Discussion of few-shot learning

Dr. Yuval Benjamini, Hebrew University

Israeli Statistical Association workshop

Dec 2022

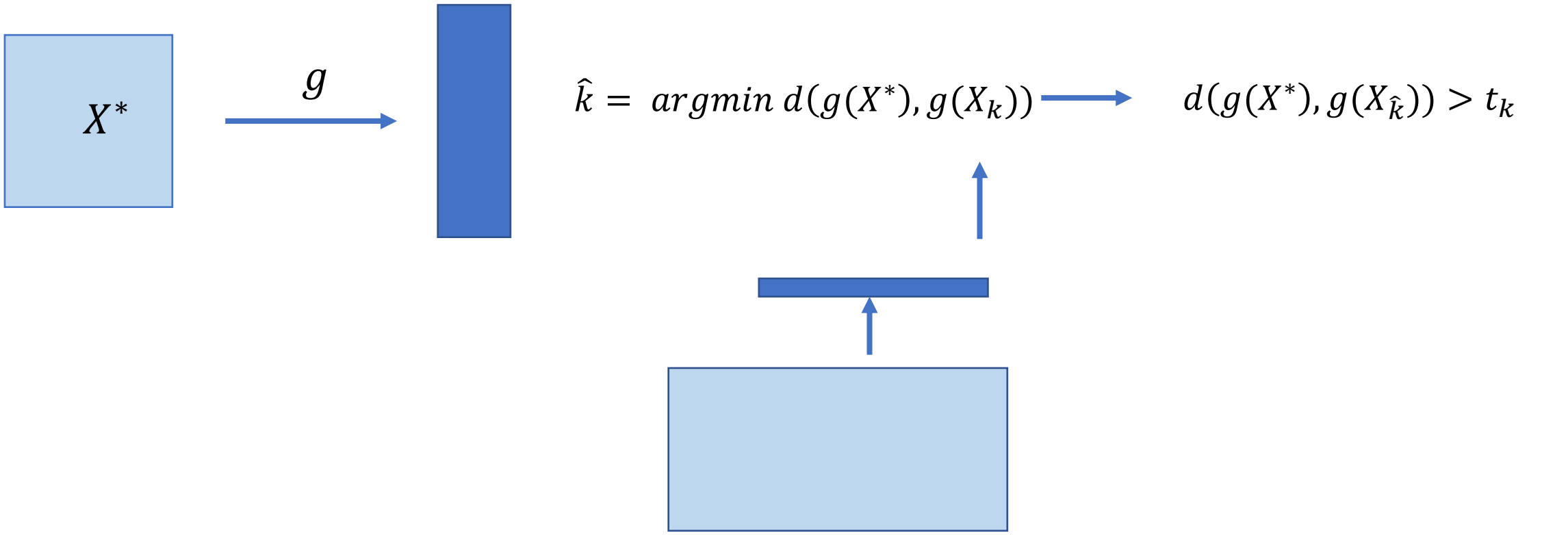
yuval.benjamini@mail.huji.ac.il
@yuvalbenj

The few-shot learning scheme

1. Represent

2. Find nearest neighbor

3. Decide if correct class



Discussions

1. The representation / metric tradeoff
2. How to decide we are in-class?

Interesting examples

A. (Annotated) Triplets for representation learning
B. Using few-shot as an approximate solution to inverse problems.

Discussion 1: The representation / metric tradeoff

- When FS introduced (early 2000s), representations mostly hand-tuned
- Overlap and correlation between features, so don't use Euclidean distance

- Metric learning (Xing 2002) looks for the best distance function, e.g.

> The goal of metric learning is to adapt some pairwise real-valued metric function, say the Mahalanobis distance $d_{\boldsymbol{M}}(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{x}')}$, to the problem of interest

- Must-link / cannot-link constraints (sometimes called positive / negative pairs):

$$\begin{aligned} \mathcal{S} &= \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be similar}\}, \\ \mathcal{D} &= \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be dissimilar}\}. \end{aligned}$$

- Relative constraints (sometimes called training triplets):

$$\mathcal{R} = \{(x_i, x_j, x_k) : x_i \text{ should be more similar to } x_j \text{ than to } x_k\}.$$

Survey of Metric Learning, Bellet et al 2013

The representation / metric tradeoff:

- But is learning a metric still necessary?

- Neural network variants greatly increase the ability to non-linearly transform data
- Often, in terms of performance, sufficient to presuppose Euclidean metric
- But if we have additional demands from representation *g(),* maybe metric is needed

- What is the optimal g() if we assume Euclidean metric ?
  Classical statistical theory provides some hints.

Survey of Metric Learning, Bellet et al 2013

The representation / metric tradeoff: A simple model

- Classes $Y \sim \tau$ with a sample $y_1, \ldots, y_k$
- Data samples $X \in R^d$ sampled $\{X|Y = y_j\} = f(y) + \epsilon \qquad E[\epsilon] = \mathbf{0}$

If we learn a few-shot representation $\hat{g}$ () with Euclidean distance,

should we expect that $\hat{g} = f$?

The representation / metric tradeoff: A simple model - LDA

- Classes $Y \sim \tau$ with a sample $y_1, \ldots, y_k$
- Data samples $X \in R^d$ sampled $\{X|Y = y_j\} = f(y) + \epsilon \qquad E[\epsilon] = \mathbf{0}$

Should we expect that $\hat{g} = f$?

In general the answer is no.

If $f(Y) \sim N_d(0, \Sigma_Y) \quad \epsilon \sim N_d(0, \Sigma_\epsilon)$ then optimal low dimensional separation is projecting f(Y) to leading eigen-vectors of $(\Sigma_\epsilon)^{-1}\,\Sigma_Y$

Directions were difference between classes dominates noise.

