Maximizing Image Diversity for fMRI studies

Charles Zheng

Stanford University

March 10, 2015

How should researchers choose images?

- *Relevance*. Using natural images, because we care about how the brain percieves typical scenes in the real world.
- *Control*. Using artificially generated images (like gratings) with known parameters
- *Diversity*. Using a broad set of images to convincingly demonstrate generalizability

In this work we only deal with the question of *diversity*

Questions about diversity

1. How can we define diversity?
2. How can researchers maximize this measure of diversity in their image set?
3. How does diversity affect metrics such as classification performance?
4. How can diversity improve generalizability?

To explore these questions, we use the dataset from *Kay et al.*

Section 2

Defining diversity

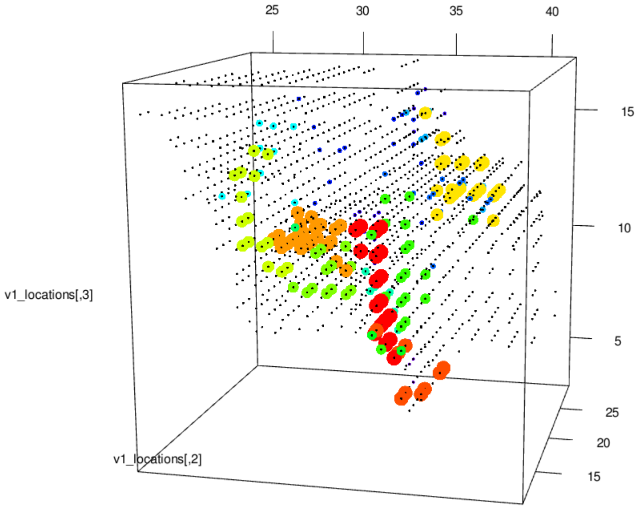
Defining diversity

- A first step in defining diversity is *dimensionality reduction* of images to a low-dimensional representation
- Next, we need a *distance metric* on the low-dimensional space
- With coordinates and a distance metric, we can define a measure of diversity

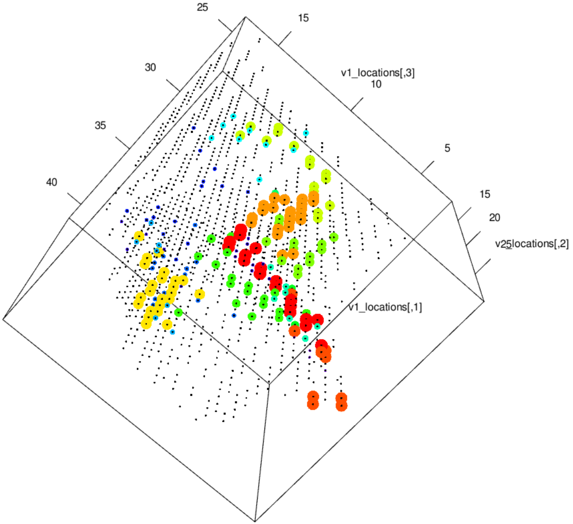
Dimensionality reduction

- Suppose we have "training data" for K images
- A $K \times p_f$ matrix X of image features (e.g. Gabor filters)
- A $K \times p_v$ matrix Y of mean voxel responses
- Use *sparse canonical correlation analysis* to find a $p_f \times d$ image basis U and a $p_v \times d$ voxel basis V
- Use the image basis U to define coordinates

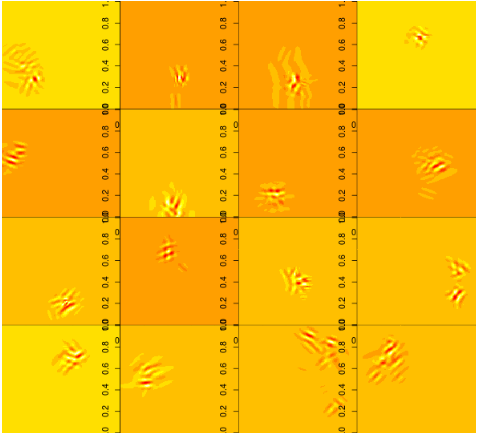
Data: V (voxel) basis



Data: V (voxel) basis



Data: U (image) basis



A measure of diversity

- We have coordinates z for any image, and a distance metric $d(z_1, z_2)$
- Suppose that the coordinates of all images form a connected, compact support set S
- We will define the diversity of a distribution P supported on S
- Let Z_1, Z_2 be random points drawn from P
- Let $\kappa(\cdot)$ be a bounded monotone decreasing function which goes to zero at infinity
- Define the diversity of P as

