Characterizing Semantic Information Content in Data

Aditya Chattopadhyay, Benjamin Haeffele, Donald Geman, René Vidal

MINDS Retreat – March 21st 2022
Why Semantic Information?

• Why do we want to quantify semantic information?
  – Such a measure could help in learning more interpretable representations of data.
  – Semantics could help characterize the complexity of a learning task, so as to compare one task with another.
  – Such a measure could help assess which data modalities are most important/relevant/informative for a task.

• Classical notions of information are insufficient: task-agnostic.

• Proposed approach: learn "semantic" representation for a task.
Prior Approach: MSI Representations

- **Prior Work**: representations are functions of the data that are
  - **Sufficient**: as informative as the data
  - **Invariant**: discount the effect of uninformative data transformations
  - **Minimal**: “simpler” than the data, ideally minimal
  - **Disentangled**?

- **Trade-off** minimal and sufficiency by minimizing information bottleneck [Tishby-Bialek-Pereira ’99]: but doesn’t generalize

\[
\min_{q(z|x)} \mathcal{L} = H_{p,q}(y|z) + \beta I(z;x)
\]
Proposed Approach: Semantic Representation

**Goal:** develop information theoretic framework for quantifying the semantic information content in multimodal data for a task

Data $\rightarrow$ Representation $\rightarrow$ Task $\rightarrow$ $\mathbf{X} \rightarrow \mathbf{Z} \rightarrow \mathbf{Y}$

*Approach:* learn minimal representation of the data that is sufficient for the task

- **Minimal, Sufficient, Invariant Representations:** functions of the data (features) that are informative for a class of tasks.

- **Latent Representations:** variables are latent, not necessarily interpretable in human language, i.e., not semantic.

- **Semantic Representations:** representations that depend on a semantic vocabulary that is relevant for the task.

An Interpretation of Shannon’s Entropy

- We denote input random variable as $X$ and the output random variable as $Y$. The joint $P_{X,Y}$ implicitly defines the learning task.

- If the task was transmission/compression, so that $Y = X$

Query set $Q$ is the set of all possible binary functions defined on $X$. $E$ refers to the coding strategy/encoder.

Each bit can be seen as an answer to a binary question on $X$.

$E[|\text{Code}^E_Q(X)|] \geq |\text{Entropy}(X)|$
From Entropy to Semantic Entropy

• Replace “bits” by (task-dependent) elementary atoms of semantic information – “semantic bits” (s-bits) from a user-defined query set \( Q \).

\[
\begin{aligned}
X & \quad \text{Encode} \quad 0 \quad 1 \quad 0 \quad \text{Decode} \quad Y \\
0 & \quad 1 & \quad 1 & \quad 0
\end{aligned}
\]

Each s-bit can be seen as an answer \( q(X) \) to a question \( q \in Q \).

\[\mathbb{E}[|\text{Code}_Q^E(X)|] \geq SE_Q(X; Y)\]

• Generalize Entropy to Semantic Entropy in \( X \) for \( Y \).

\[
SE_Q(X; Y) := \min_{\text{Coding Strategy } E} \mathbb{E}[|\text{Code}_Q^E(X)|]
\]

s.t. \( p(y | \text{Code}_Q^E(x)) = p(y | x) \ \forall x, y \)
Potential Query Sets

1. Q: Is there a person in the blue region? A: yes
2. Q: Is there a unique person in the blue region? A: yes
   (Label this person 1)
3. Q: Is person 1 carrying something? A: yes
5. Q: Is person 1 walking on a sidewalk? A: yes
6. Q: Is person 1 interacting with any other object? A: no

Visual Semantic Information

Salient parts of the image
**Defintion: Semantic Entropy**

\[ X \xrightarrow{\text{Encoder (E)}} p(Y \mid \text{Code}_Q^E(X)) \xrightarrow{\text{Output}} Y \]

**Queries**

- \( q_i \in Q \) from a user-specified query set

**Semantic Entropy**

- **Minimal:** \( \min_E \mathbb{E}_X [||\text{Code}_Q^E(X)||] =: SE_Q(X;Y) \)
- **Sufficient:** s.t. \( p(y \mid \text{Code}_Q^E(x)) = p(y \mid x) \ \forall x \in X, y \in Y \)

\[ \text{Code}_Q^E(x) := \{(q_1, q_1(X), \ldots, q_5, q_5(X))\} \]
Approximating Semantic Entropy using IP

- **Computing** $SE_Q(X; Y)$ is generally intractable.
- **Information Pursuit (IP):** greedy strategy where the encoder chooses queries sequentially in order of information gain.

\[ \text{Definition: IP Encoder} \]

Queries are chosen according to observed $x$.

- First query: $q_1 = \arg \max_{q \in Q} I(q(X); Y)$
- Next query: $q_{k+1} = \arg \max_{q \in Q} I(q(X); Y | q_{1:k}(x))$
- Termination: $q_{L+1} = q_{STOP}$ if $\max_{q \in Q} I(q(X); Y | q_{1:L}(x)) = 0$

$q_{1:k}(x)$ is the event that contains all realizations of $X$ that agree on the first $k$ query-answers for $x$.

- **Theorem:** If $Y$ is a discrete-valued function of $X$ and $Q$ is the set of all binary queries on $X$, $SE_Q^{IP}(X; Y) \leq SE(X; Y) + 1$. 
Computing Mutual Information is Intractable

- Selecting the first query requires computing $I(q(X); Y)$
  - Need a joint distribution of $q(X)$ and $Y$.
- Later queries require computing $I(q(X); Y \mid q_{1:k}(x))$
  - Need a joint distribution of $(q(X), Y)$ given History.
  - As histories get longer, we run out of samples that match History.
- The above two problems need to be solved $\forall q \in Q$, which scales linearly with the number of queries.
- What do we assume to make computation tractable?