The representation / metric tradeoff

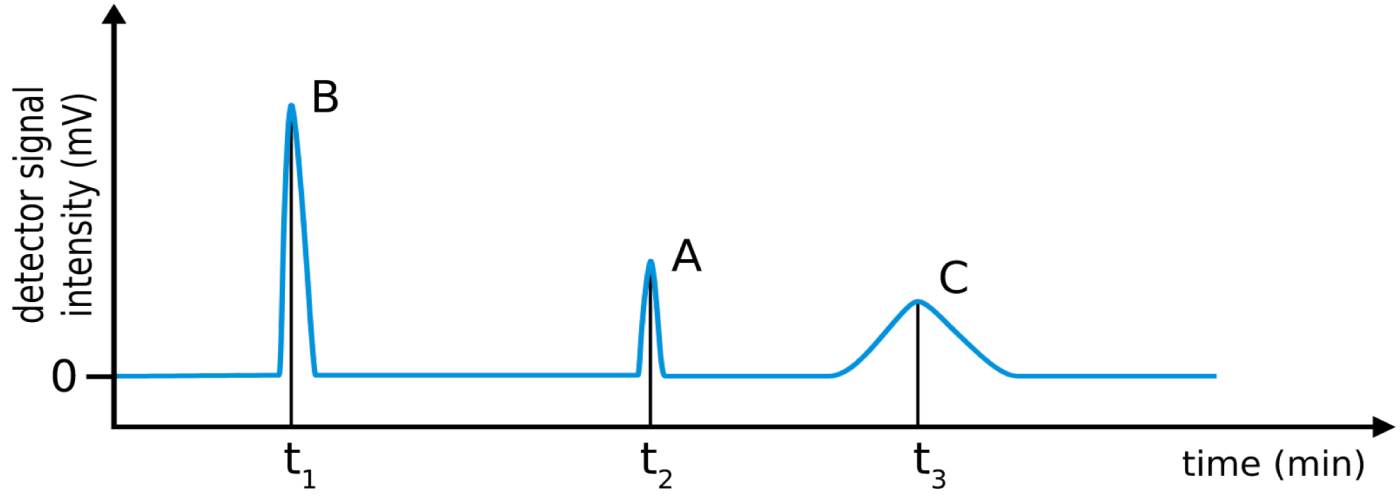
- When we want a meaningful representation,

  we might need to learn the metric as well…

- Examples:
  - MAZAP
  - Brain decoding (take 1)

Example 1: Detecting similar drug-compound mixtures

- Mass spectroscopy for a compound mixtures gives peak signals for each compound that are proportional to the quantity.
- When a new substance is caught, look for matches in previous samples



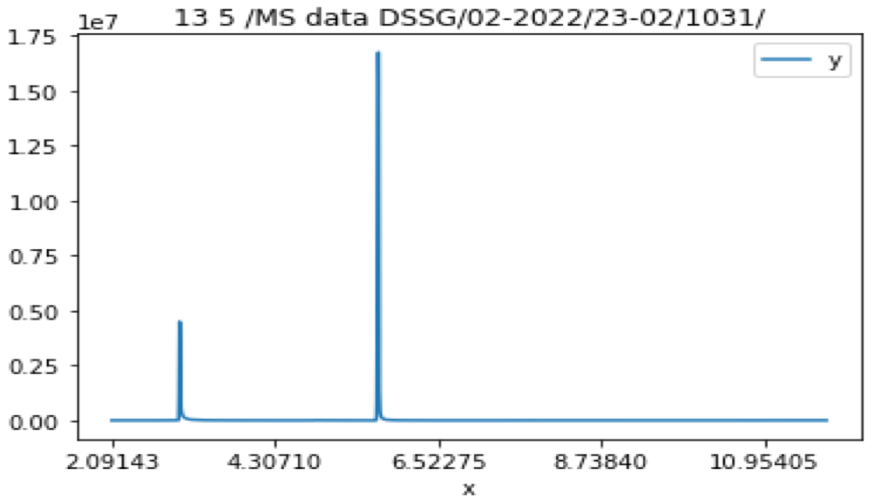
From Data Science for Social Good collaboration

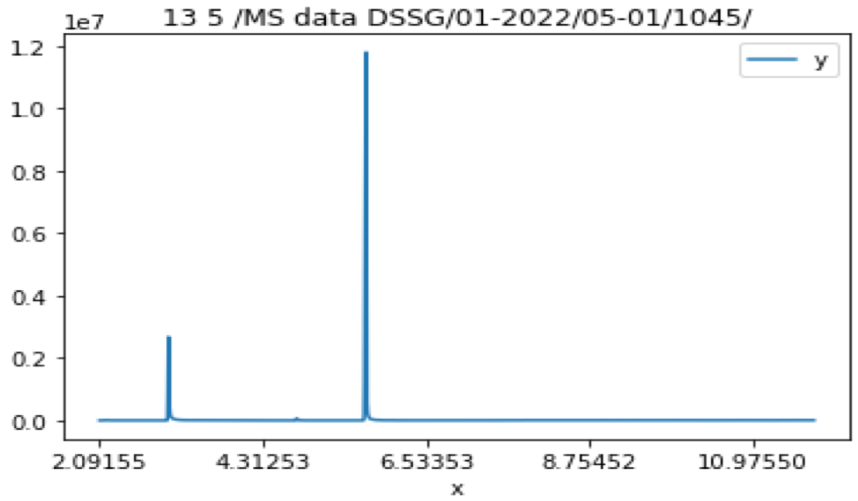
Example 1: Detecting similar drug-compound mixtures

- Perhaps *better* to stay in original space and estimate noise model:

