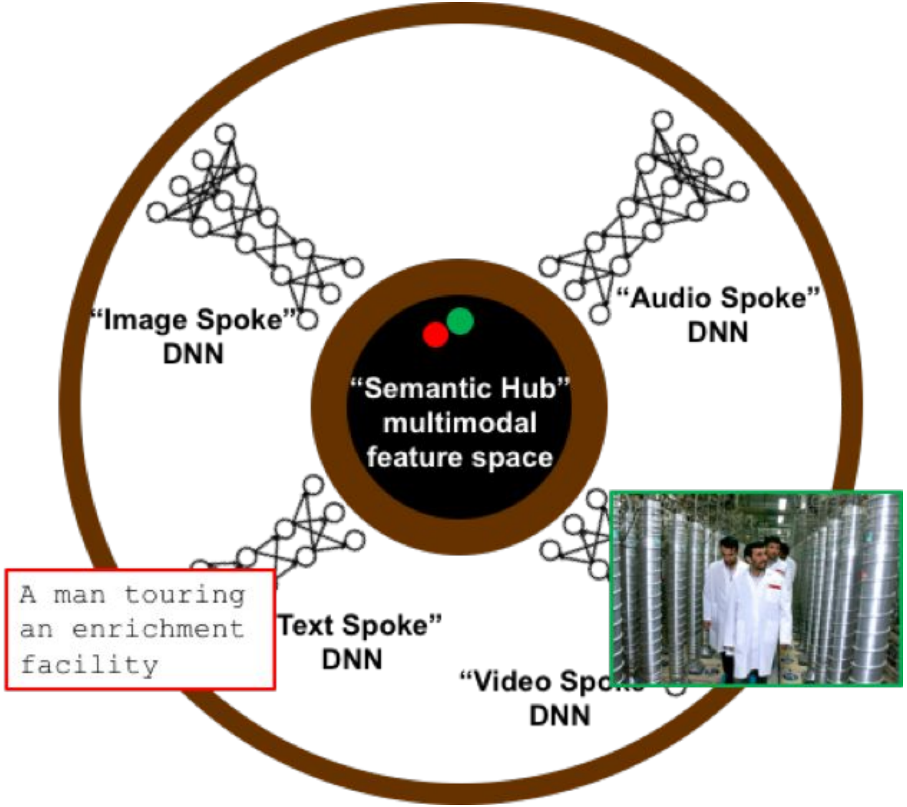
An Interpretable Multimodal Retrieval Tool

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.



"Image Spoke" DNN
"Audio Spoke" DNN
"Semantic Hub" multimodal feature space
Text Spoke" DNN
"Video Spo
DNN
A man touring an enrichment facility

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]

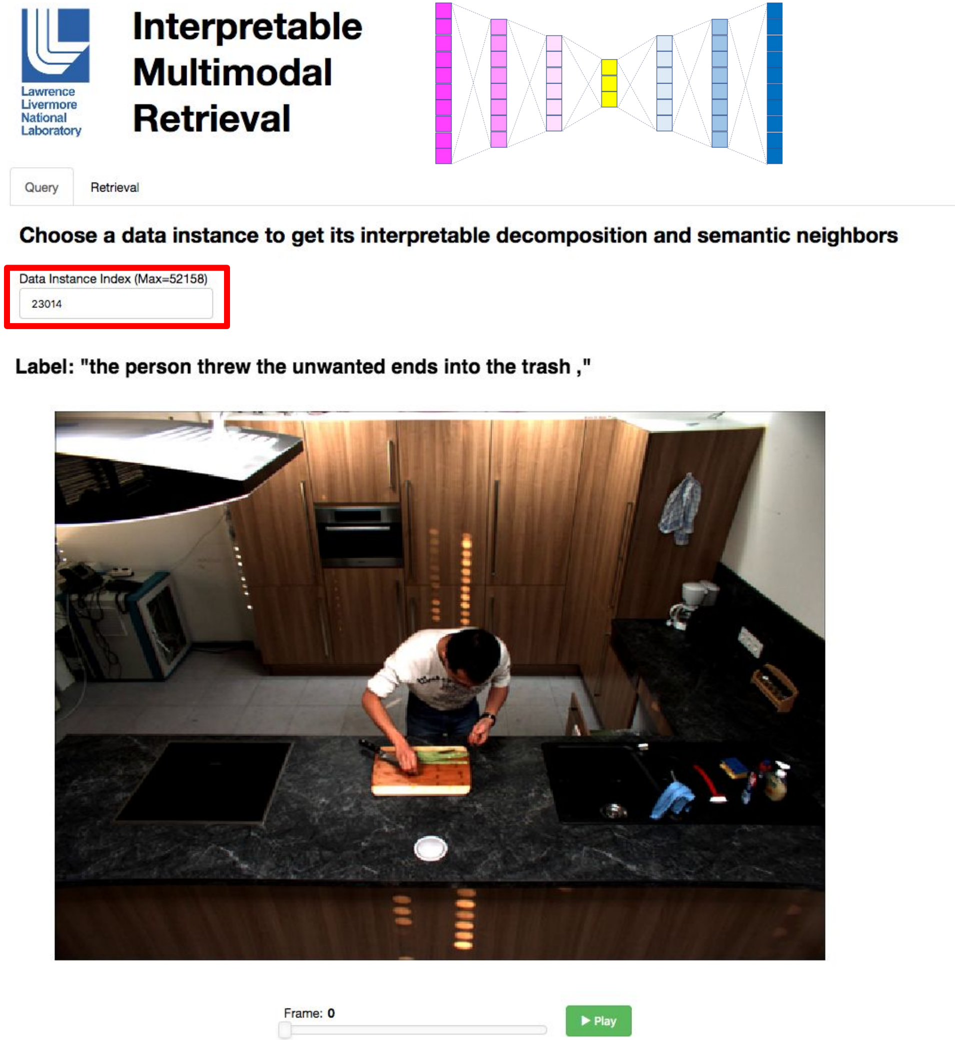
- Over 50,000 human descriptions of cooking activities displayed in the Cooking Activities 2.0 dataset
- Multiple descriptions per video subclip

A Frame from Video s21-d42

'the person adjusted burner temperature'
'the person turned the stove on'

Tool Walkthrough

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



Lawrence Livermore National Laboratory

Interpretable Multimodal Retrieval

Choose a data instance to get its interpretable decomposition and semantic neighbors

Data Instance Index (Max=52158)