$$-\mathbb{E}[\kappa(d(Z_1, Z_2))]$$

A measure of diversity

- Let $\kappa(\cdot)$ be a bounded monotone decreasing function which goes to zero at infinity
- Define the diversity of P as

$$-\mathbb{E}[\kappa(d(Z_1, Z_2))]$$

- We did not define the function κ... but in some sense the choice is unimportant!
- Claim: let h be a bandwidth, and let

$$P_{\kappa,h} = \text{argmin}\mathbb{E}[k(d(Z_1, Z_2)/h)]$$

  Then let $P_\kappa = \lim_{h\to 0} P_{\kappa,h}$. There exist a unique distribution U such that for all κ satisfying a *postive definite* condition, $P_\kappa = U$. We call U the *uniform distribution*

Example

- Choose $\kappa(d) = \exp(-d^2)$, and various h
- Let $d_{i,j}$ be the pairwise distances. Measure the diversity of the image set by
$$-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\exp(-(d_{i,j}/h)^2)$$
- For $h = 1$, diversity of the validation set (120 images) is -0.008, diversity of the training set (1750 images) is -0.00057.
- For $h = 5$, diversity of the validation set (120 images) is -0.01, diversity of the training set (1750 images) is -0.003.
- The training set is more diverse

Section 3

Maximizing diversity

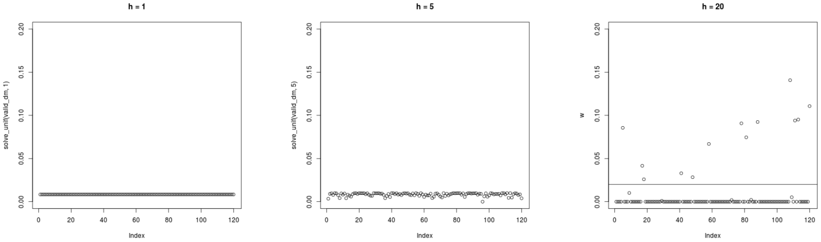
Choosing a diverse image set

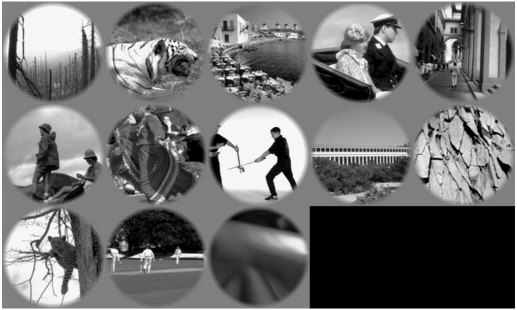
- Consider a database of $i = 1, \ldots, N$ images
- Any selection of images is a uniform distribution on a subset of $\{1, \ldots, N\}$ which can be represented by a vector w
- Given a kernel κ and bandwidth h, find the most diverse w by

$$\min_w w^T K w \text{ st } w \geq 0, ||w||_1 = 1$$

  where K has entries $k_{ij} = \kappa(d_{ij}/h)$.
- Finding the *optimal* weight vector w of a given sparsity is NP-hard. However by increasing the bandwidth h we naturally get sparse solutions.

Optimal w in 120 validation images

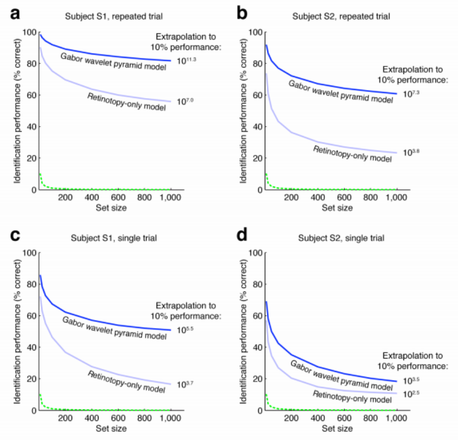




Section 4

Effect of diversity on misclassification

Misclassification rates from Kay et al



Setup for misclassification rates

- Fix a number n of images
- Pick n images at random *with replacement* from the database of N images
- (Sampling with replacement discourages deliberately using a small image database)
- Assess misclassification rate of images from fMRI data using whatever method (encoding/decoding method, or standard classification)

Distribution for the images?