$X \sim p_\mu(x)$ for a sparse μ vector

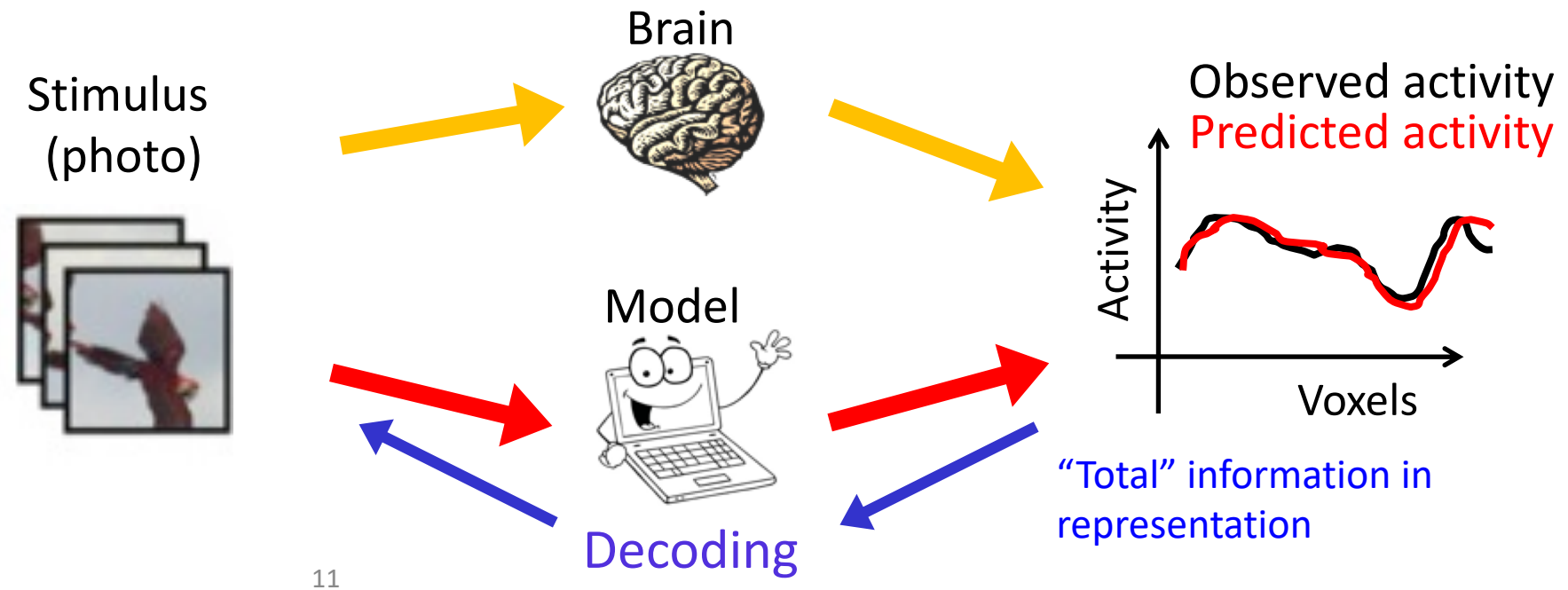
$$d(X^*, X_k) = argmin_{\mu \in M}(p_\mu(X^*) \cdot p_\mu(X_k))$$





Example 2: Identifying stimuli from brain response

- Encoding: Predict the neural activity evoked by a stimulus (photo).
- Decoding: Recovering stimulus from neural activity



Kay et al 2009

Example 2: Identifying stimuli from brain response

In this world:

X is a vector of brain measurements while viewing an image

Y_1,…,Y_1750 the image being shown (each image repeated 10 times)

Y*_1,…,Y*_120 new images as query sets

Can we identify the unseen image from the brain data?

CAN WE READ MINDS?

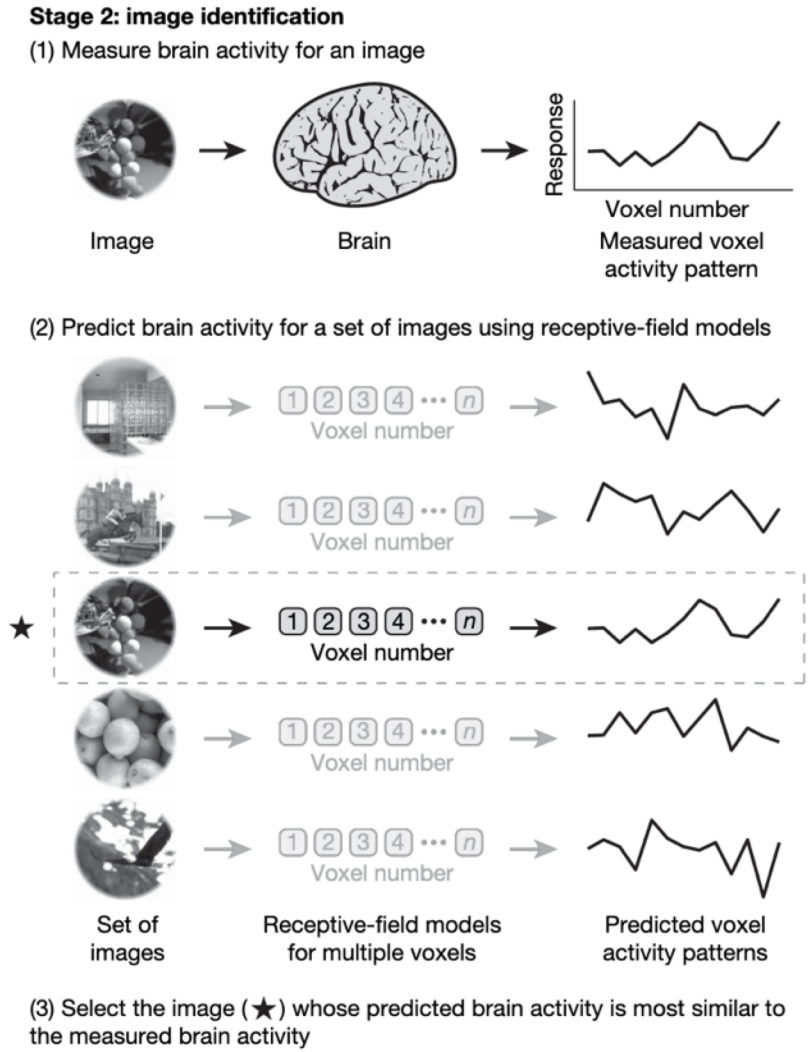
Example 2: Identifying stimuli from brain response