23014

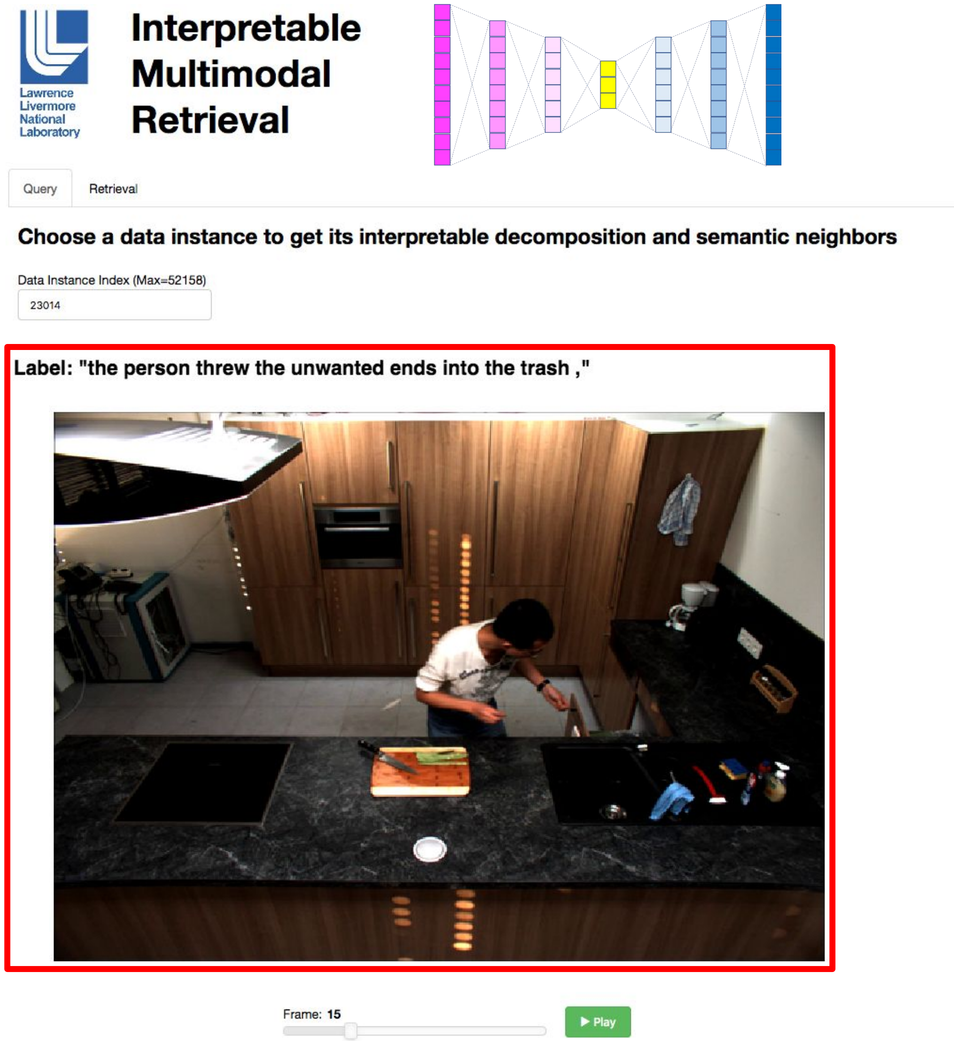
Label: "the person threw the unwanted ends into the trash ,"

Frame: 0

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.



Interpretable Multimodal Retrieval

Choose a data instance to get its interpretable decomposition and semantic neighbors

Data Instance Index (Max=52158): 23014

Label: "the person threw the unwanted ends into the trash ,"

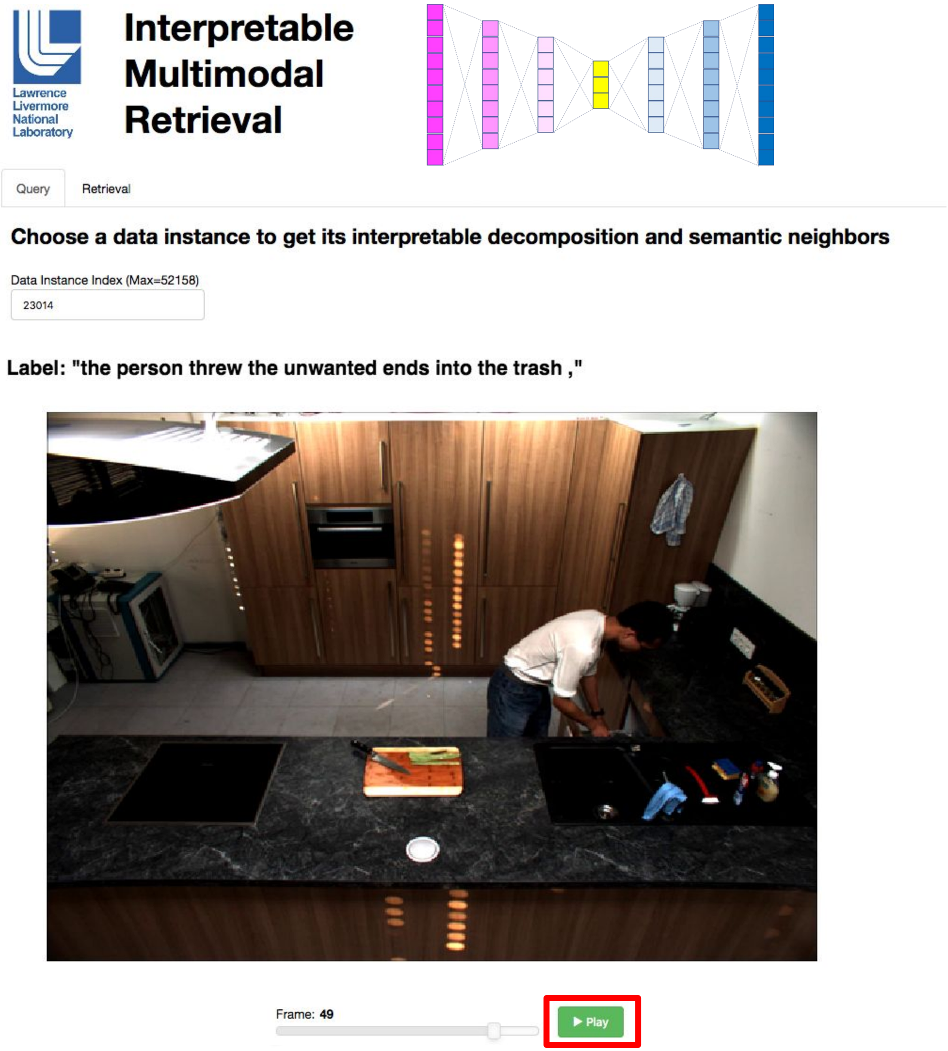
Frame: **15**

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!



