Visual Explainability in Machine Learning

Lecture 8: Concept Vectors: Utility in Training and Testing

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 7, 2023

Short Course Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/sps-education-short-course/

{alregib, mohit.p}@gatech.edu

Title: Visual Explainability in Machine Learning

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

https://alregib.ece.gatech.edu/

Short Course

Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

Day 1

Lecture 1: Why Explainability?

Lecture 2: What is Explainability?

Lectures 3 and 4: Popular Methods of Explainability

Day 2

Lecture 5: Evaluating Explainability

Lecture 6: Robustness: A proxy for Explainability

Day 3

Lectures 7 and 8: Reassessing Explainability via Uncertainty and Concept vectors

Lectures 9 and 10: Causality and Generative AI

Outline

Lecture 8: Concept Vectors: Utility in Training and Testing

- Concept Vectors for Explainability
- Testing with Concept Activation Vectors
- Concept Retrieval
  - Case study in seismic interpretability
  - Training for concept retrieval
- Concept Weights
  - Regularization-based concepts
  - Preprocessing-based concepts
  - Sparsity-based concepts
  - Color space-based concepts
  - Texture-based concepts
- Takeaways

Outline

Lecture 8: Concept Vectors: Utility in Training and Testing

- Concept Vectors for Explainability
- Testing with Concept Activation Vectors
- Concept Retrieval
  - Case study in seismic interpretability
  - Training for concept retrieval
- Concept Weights
  - Regularization-based concepts
  - Preprocessing-based concepts
  - Sparsity-based concepts
  - Color space-based concepts
  - Texture-based concepts
- Takeaways

Concept Vectors

What are concept vectors?

Interpretable semantic notions of an abstract idea that can be represented mathematically

- Textures
  - Soft
  - Rough
  - …
- Shapes
  - Circular
  - Cube
  - …
- Semantic concepts
  - Nose
  - Beak
  - …

Label: Tabby cat

Relevant concepts:

- Orange fur
- Cat-like shape
- Nose
- Eyes
- Whiskers

Concept Vectors

Explanations via Concept Vectors

Explanations: Weight the importance of concepts for the task at hand

- Textures
    - Soft
    - Rough
    - …
- Shapes
    - Circular
    - Cube
    - …
- Semantic concepts
    - Nose
    - Beak
    - …

Label: Tabby cat

For classification, how relevant is the texture of the fur as compared to the shape of the animal?

Concept Vectors

Explanations via Concept Vectors

Imagenet-trained Neural Networks are biased to texture rather than shape

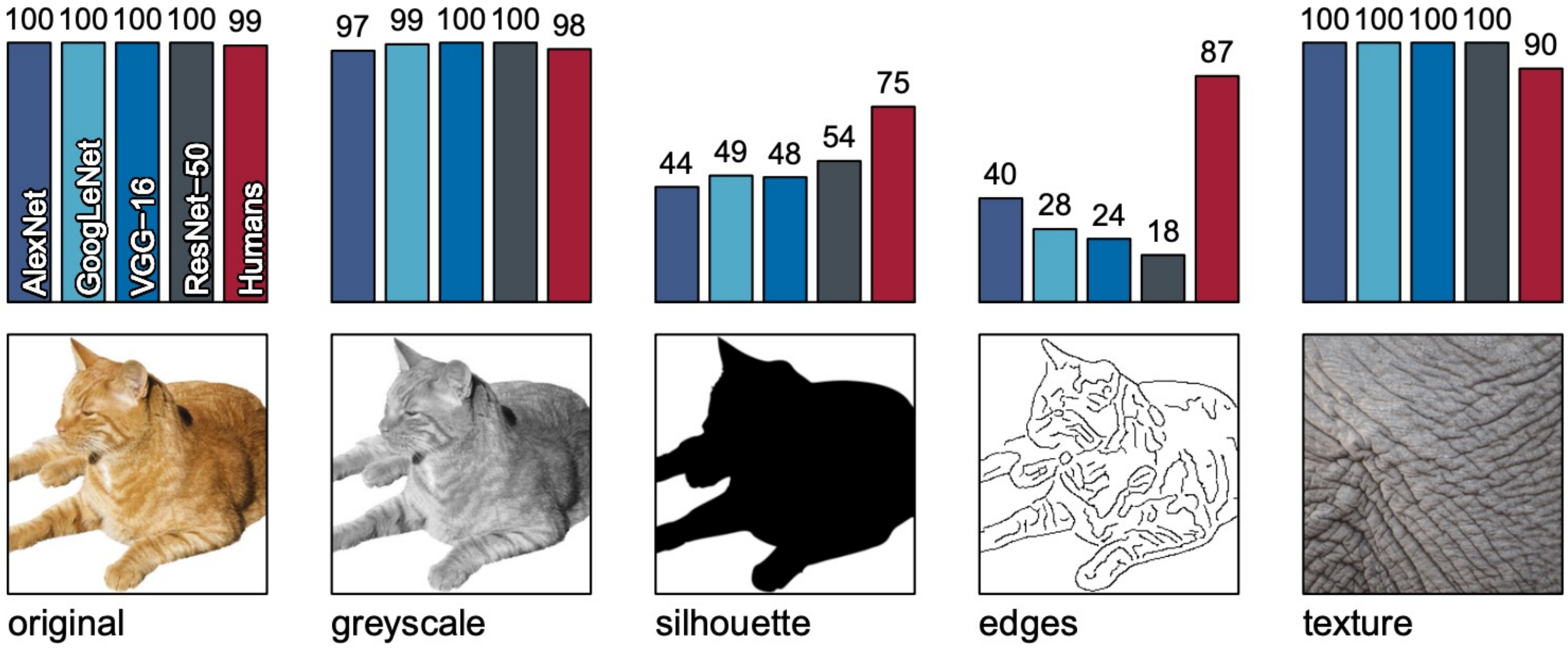
(a) Texture image		(b) Content image		(c) Texture-shape cue conflict	
81.4%	**Indian elephant**	71.1%	**tabby cat**	63.9%	**Indian elephant**
10.3%	indri	17.3%	grey fox	26.4%	indri
8.2%	black swan	3.3%	Siamese cat	9.6%	black swan