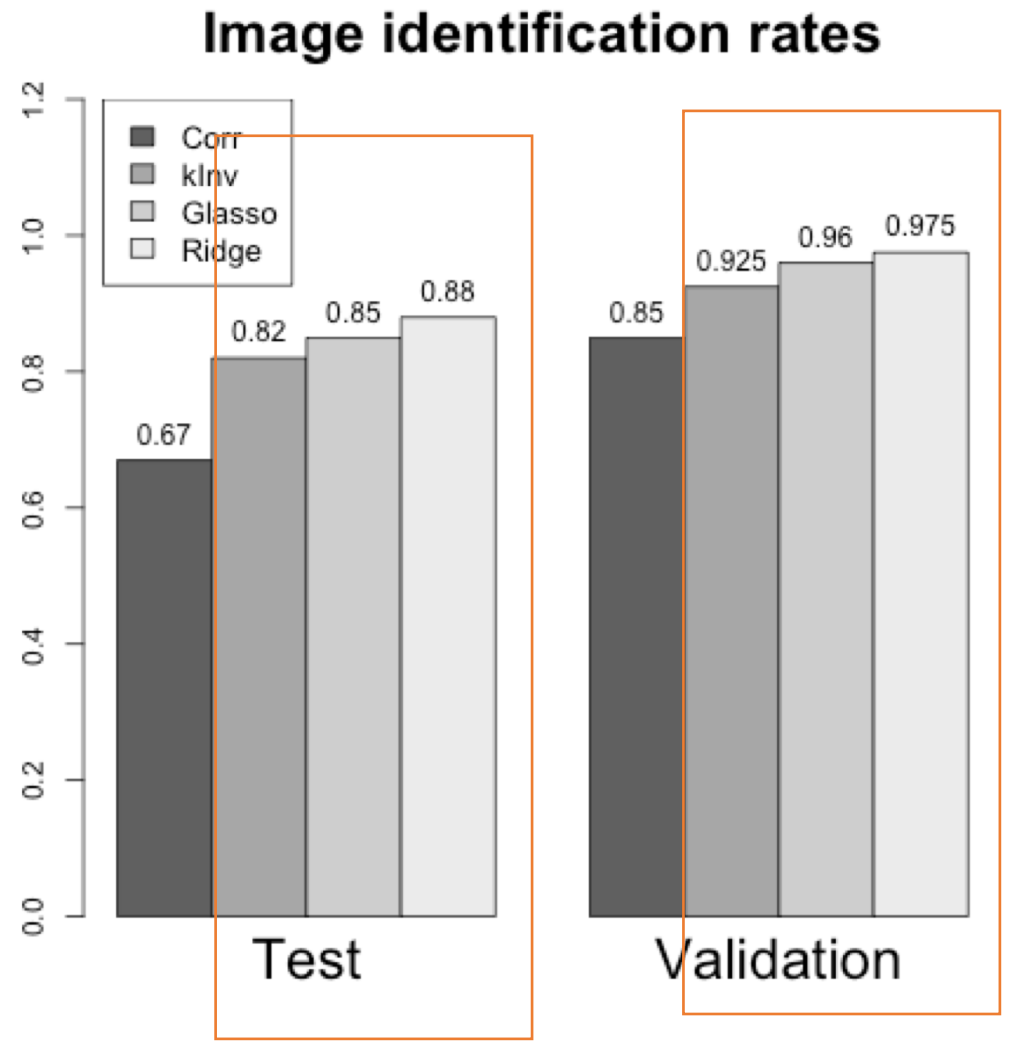
Their idea:

- Fit a multivariate regression model f: Y -> X using the training set.
- Estimate the covariance of the error S = Cov(f(Y)- X)
- For new brain signal X*:
  - each class y_1,…,y_k, give score

$$(y_j - f(X^*))' S^{-1} (y_j - f(X^*))$$

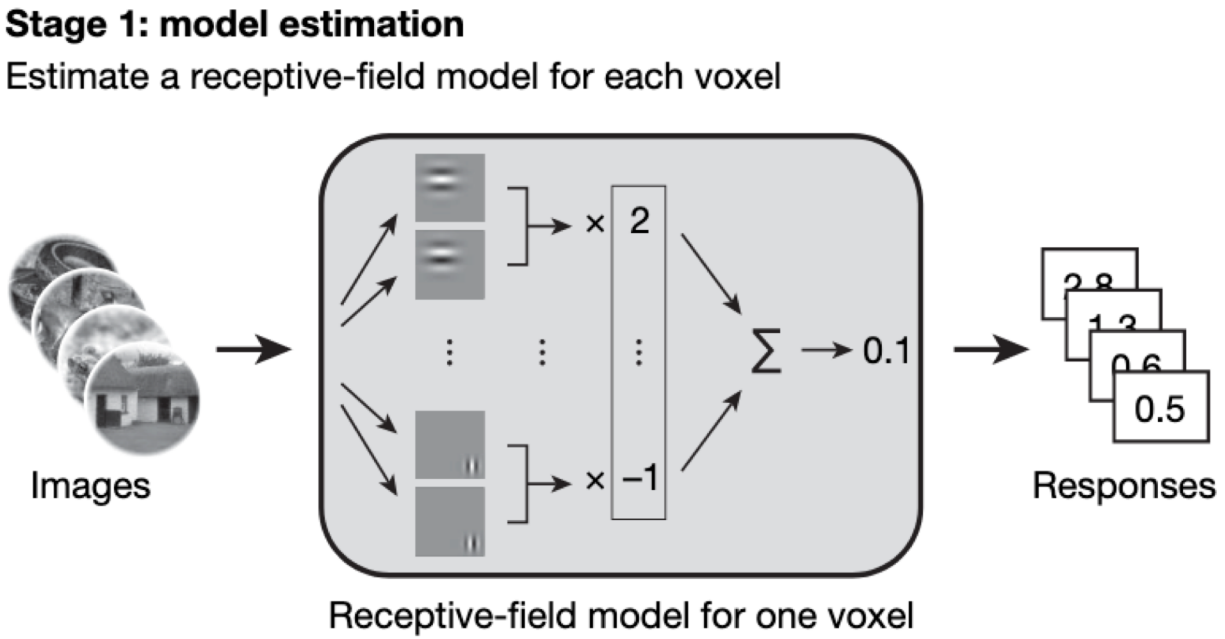
Example 2: Identifying stimuli from brain response