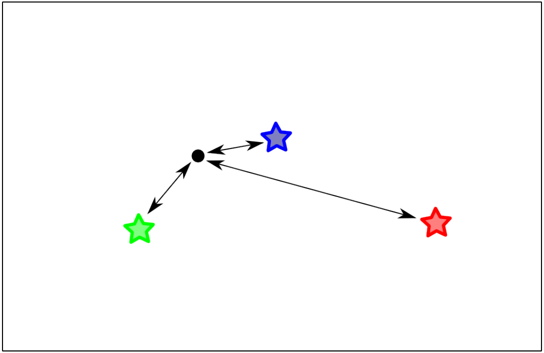
- Standard approach: pick n images uniformly from database
- What if your database has an unusually high proportion of a certain kind of image?
- Alternative proposal: pick n images with *probability weighted* by an arbitrary w
- Result: researchers would be incentivized to choose optimally uniform w.

Theoretical model

A theoretical model for misclassification with random images:

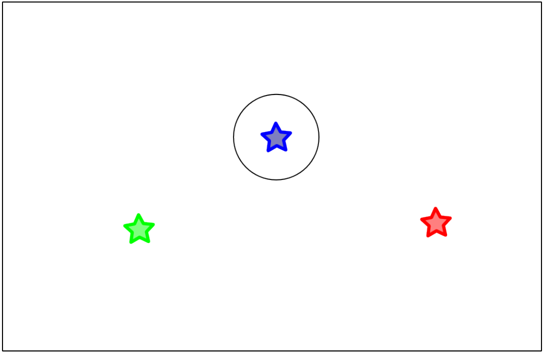
- Let P be a distribution with support S, with density $p(x)$
- Generate n random images: their mean fMRI responses are $\mu_1, \ldots, \mu_n \sim P$
- The observed fMRI responses for the ith image are $y_i^1, \ldots, y_i^{m_i} \sim N(\mu_i, \sigma^2 I)$

Theoretical model



Suppose further that the true μ_i are known. Then, points are classified based on nearest μ_i.

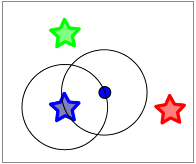
Theoretical model



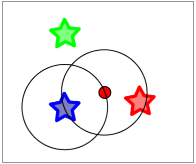
In a high dimension d, the distance of y_i^ℓ from μ_i approaches a constant, r_d due to *concentration of measure*.

Theoretical model

Correct!



Incorrect...



Whether a given point y_i^ℓ is classified correctly depends on whether or not there exist any other μ_j within radius r_d of y_i^ℓ.

Theoretical model

- An exact expression for average misclassification rate is

$$1 - \int_S \underbrace{p(\mu)}_{\text{distribution of } \mu} \left[\int_S \underbrace{\phi((\mu - y)/\sigma)}_{\text{distribution of } y} \underbrace{\left[1 - \int_{B_{||\mu - y||}} p(\tau) d\tau \right]^n}_{\text{pr. } y \text{ correctly classified}} dy \right] d\mu$$

- Under asymptotics $n \to \infty$, $\sigma = \sqrt[d]{K/n}$, approximated by

$$1 - \int_S C_1 \underbrace{p(\mu)}_{\text{distribution of } \mu} \underbrace{\exp[-C_2 p(\mu)]}_{\text{pr. } y \text{ correctly classified}}$$

Theoretical model

Conclusion: For optimal classification, the density $p(x)$ should maximize

$$\int_S C_1 p(\mu) \exp[-C_2 p(\mu)] d\mu$$

A simple argument shows that $p(\mu)$ should be constant (uniform).

Example in data

- Gaussian classification with shared covariance, validation data
- All data was used for covariance estimate
- 13 repeats per image, 11 in training and 2 for test
- Select 30 images randomly and get misc. error

Results in validation data

Averaged over 10000 trials

- Average misclassification (uniform)

$$0.1909 \pm 0.001$$

- Average misclassification (w from $h = 0.5$)

$$0.1935 \pm 0.001$$

- Average misclassification (w from $h = 0.2$)

$$0.1909 \pm 0.001$$

Conclusion: validation data is already close to uniform *with respect to itself*.
Further work: applying to test data, requires implementing encoding/decoding model rather than Gaussian classification.