [1 3]

[1]The Center for Brains, Minds and Machines, MIT and The McGovern Institute for Brain Research at MIT [2]Currently with DeepMind [3]Currently with X, Alphabet

Robust image representations make recognition easy (and cheap)

- Learning in small sample regimes is a remarkable feature of human perception.
- **Low-sample complexity learning** is related (computationally and statistically) to robustness to transformations.
- **Transformation-robust visual representations** have been linked in neuroscience with modes of visual experience equivalent to weak- or self-supervision.
  - (e.g. spatial proximity or sequential presentation)
- We explore systems that rely on **representations of images** learned through **weak supervision** prior to downstream **supervised tasks** (face/image recognition, one-shot learning).
- Our image representation are **neural network embeddings** learned using **unlabeled image sets and video sequences**.

Main idea: Orbit sets for weak supervision

TOP ROW: Images of an orbit sequence from the ``Late Night'' video face dataset. **MIDDLE ROW:** random samples from distinct orbits. **BOTTOM ROW:** their detected canonical, frontal view.

Generic orbit associated to $x \in \mathcal{X}$ is given by the equivalence relation: $\mathcal{O}_x = \{x' \in \mathcal{X} | x \sim x'\} \subset \mathcal{X}, c : \mathcal{X} \to \mathcal{C}$ such that $x \sim x' \Leftrightarrow c(x) = c(x')$.

TL;DR (summary)

Humans can recognize someone in a crowd after seeing a single image; artificial perception systems on the other hand require thousands of examples to achieve satisfactory levels of accuracy. However, this comparison is not fair: *humans bring to bear years of perceptual experience* when learning new visual tasks.

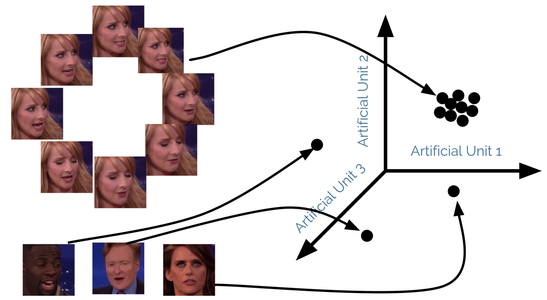
We explore the idea of having systems learn generic and robust representations from unconstrained visual experience and show that these make new, low-sample visual recognition easy.

Main contributions:

1. Building on prior theoretical work, we define **generic orbits**: sets of images that are generated by some unknown transformation acting on a canonical image (e.g. 3D face rotation).
2. We introduce a **new loss function** that combines a discriminative and a generative term and use it to learn representations from data-sets partitioned by *orbit sets* rather than class-label sets.
3. We empirically demonstrate the existence of a **trade off between weakly supervised and supervised learning**. That is, for a fixed recognition accuracy one can reduce the size of the supervised dataset by increasing the size of the representation dataset.

Results Learning representations using our loss function improves performance on downstream, low-sample tasks compared to other, state-of-the-art, weakly supervised methods.

Discriminate-and-Rectify loss



Triplets from sets of orbits:

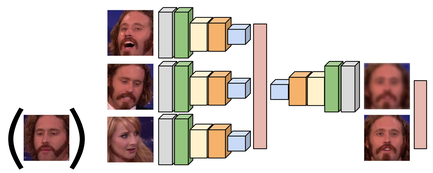
- *positive example* x_p (in-orbit), i.e. $x_i \sim x_p \Leftrightarrow x_i, x_p \in \mathcal{O}_{x_i}$
- *negative example* x_q (out-of-orbit), i.e. $\mathcal{O}_{x_q} \cap \mathcal{O}_{x_i} = \emptyset$.

$$\mathcal{T} \subset \{(x_i, x_p, x_q) \,|\, x_i \in \mathcal{X}_n, x_p \in \mathcal{O}_{x_i}, x_q \in \mathcal{O}_{x_q}; \mathcal{O}_{x_i} \cap \mathcal{O}_{x_p} = \emptyset\} \tag{1}$$