Example 2: Identifying stimuli from brain response

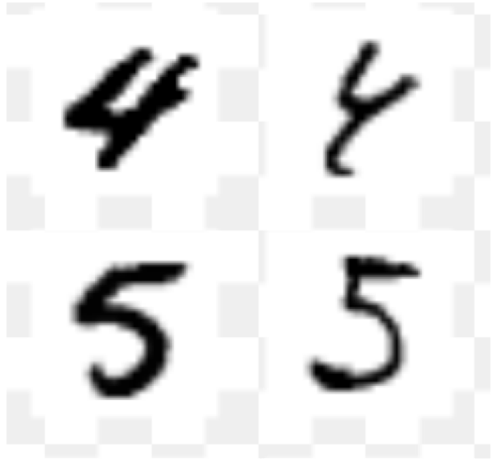
1. Show images to the subject and measure brain response
2. Code image into a vector of features.
3. Fit a function



Kay et al 2009

Example 3: Multi-task representation

- Often, we want to work in a representation that fits multiple tasks



MNIST
(hand-written digits)

Task 1: Identifying digit (OCR)

Task 2: Identifying writer

Discussion 2: The power of maybe

- “The largest benefit of using few shot learning is that we get confidence in our classification”
- How should we set the threshold for in-class?

Q: What do we call the problem of deciding whether x ~ P_y(x) was generated from a specific y’ ?

Proposal 1

- How should we set the threshold for in-class?

Proposal 1:

- In the binary description of the few-shot, we can use an ROC to threshold

Proposal 2

- How should we set the threshold for in-class?

Proposal 1:

- In the binary description of the few-shot, we can use an ROC to threshold

Proposal 2:

- For every class y, we can set a level beyond which we will not accept the classification.

Do these proposals give the same type of confidence?

Calibration on average vs. conditional calibration

- Proposal 1 gives calibration on average for pairs (x,y)

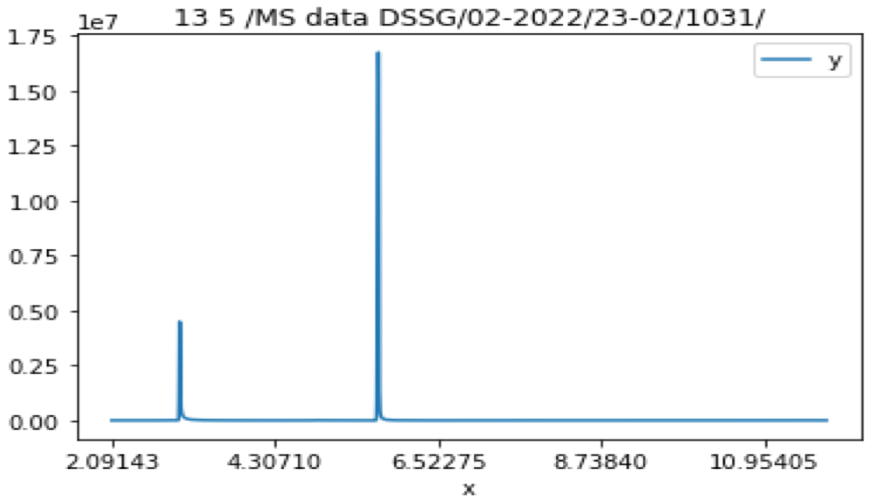
  But could be that for a given class, we are always wrong.

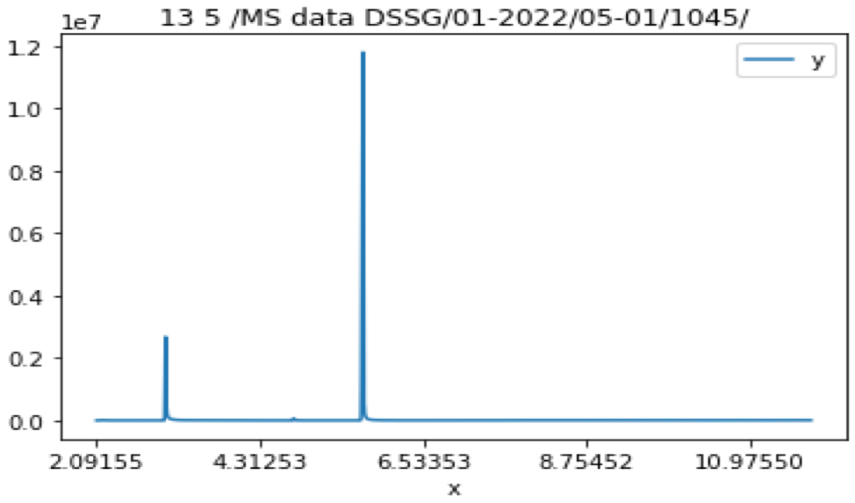
- Proposal 2 gives conditional calibration:
  For each class y, it ascertains that we get the correct calibration

- To check if Prop 1 gives conditional calibration,
  we should make sure to sample at least a few classes deeply.

Example 1: Detecting similar drug-compound mixtures

- In this space, perhaps easier to model the effect of μ on p_μ and get tailored probabilities (p-values).





Dense vs Sparse settings in few-shot

- We can make two kinds of error:
  - We need to decide between true class to nearest competitor class
  - We need to decide if this is true class, or a unseen class
- Dense problems: enough classes to sit inside the competitors noise
- Sparse problems: Classes are well separated, but need to decide whether good fit.