Interpretable Multimodal Retrieval

Choose a data instance to get its interpretable decomposition and semantic neighbors

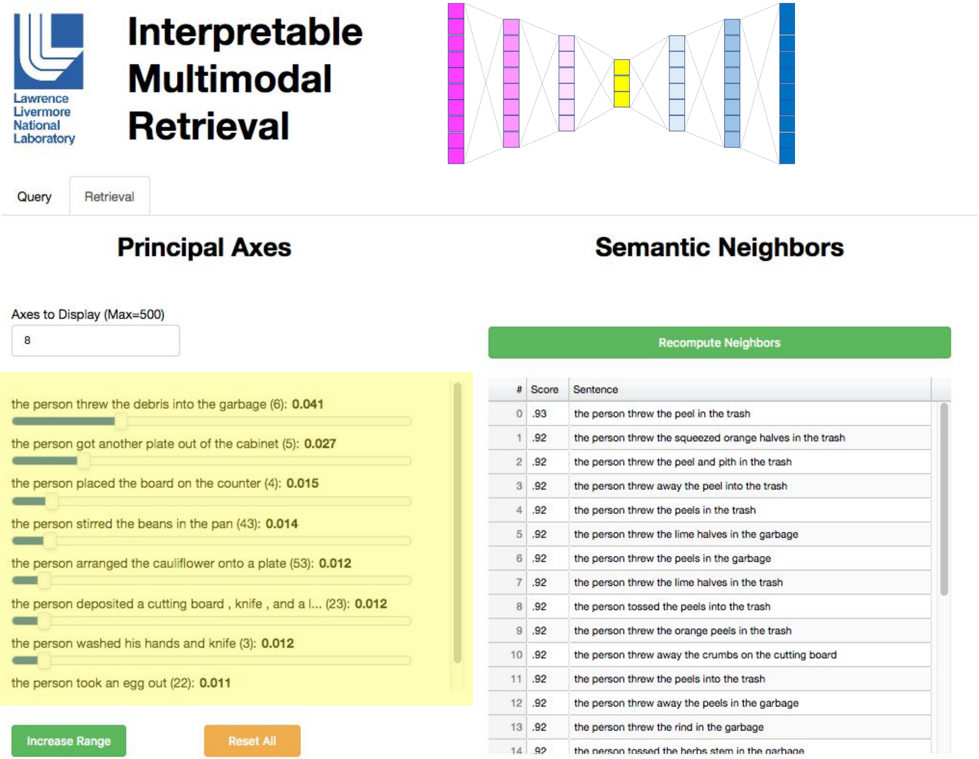
Data Instance Index (Max=52158): 23014

Label: "the person threw the unwanted ends into the trash ,"

Frame: 49

Tool Walkthrough

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



Lawrence Livermore National Laboratory
Interpretable Multimodal Retrieval
Query
Retrieval

Principal Axes

Axes to Display (Max=500): 8

the person threw the debris into the garbage (6): **0.041**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.014**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

the person washed his hands and knife (3): **0.012**

the person took an egg out (22): **0.011**

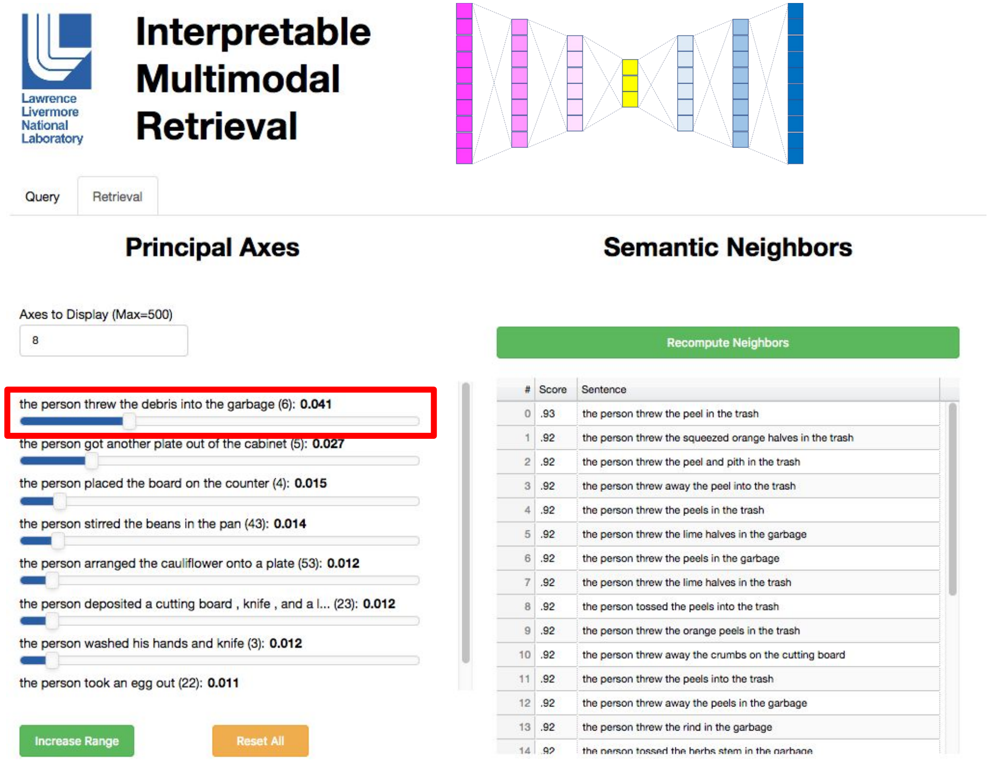
Semantic Neighbors

#	Score	Sentence
0	.93	the person threw the peel in the trash
1	.92	the person threw the squeezed orange halves in the trash
2	.92	the person threw the peel and pith in the trash
3	.92	the person threw away the peel into the trash
4	.92	the person threw the peels in the trash
5	.92	the person threw the lime halves in the garbage
6	.92	the person threw the peels in the garbage
7	.92	the person threw the lime halves in the trash
8	.92	the person tossed the peels into the trash
9	.92	the person threw the orange peels in the trash
10	.92	the person threw away the crumbs on the cutting board
11	.92	the person threw the peels into the trash
12	.92	the person threw away the peels in the garbage
13	.92	the person threw the rind in the garbage
14	.92	the person tossed the herbs stem in 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”.



