Learning Intermediate-Level Representations of Form and Motion from Natural Movies

Charles F. Cadieu[1], Bruno A. Olshausen[2]
[1]MIT, [2] UC-Berkeley
cadieu@mit.edu, baolshausen@berkeley.edu

We present a model of intermediate-level visual representation that is based on learning invariances from movies of the natural environment. The model is composed of two stages of processing: an early feature representation layer, and a second layer in which invariances are explicitly represented. Invariances are learned as the result of factoring apart the temporally stable and dynamic components embedded in the early feature representation. The structure contained in these components is made explicit in the activities of second-layer units that capture invariances in both form and motion. When trained on natural movies, the first-layer produces a factorization, or separation, of image content into a temporally persistent part representing local edge structure and a dynamic part representing local motion structure, consistent with known response properties in early visual cortex (area V1). This factorization *linearizes* statistical dependencies among the first-layer units, making them learnable by the second layer. The second-layer units are split into two populations according to the factorization in the first-layer. The form-selective units receive their input from the temporally persistent part (local edge structure) and after training result in a diverse set of higher-order shape features consisting of extended contours, multi-scale edges, textures, and texture boundaries. The motion-selective units receive their input from the dynamic part (local motion structure) and after training result in a representation of image translation over different spatial scales and directions, in addition to more complex deformations.

In addition to presenting the structure of the model and the characteristics of the learned representation, we show

- How the learned functions in the *second layer* may be condensed into a compact 'second-layer Gabor' function akin to the Gabor function commonly used to describe the structure learned by single layer sparse coding models,
- How experimental results in primate visual area V2 may be consistent with structure in the second layer, and
- That using the learned representation produces competitive results on classification benchmarks, such as STL-10 where the first and second layers produce an average test set accuracy of 59.5% (±0.3).

Topics: vision, biological learning, unsupervised learning
Preference: Oral
Presenting Author: Cadieu