Dense vs Sparse settings in few-shot

- We can make two kinds of error:
  - We need to decide between true class to nearest competitor class
  - We need to decide if this is true class, or a unseen class

- Dense problems: Enough classes to sit inside the competitors noise
- Sparse problems: Classes are well separated, but need to decide whether good fit.

- Changes with number of classes + magnitude of noise.

Pirsomet to our research

- Yuli and I have a paper about the effect of the # of classes on accuracy, in the dense setting (☺)

Published as a conference paper at ICLR 2021

PREDICTING CLASSIFICATION ACCURACY WHEN ADDING NEW UNOBSERVED CLASSES

Yuli Slavutsky, Yuval Benjamini
Department of Statistics and Data Science
The Hebrew University of Jerusalem
Jerusalem, Israel
{yuli.slavutsky, yuval.benjamini}@mail.huji.ac.il

Interesting example A:
From FS learning to representation learning

- Often, we do not have true in-class / out-of-class.
- Contrastive/triplet loss and it’s variants can still be used
- This is really an extension of MDS (multi-dimensional-scaling)

Example: what is the space of objects?

- X: 1,854 objects in the THINGS database
- Look for x so that $d(y_i, \ y_j) \approx \ d(x_i, x_j)$
- Check what the dimensions of x represent in the real world

Revealing the multidimensional mental representations of natural objects, Martin 2020

Example: what is the space of objects?

- Annotate using triplets…

Broad sampling
of images
(1,854 objects)

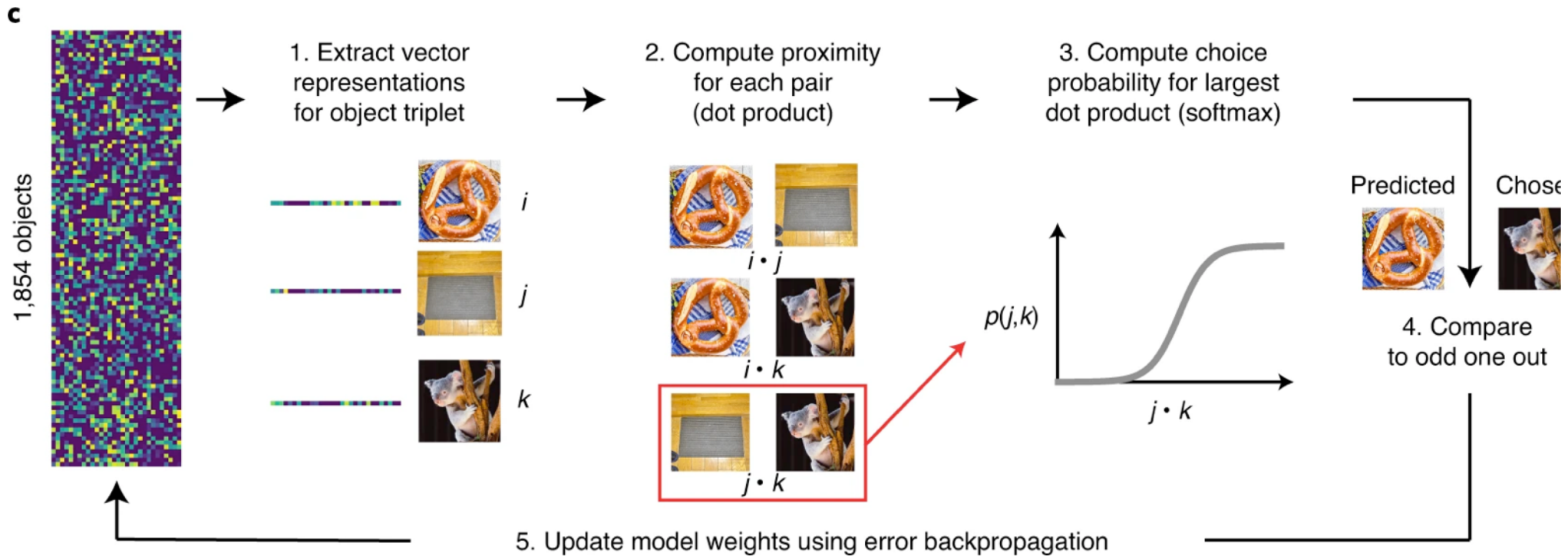
Large-scale
online crowdsourcing
(n = 1.46 million trials)

sentations of natural objects, Martin 2020

Example: what is the space of objects?

- Optimize x's with a probit model

$$\sum^{n} \log\left(\frac{\exp(\mathbf{x}_i\mathbf{x}_j)}{\exp(\mathbf{x}_i\mathbf{x}_j)+\exp(\mathbf{x}_i\mathbf{x}_k)+\exp(\mathbf{x}_j\mathbf{x}_k)}\right)+\lambda\sum^{m}||\mathbf{x}||_1$$



Revealing the multidimensional mental representations of natural objects, Martin 2020

Example: what is the space of objects?

Example: what is the space of objects?

Extended Data Fig. 2: Labels and word clouds for all 49 model dimensions.

From: Revealing the multidimensional mental representations of natural objects underlying human similarity judgements

Interesting Example B: Approximate inverse problem

- Suppose we know classes from high-dimensional space
- $y_1, \cdots, y_k$
- Data samples $X \in R^d$ sampled $\left\{X \middle| Y = y_j\right\} = f(y) + \epsilon \qquad E[\epsilon] = \mathbf{0}$

- Inverse problem: Given X find the y that generated it.

  De-blurring

  Standardizing faces

  Real mind reading (“Brain Decoding”)

Example: Decoding text for brain (Tang et al 2022)

Subjects listen to 16 h of podcasts in the FMRI.