Interpretable Multimodal Retrieval

Principal Axes

Axes to Display (Max=500): 8

the person threw the debris into the garbage (6): **0.041**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.014**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

the person washed his hands and knife (3): **0.012**

the person took an egg out (22): **0.011**

Semantic Neighbors

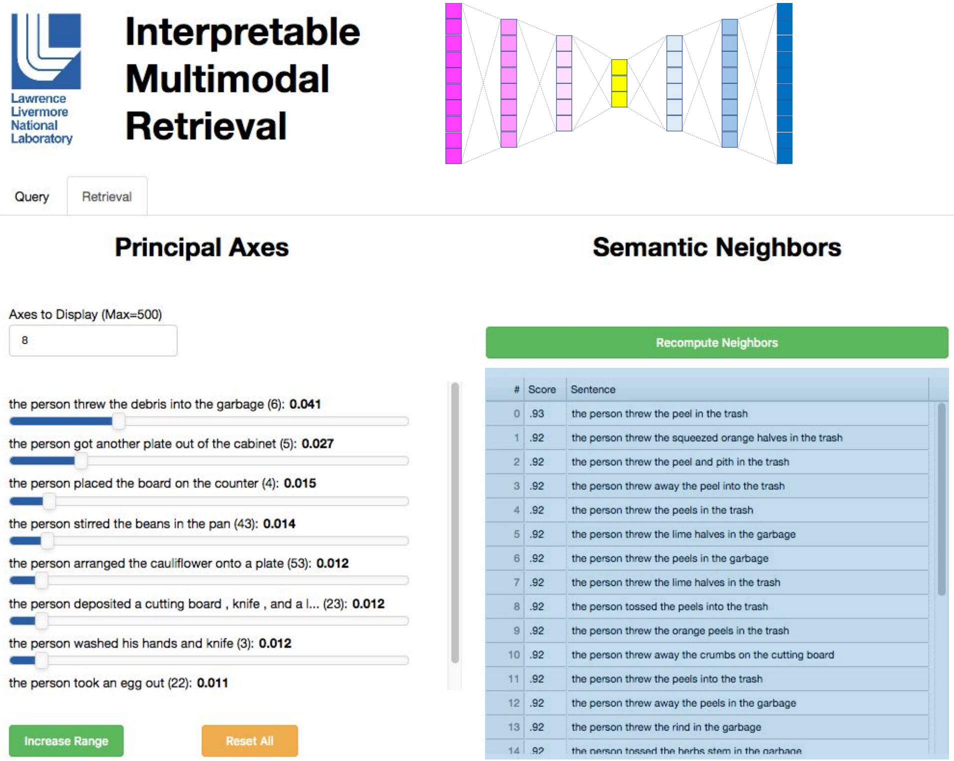
#	Score	Sentence
0	.93	the person threw the peel in the trash
1	.92	the person threw the squeezed orange halves in the trash
2	.92	the person threw the peel and pith in the trash
3	.92	the person threw away the peel into the trash
4	.92	the person threw the peels in the trash
5	.92	the person threw the lime halves in the garbage
6	.92	the person threw the peels in the garbage
7	.92	the person threw the lime halves in the trash
8	.92	the person tossed the peels into the trash
9	.92	the person threw the orange peels in the trash
10	.92	the person threw away the crumbs on the cutting board
11	.92	the person threw the peels into the trash
12	.92	the person threw away the peels in the garbage
13	.92	the person threw the rind in the garbage
14	.92	the person tossed the herbs stem in 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.


Lawrence Livermore National Laboratory
Interpretable Multimodal Retrieval
Query
Retrieval
Principal Axes
Semantic Neighbors
Axes to Display (Max=500)
8
Increase Range
Reset All
Recompute Neighbors

the person threw the debris into the garbage (6): **0.041**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.014**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

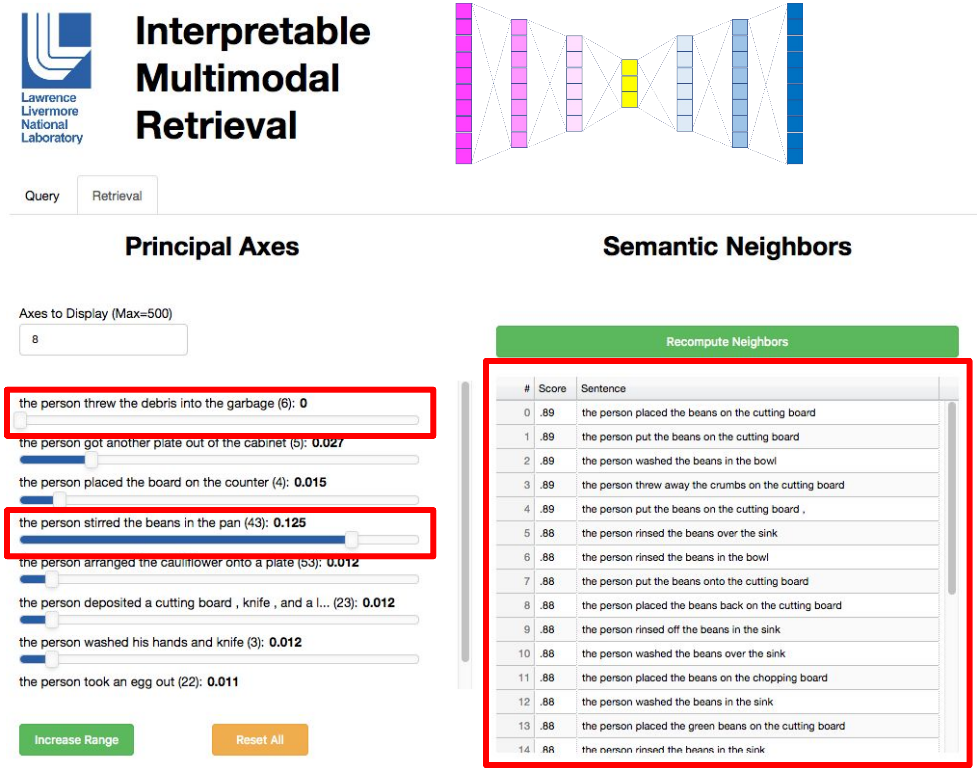
the person washed his hands and knife (3): **0.012**

the person took an egg out (22): **0.011**

#	Score	Sentence
0	.93	the person threw the peel in the trash
1	.92	the person threw the squeezed orange halves in the trash
2	.92	the person threw the peel and pith in the trash
3	.92	the person threw away the peel into the trash
4	.92	the person threw the peels in the trash
5	.92	the person threw the lime halves in the garbage
6	.92	the person threw the peels in the garbage
7	.92	the person threw the lime halves in the trash
8	.92	the person tossed the peels into the trash
9	.92	the person threw the orange peels in the trash
10	.92	the person threw away the crumbs on the cutting board
11	.92	the person threw the peels into the trash
12	.92	the person threw away the peels in the garbage
13	.92	the person threw the rind in the garbage
14	.92	the person tossed the herbs stem in the garbage

Tool Walkthrough

The user can alter the multimodal encoding (by adjusting the basis-vector weights) to get new results!



Interpretable Multimodal Retrieval

Principal Axes

Axes to Display (Max=500)

8

the person threw the debris into the garbage (6): **0**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.125**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

the person washed his hands and knife (3): **0.012**

the person took an egg out (22): **0.011**

Semantic Neighbors

#	Score	Sentence
0	.89	the person placed the beans on the cutting board
1	.89	the person put the beans on the cutting board
2	.89	the person washed the beans in the bowl
3	.89	the person threw away the crumbs on the cutting board
4	.89	the person put the beans on the cutting board ,
5	.88	the person rinsed the beans over the sink
6	.88	the person rinsed the beans in the bowl
7	.88	the person put the beans onto the cutting board
8	.88	the person placed the beans back on the cutting board
9	.88	the person rinsed off the beans in the sink
10	.88	the person washed the beans over the sink
11	.88	the person placed the beans on the chopping board
12	.88	the person washed the beans in the sink
13	.88	the person placed the green beans on the cutting board
14	.88	the person rinsed the beans in the sink

Technical Approach

Why neural networks?