Representation learning (OJ):

$$\min_{\Phi,\tilde{\Phi}} \sum_{i=1}^{|\mathcal{T}|} \left(\frac{\lambda_1}{k} L_{\mathrm{t}}(x_i, x_p, x_q) + \frac{\lambda_2}{d} L_{\mathrm{e}}(x_i, x_c) \right), \tag{2}$$

1. **Discriminative term (OT)**: based on triplet loss, uses distances on the feature space.
$$L_{\mathrm{t}}(x_i, x_p, x_q) = \left| \left\| \Phi(x_i) - \Phi(x_p) \right\|^2_{\mathbb{R}^k} + \alpha - \left\| \Phi(x_i) - \Phi(x_q) \right\|^2_{\mathbb{R}^k} \right|_+, \tag{3}$$
2. **Reconstruction error (OE)**: distance on input space between decoder output and orbit canonical
$$L_{\mathrm{e}}(x_i, x_c) = \left\| x_c - \tilde{\Phi} \circ \Phi(x_i) \right\|^2_{\mathbb{R}^d}, x_c \in \mathcal{O}_{x_i}. \tag{4}$$

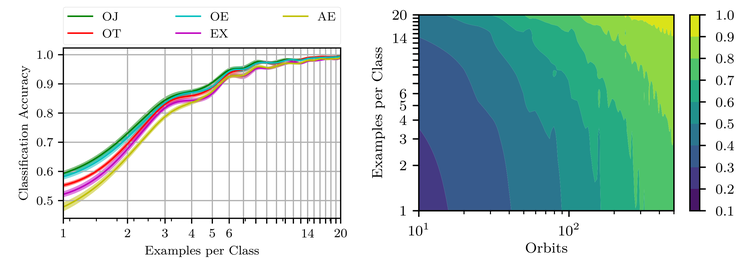


Experimental set-up (embedding, train, test)

1. Use a large embedding set, partitioned into *generic orbits* to learn Φ using loss (3) and baselines.
2. Note that instances of the **same class** can appear in **separate orbits**!
3. Use a very small supervised set to train a simple classifier on a recognition task.
4. Assess performance on large test set.
5. There is **no overlap** between the embedding, training and test set **at the level of orbits**!

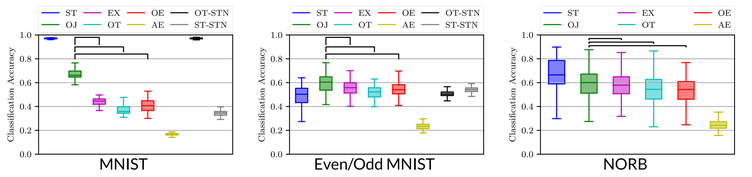
"Late Night" face dataset: Automatic orbit extraction

Trade-off between weak and exact, full supervision

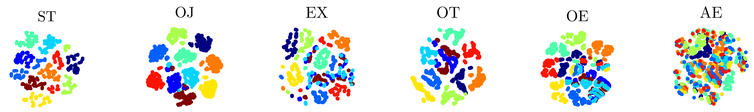


(LEFT PLOT) Sample complexity: Accuracy (across 10 re-samples) vs. training set size on the **Late Night** face dataset. Task is a 28-way face discrimination (linear SVM), with embeddings learned on a separate set (500 orbits).
(RIGHT PLOT) Embedding and sample complexity trade-off: Accuracy map (mean across 10 train/test re-splits of the validation set) of OJ for 1-20 labeled examples per class for classifier learning and 10-500 orbits for embedding learning.

One-shot learning on standard image benchmarks



Distances in embedding spaces.



2D t-SNE projections of face image embeddings in **Multi-PIE** (10 random classes, coded by different colors).

Acknowledgements

We thank **Tomaso Poggio** for the advice, inspiration and overall supervision throughout the project and the **McGovern Institute for Brain Research at MIT** for supporting this research. **This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.**