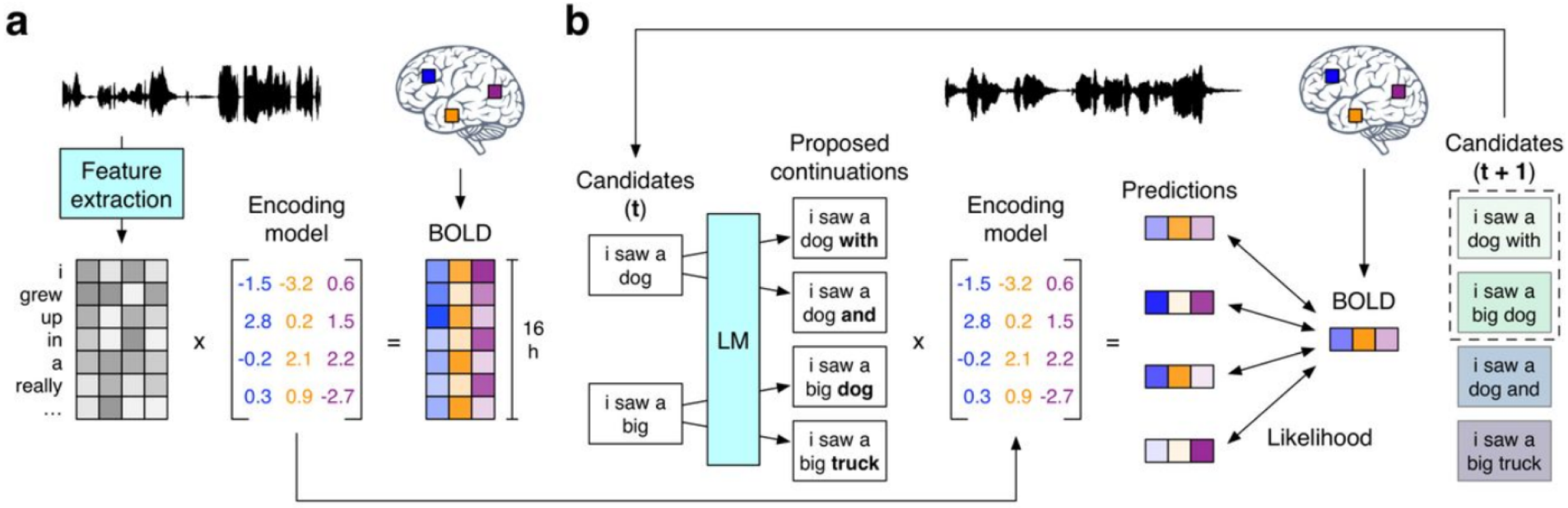
Texts are decoded from brain .

Example: Decoding text for brain (Tang et al 2022)

Code texts using language models.

Fit multivariate regression to brain space.

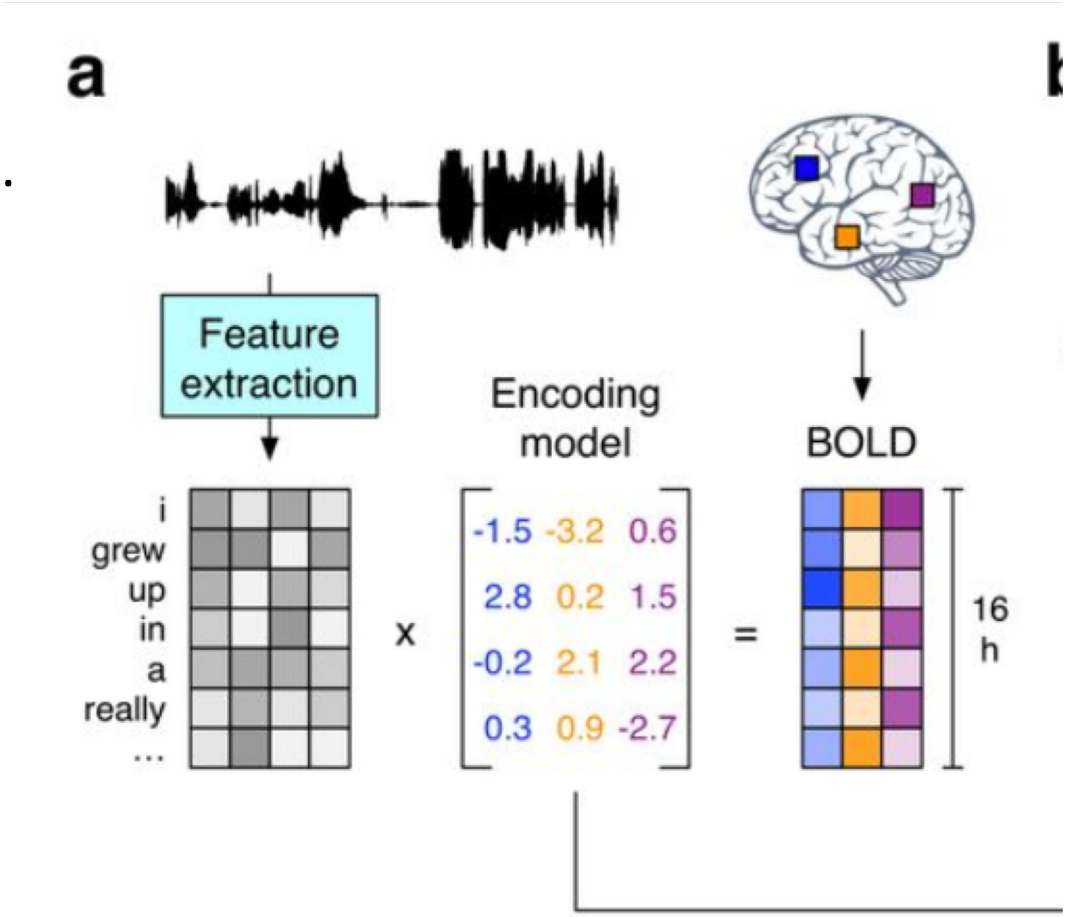
Learn covariance of error.