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

Technical Approach

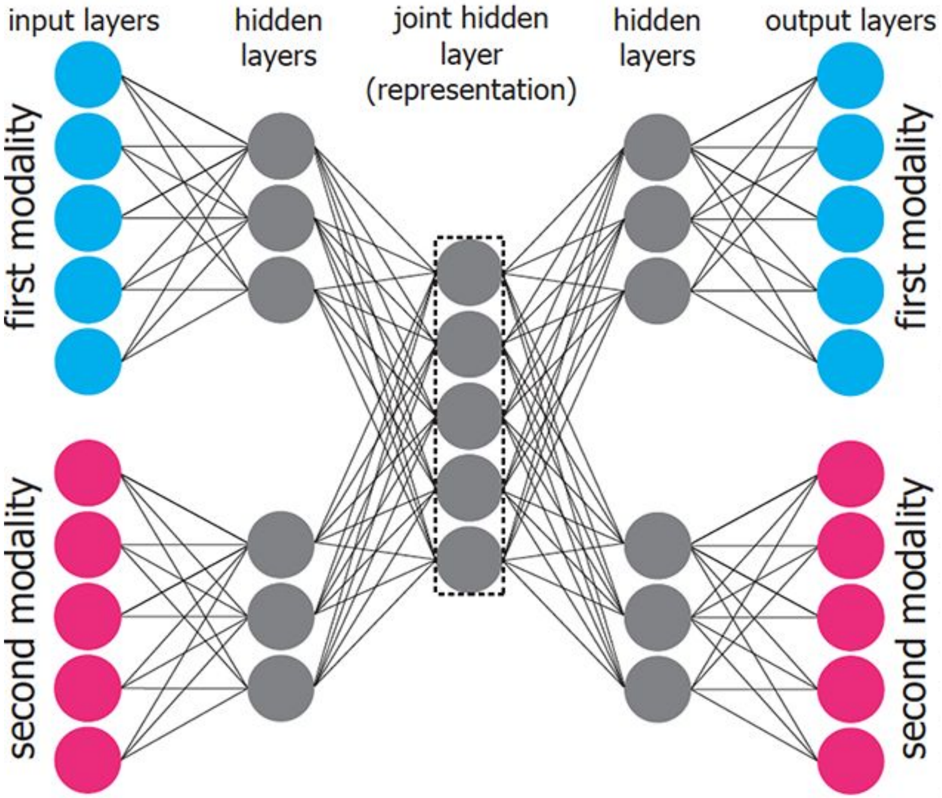
Why neural networks?

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

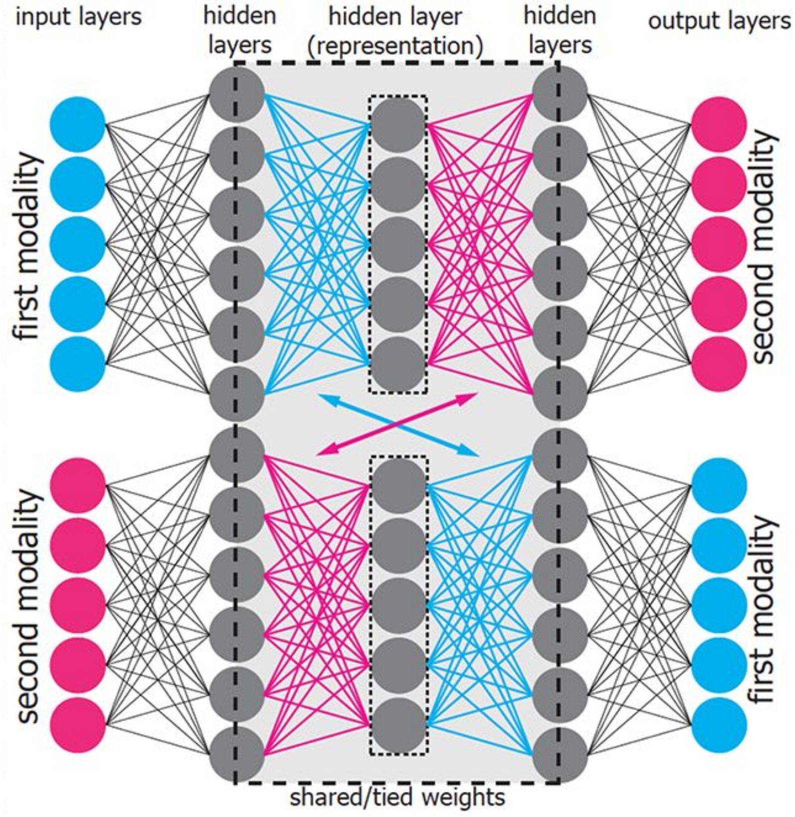
Why interpretability?

- 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

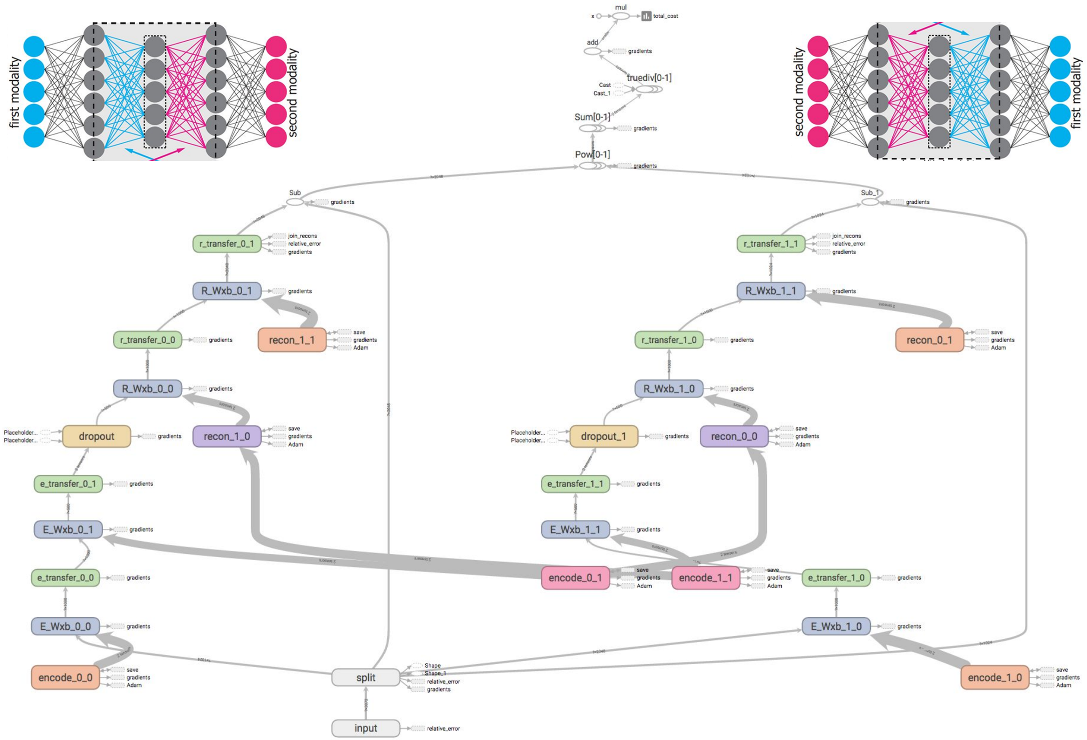
Multimodal Autoencoder



Bidirectional Deep Neural Network [3]

