An Interpretable Multimodal Retrieval Tool

Brian Bartoldson, Brenda Ng, Barry Chen
Lawrence Livermore National Laboratory
Motivation

Help nonproliferation analysts retrieve important multimodal data from a sea of unlabeled open-source data using multimodal semantic feature spaces created by HPC-accelerated Deep Learning.
Multimodal Data
from Max Planck Institute for Informatics

Video Modality
Cooking Activities 2.0 [1]
- Over 15 hours of video (185 videos with average duration of 5 minutes)
- Videos differ in human subject and dish prepared

Text Modality
Textually Annotated Cooking Scenes [2]
- Over 50,000 human descriptions of cooking activities displayed in the Cooking Activities 2.0 dataset
- Multiple descriptions per video subclip
Tool Walkthrough

On the query tab, the user selects a video to encode into the multimodal feature space.
Tool Walkthrough

On the query tab, the user selects a video to encode into the multimodal feature space.

The video and one of its text descriptions is displayed.
Tool Walkthrough

On the query tab, the user selects a video to encode into the multimodal feature space.

The video and one of its text descriptions is displayed.

Click play to watch the video!
Tool Walkthrough

On the retrieval tab, the encoded video is presented as a linear combination of labeled basis vectors for multimodal feature space.
Tool Walkthrough

On the retrieval tab, the encoded video is presented as a linear combination of labeled basis vectors for multimodal feature space.

✓ The highest-weighted vector is labeled “the person threw the debris into the garbage”.

---

The highest-weighted vector is labeled “the person threw the debris into the garbage”.
Tool Walkthrough

On the retrieval tab, the encoded video is presented as a linear combination of labeled basis vectors for multimodal feature space.

✓ The highest-weighted vector is labeled “the person threw the debris into the garbage”.

The linear combination is decoded into the text space. The nearest neighbors of this decoding are retrieved.
The user can alter the multimodal encoding (by adjusting the basis-vector weights) to get new results!
Technical Approach

Why neural networks?

- Neural networks can learn to extract modality-independent semantic features.
Technical Approach

Why neural networks?

- Neural networks can learn to extract modality-independent semantic features.

Why interpretability?

- Makes the neural network’s logic more transparent to the user
- Allows the user to modify queries with an understanding of how retrieved results will change
Multimodal Autoencoder
Bidirectional Deep Neural Network [3]
TensorFlow Implementation

TensorBoard Visualization of our BiDNN
TensorFlow Implementation

Minimization of BiDNN’s Cost Function

TensorBoard Visualization of our BiDNN
Interpretability

✓ Use multimodal data to train bidirectional deep neural network.

‘the person turned the stove on’

ELMo Word Vectors [4]

ResNet 152 V2 CNN [5]
Interpretability

✓ Use multimodal data to train bidirectional deep neural network.

Create basis for multimodal feature space via PCA of trained network’s video encodings.

‘the person turned the stove on’

ELMo Word Vectors [4]
Given a video encoding, we can find how much of each PCA axis is present by solving for $x$:

$$Ax = b$$

where the columns of $A$ are the principal axes, and $b$ is the video encoding.
Interpretability

Each axis is decoded into text space. The axis’s interpretable label is the nearest text-description neighbor to its decoding.
Thus, users can adjust their queries’ interpretable encodings to obtain a predictable effect on retrieved results.
Summary

- Observations of natural phenomena often possess multiple modalities.

- We seek to map multimodal data to a latent feature space that semantically characterizes data of any modality.

- We use multimodal neural networks to learn this feature space.

- For each data instance (text, image, or video), this model yields a (modality-agnostic) representation as a vector of latent variables.

- By applying PCA to the representations, and decoding the PCA axes, we obtain an interpretable basis that can be used to illuminate the neural network’s logic and subsequently fine-tune retrievals.
References