Example: Decoding text for brain (Tang et al 2022)

As before:

1. Code texts using language models.
2. Fit multivariate regression to brain space.
3. Learn covariance of error.

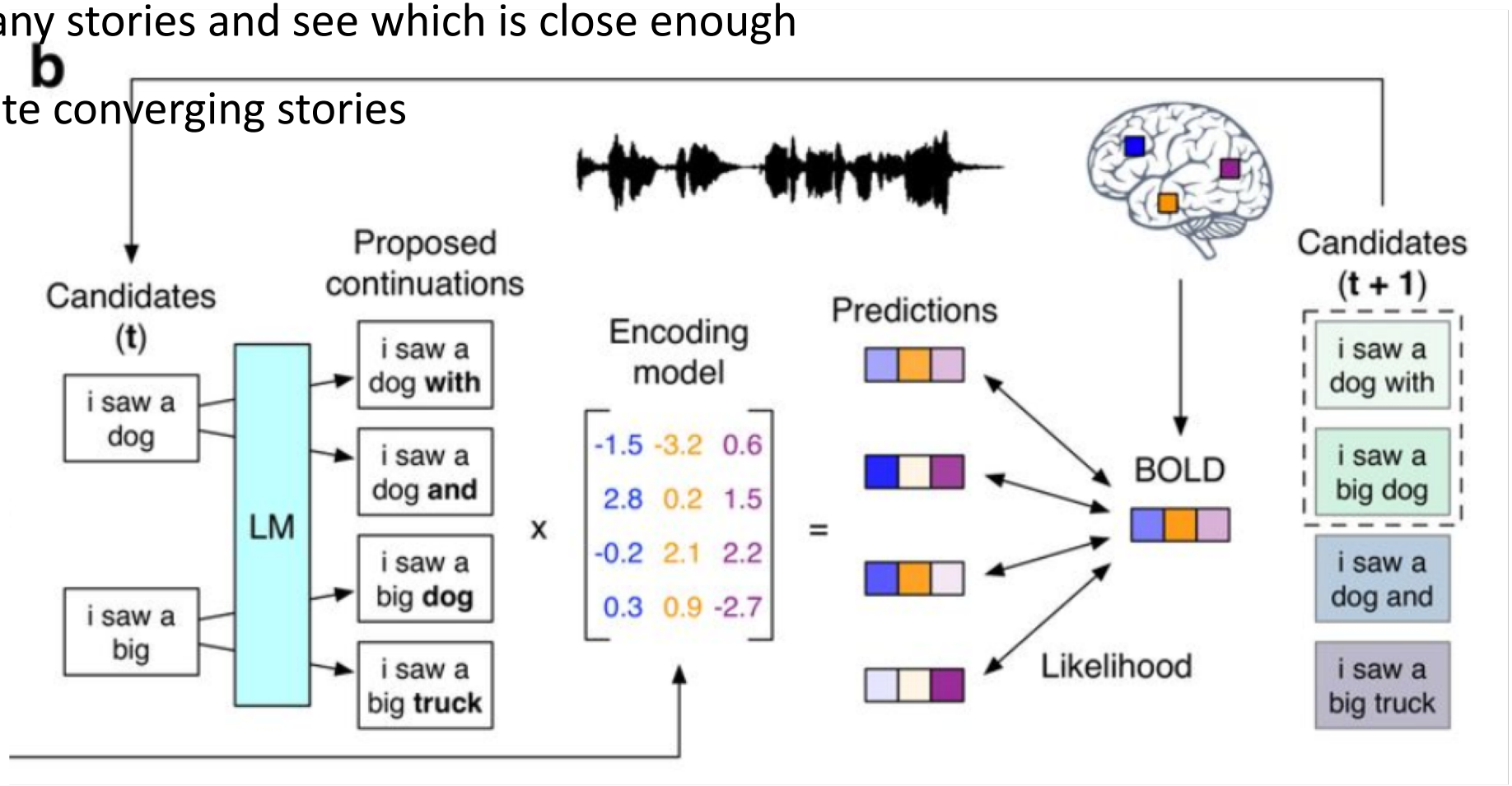


Example: Decoding text for brain (Tang et al 2022)

We know how to choose best story

So… we can try many stories and see which is close enough

Use GPT to generate converging stories



Summary

- The optimal transformation should depend on the noise distribution
- If representation is constrained, the metric should account for noise
- When deciding if class is correct,

  difference between average calibration and conditional callibration
- Contrastive / triplet loss can be used for representation learning,

  not only for classification.

References

Some early few shot papers

- Fe-Fei, L., 2003, October. A Bayesian approach to unsupervised one-shot learning of object categories. In *proceedings ninth IEEE international conference on computer vision* (pp. 1134-1141). IEEE.

- Ferencz, A., Learned-Miller, E.G. and Malik, J., 2005, October. Building a classification cascade for visual identification from one example. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 1, pp. 286-293). IEEE.

- Fink, M., 2004. Object classification from a single example utilizing class relevance metrics. *Advances in neural information processing systems*, *17*.

Metric learning:

- Bellet, A., Habrard, A. and Sebban, M., 2013. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.
- Xing, E., Jordan, M., Russell, S.J. and Ng, A ., 2002. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, *15*.
- Chopra, S., Hadsell, R. and LeCun, Y., 2005, June. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 539-546). IEEE.

Extrapolation:

- Slavutsky, Y. and Benjamini, Y., 2020, September. Predicting Classification Accuracy When Adding New Unobserved Classes. In *International Conference on Learning Representations*.

Examples:

- Kay, K., Naselaris, T., Prenger, R. *et al.* Identifying natural images from human brain activity. *Nature* **452**, 352–355 2008.

- Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B. and Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, *21*(19), pp.1641-1646.

- Zheng, Charles Y., Francisco Pereira, Chris I. Baker, and Martin N. Hebart. "Revealing interpretable object representations from human behavior." In *International Conference on Learning Representations*. 2018.

- Tang, J., LeBel, A., Jain, S. and Huth, A.G.Semantic reconstruction of continuous language from non-invasive brain recordings. *bioRxiv*. 2022.