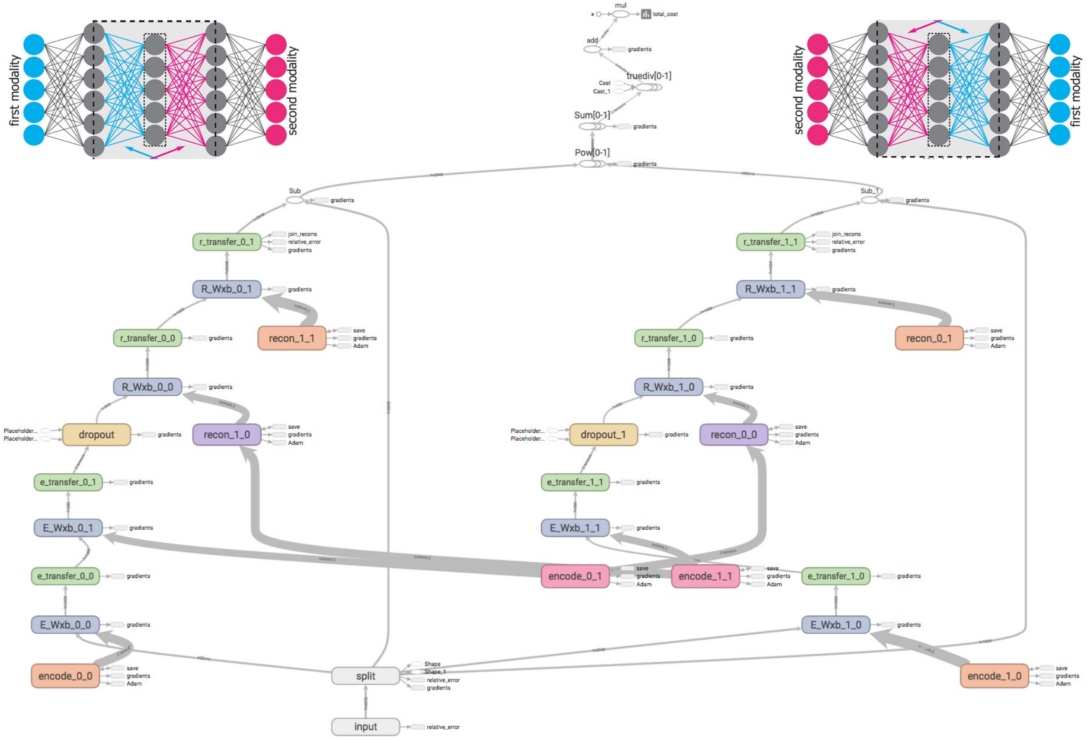


TensorFlow Implementation

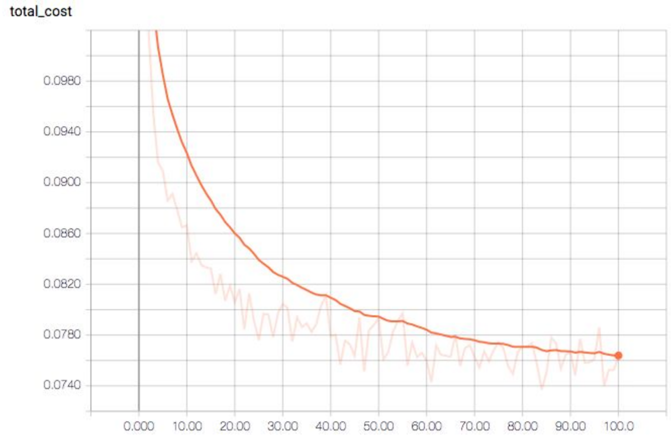


TensorBoard Visualization of our BiDNN

TensorFlow Implementation



TensorBoard Visualization of our BiDNN



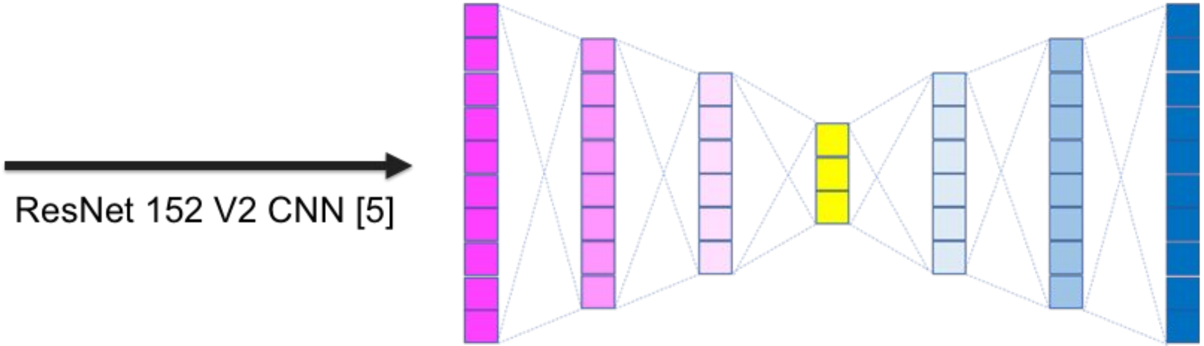
Minimization of BiDNN's Cost Function

Interpretability

✓ Use multimodal data to train bidirectional deep neural network.