Queries are Independent Given Nuisances

• **Assumption**: query answers are conditionally independent given target variable Y and “some” latent nuisance variable Z

\[ p(Q(X), Z, Y) = \prod_{q} p(q(X) | Z, Y) p(Z) p(Y) \]

\[ Q(X) = \{q(x) : q \in Q\} \]

• Reasonable assumption unless queries are causally related.

• **Examples**:
  - Z = pose and lighting conditions.
  - Z = phonemes in speech.

![Diagram](image-url)
Learn a Generative Model for IP

• We learn this joint distribution from data using a VAE.

\[ p(Q(X), Z, Y) = \prod_{q} p(q(X) \mid Z, Y)p(Z)p(Y) \]

• Assuming conditional independence makes estimating

\[ I(q(X); Y \mid q_{1:k}(x)) \] tractable using MCMC sampling.
Semantic Information Pursuit for Binary Image Classification

Aditya Chattopadhyay, Donald Geman, Benjamin Haeffele, René Vidal

Mathematical Institute for Data Science
Johns Hopkins University
IP for binary image classification

• Task is image classification.

• Queries $q_i$: “What are the image intensities at the $i^{th}$ patch?”

MNIST  |  KMNIST  |  Fashion MNIST  |  Caltech Silhouettes
Experiments: IP in action (Iteration 0)
Experiments: IP in action (Iteration 1)
Experiments: IP in action (Iteration 2)
Experiments: IP in action (Iteration 3)
Experiments: IP in action (Iteration 4)
Experiments: IP in action (Iteration 5)
Experiments: IP in action (Iteration 6)
Experiments: IP in action (Iteration 7)
Experiments: IP in action (Iteration 8)
Experiments: IP in action (Iteration 9)
Experiments: IP in action (Iteration 10)
Experiments: IP in action (Iteration 11)
Experiments: Binary Image Classification

- Semantic Entropy correlates with task complexity.

**Figure 2.** The results conform with intuition of more complex datasets having higher semantic entropy. For instance, Caltech Silhouettes, a dataset of binarized images of 101 classes from the Caltech dataset is obviously semantically more complex than handwritten digits in the MNIST dataset.
IP for Interpretable Decision-Making

Aditya Chattopadhyay, Stewart Slocum, Benjamin Haeffele, Donald Geman, René Vidal

Annual MURI Review,
December 1-2, 2021

Work supported by the Army Research Office (ARO) under W911NF-17-1-0304. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative (MURI).
Interpretability Crisis

Reality

MRI Scan

Black-Box

Patient has Alzheimer's disease with 98.6% probability

Desire

MRI Scan

Black-Box

"Since this region is abnormally dilated…"

Patient has Alzheimer's disease with 98.6% probability
Post-Hoc interpretability: The norm

• Current trend is to interpret black-box models post-hoc.

• **The Good:** No need to retrain model, accuracy maintained.

• **The Bad:**
  – Explanations generated are unreliable; not faithful to the model it tries to explain.\(^1\)
  – Salient parts of image might not be most informative to end-users.\(^2\)

---

Explainable by Design

• Need a new framework for learning that is “explainable by design”.

• "Explainable" entails a description in words, symbols or patterns of the reasoning leading to the decision.

• Useful explanations are often domain and task-dependent.

• One way to capture this is by specifying a query set Q.
  – Set of user-defined functions of data, each interpretable to the user.
Explainable by Design (cont.)

• Given Q, efficiently compose queries to explain predictions concisely.

Input image  

Composing explainable queries
1. Has shape perching-like?  Yes
2. Has bill shape all-purpose?  Yes
3. Has belly color yellow?  Yes
4. Has upperparts color yellow?  No
5. Has throat color yellow?  No
6. Has breast color black?  Yes
7. Has belly color olive?  Yes

Predicted bird species  Green Jay
CUB Birds Attribute Dataset

- 11,788 images, 200 bird species
- 312 binary attributes per image
- Difficult, fine-grained task, too hard for non-experts
- **Query set**: One binary query for each attribute.

Caltech-UCSD Bird Species Classification

- 11,788 images, 200 bird species
- 312 binary attributes per image
- Difficult, fine-grained task, too hard for non-experts
- **Query set**: One binary query for each attribute.
IP in action

Great Crested Flycatcher

$p(Y | s_k^{IP}(x_0))$

1. shape::perching-like?
2. bill_shape::all-purpose?
3. belly_color::yellow?
4. upperparts_color::yellow?
5. throat_color::yellow?
6. breast_color::black?
7. back_color::buff?
8. throat_color::grey?
9. bill_length::about_the_same_as_head?
HuffPost dataset

- Dataset: Huffington Post news headlines, $132K$ samples
- Task is to identify topics of newspaper articles from headlines
- Total of 10 topics (e.g. Entertainment, Politics, Food & Drink)
- Given a vocabulary of possible words in the headline, query $q_i$: “Is the $i^{th}$ word present in the headline?”
**Category:** Travel

**Headline:** Where Chefs, Bartenders and Sommeliers Eat and Drink in New York

**Short Description:** With over 25,000 restaurants and bars in New York City, it isn’t easy to navigate the dining landscape in the city that never sleeps. We asked the industry pros where they go. Here are restaurants and bars that chefs, bartenders and sommeliers recommend visiting.
Future Directions

- Explore more scalable algorithms.
Future Directions

- Explore more semantic tasks

Original Image

Segmentation Label

IP Unsupervised Segmentation
More Information,

Research supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345.

Vision Lab @ JHU
http://www.vision.jhu.edu

Center for Imaging Science @ JHU
http://www.cis.jhu.edu

Mathematical Institute for Data Science @ JHU
http://www.minds.jhu.edu

Thank You!