‘the person turned the stove on’

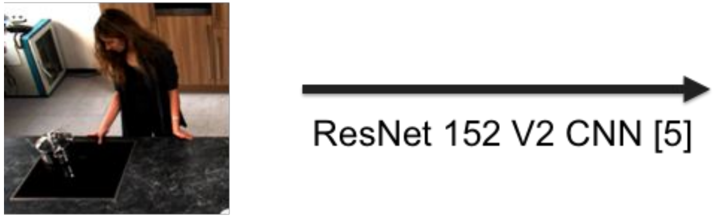
ELMo Word Vectors [4]

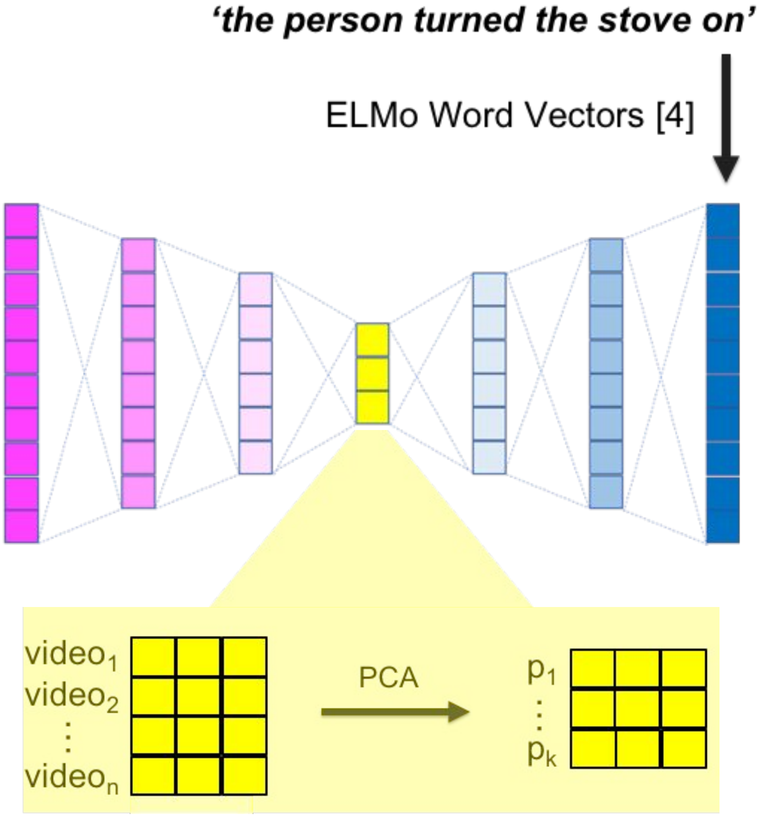


ResNet 152 V2 CNN [5]

Interpretability

✓ Use multimodal data to train bidirectional deep neural network.





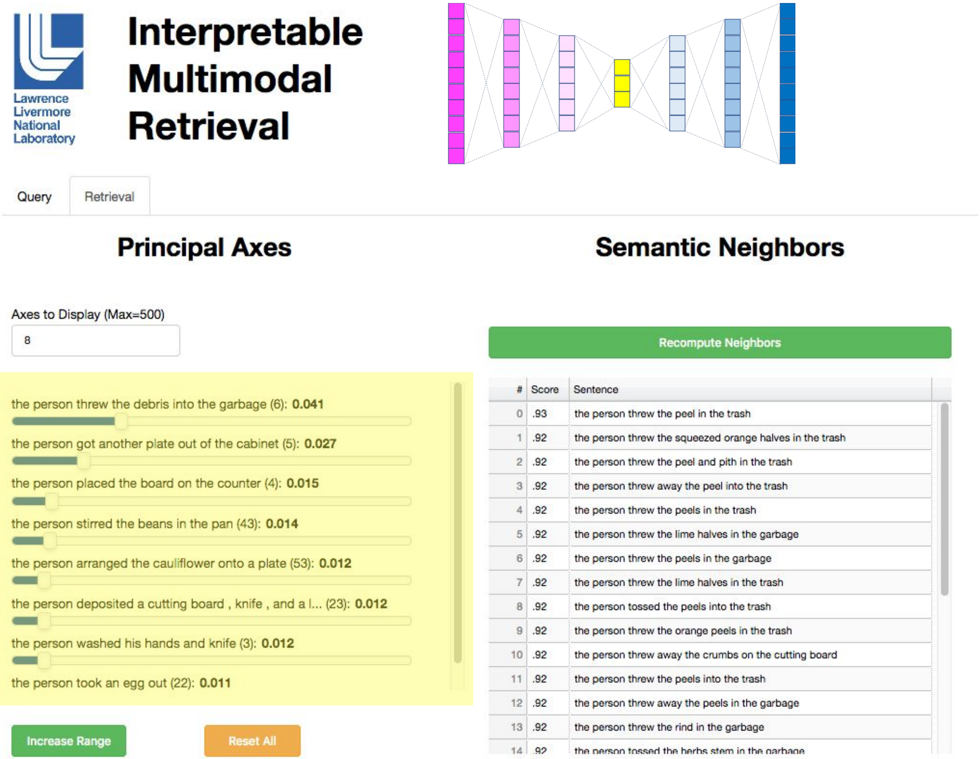
Create basis for multimodal feature space via PCA of trained network’s video encodings.

Interpretability

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.



Interpretable Multimodal Retrieval

Principal Axes

Axes to Display (Max=500)

8

the person threw the debris into the garbage (6): **0.041**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.014**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

the person washed his hands and knife (3): **0.012**

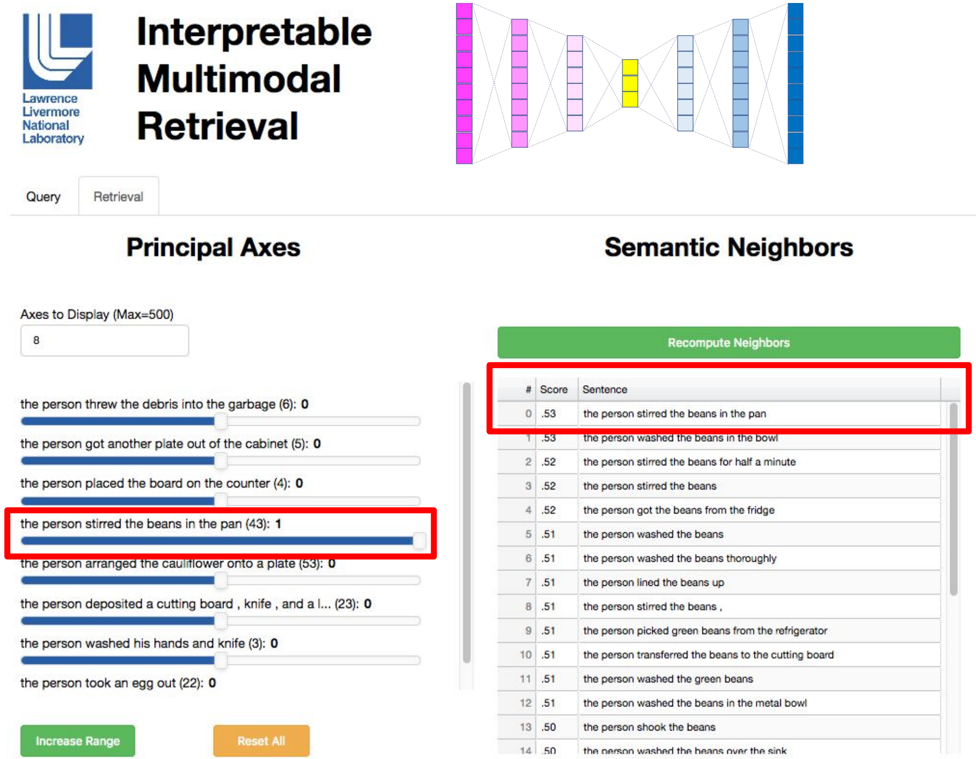
the person took an egg out (22): **0.011**

Semantic Neighbors

#	Score	Sentence
0	.93	the person threw the peel in the trash
1	.92	the person threw the squeezed orange halves in the trash
2	.92	the person threw the peel and pith in the trash
3	.92	the person threw away the peel into the trash
4	.92	the person threw the peels in the trash
5	.92	the person threw the lime halves in the garbage
6	.92	the person threw the peels in the garbage
7	.92	the person threw the lime halves in the trash
8	.92	the person tossed the peels into the trash
9	.92	the person threw the orange peels in the trash
10	.92	the person threw away the crumbs on the cutting board
11	.92	the person threw the peels into the trash
12	.92	the person threw away the peels in the garbage
13	.92	the person threw the rind in the garbage
14	.92	the person tossed the herbs stem in the garbage

Interpretability

Each axis is decoded into text space. The axis's interpretable label is the nearest text-description neighbor to its decoding.


Lawrence Livermore National Laboratory
Interpretable Multimodal Retrieval
Query
Retrieval

Principal Axes

Axes to Display (Max=500): 8

the person threw the debris into the garbage (6): **0**

the person got another plate out of the cabinet (5): **0**

the person placed the board on the counter (4): **0**

the person stirred the beans in the pan (43): **1**

the person arranged the cauliflower onto a plate (53): **0**

the person deposited a cutting board , knife , and a l... (23): **0**

the person washed his hands and knife (3): **0**

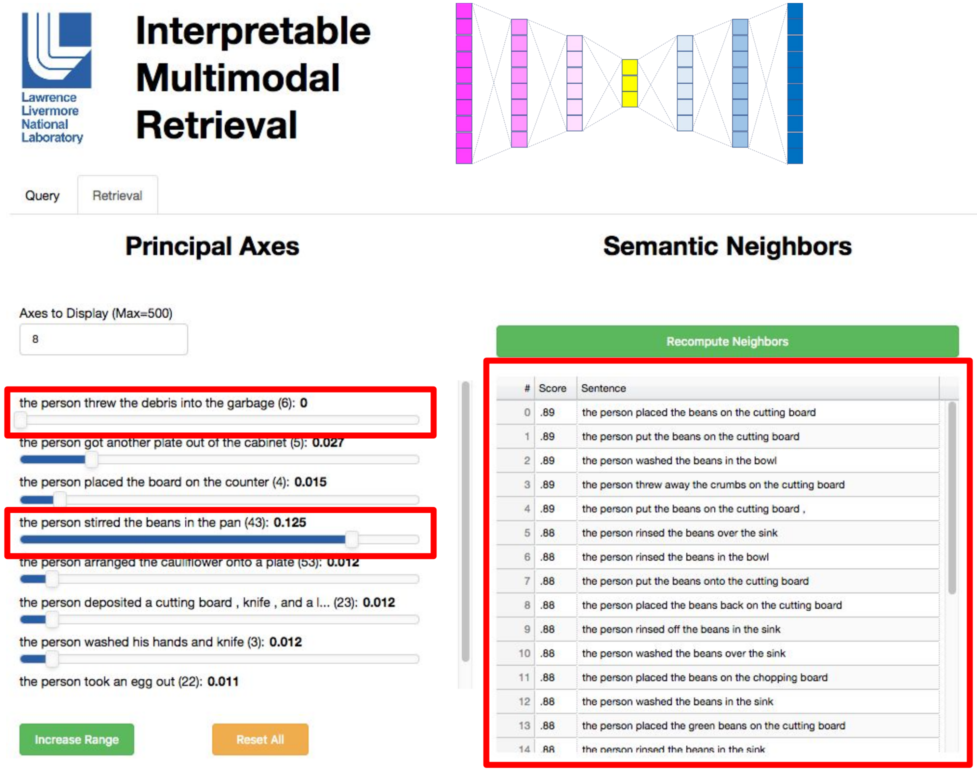
the person took an egg out (22): **0**

Semantic Neighbors

#	Score	Sentence
0	.53	the person stirred the beans in the pan
1	.53	the person washed the beans in the bowl
2	.52	the person stirred the beans for half a minute
3	.52	the person stirred the beans
4	.52	the person got the beans from the fridge
5	.51	the person washed the beans
6	.51	the person washed the beans thoroughly
7	.51	the person lined the beans up
8	.51	the person stirred the beans ,
9	.51	the person picked green beans from the refrigerator
10	.51	the person transferred the beans to the cutting board
11	.51	the person washed the green beans
12	.51	the person washed the beans in the metal bowl
13	.50	the person shook the beans
14	.50	the person washed the beans over the sink

Summary

Thus, users can adjust their queries' interpretable encodings to obtain a predictable effect on retrieved results.



Interpretable Multimodal Retrieval

Principal Axes

Axes to Display (Max=500): 8

the person threw the debris into the garbage (6): **0**

the person got another plate out of the cabinet (5): **0.027**

the person placed the board on the counter (4): **0.015**

the person stirred the beans in the pan (43): **0.125**

the person arranged the cauliflower onto a plate (53): **0.012**

the person deposited a cutting board , knife , and a l... (23): **0.012**

the person washed his hands and knife (3): **0.012**

the person took an egg out (22): **0.011**

Semantic Neighbors

#	Score	Sentence
0	.89	the person placed the beans on the cutting board
1	.89	the person put the beans on the cutting board
2	.89	the person washed the beans in the bowl
3	.89	the person threw away the crumbs on the cutting board
4	.89	the person put the beans on the cutting board ,
5	.88	the person rinsed the beans over the sink
6	.88	the person rinsed the beans in the bowl
7	.88	the person put the beans onto the cutting board
8	.88	the person placed the beans back on the cutting board
9	.88	the person rinsed off the beans in the sink
10	.88	the person washed the beans over the sink
11	.88	the person placed the beans on the chopping board
12	.88	the person washed the beans in the sink
13	.88	the person placed the green beans on the cutting board
14	.88	the person rinsed the beans in the sink

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

[1] Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., & Schiele, B. (2016). Recognizing fine-grained and composite activities using hand-centric features and script data. International Journal of Computer Vision, 119(3), 346-373.

[2] Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014, September). Coherent multi-sentence video description with variable level of detail. In German conference on pattern recognition (pp. 184-195). Springer, Cham.

[3] Vukotić, V., Raymond, C., & Gravier, G. (2016, October). Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. In Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion (pp. 37-44). ACM.

[4] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.