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Concept Vectors

Explanations via Concept Vectors

Imagenet-trained Neural Networks are biased to texture rather than shape



Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Utilizing Concept Vectors

Training on Style-Transferred Images

Key Insight: Neural Networks overfit to lower-order concepts like color and texture rather than higher-order concept like shape

The above insight suggests training on style-transferred images where shape remains the same but texture differs is more robust

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Outline

Lecture 8: Concept Vectors: Utility in Training and Testing

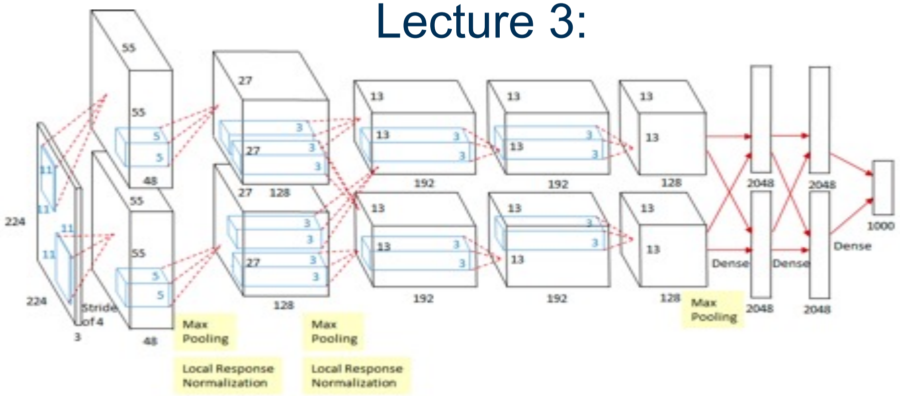
- Concept Vectors for Explainability
- Testing with Concept Activation Vectors
- Concept Retrieval
    - Case study in seismic interpretability
    - Training for concept retrieval
- Concept Weights
    - Regularization-based concepts
    - Preprocessing-based concepts
    - Sparsity-based concepts
    - Color space-based concepts
    - Texture-based concepts
- Takeaways

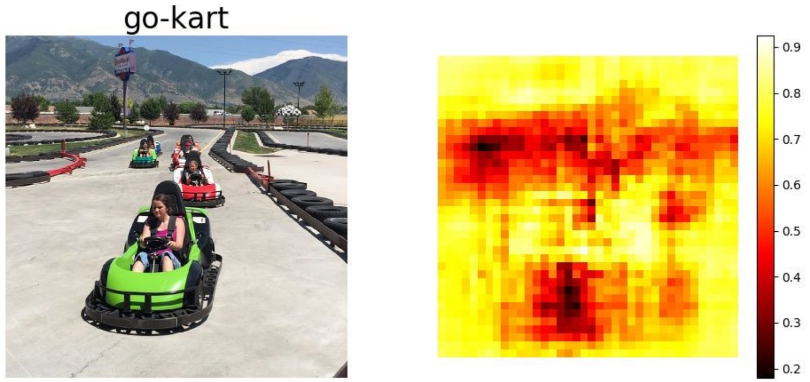
As an Aside..

Constructing Explanations vs Evaluating Explanations

Lecture 3: Construct explanations via occlusion; Lecture 5: Evaluate explanations via occlusion

Lecture 3:





Lecture 5:

→ Trained Model → Spoonbill

Evaluate the effect of Explainability in Network Evaluation

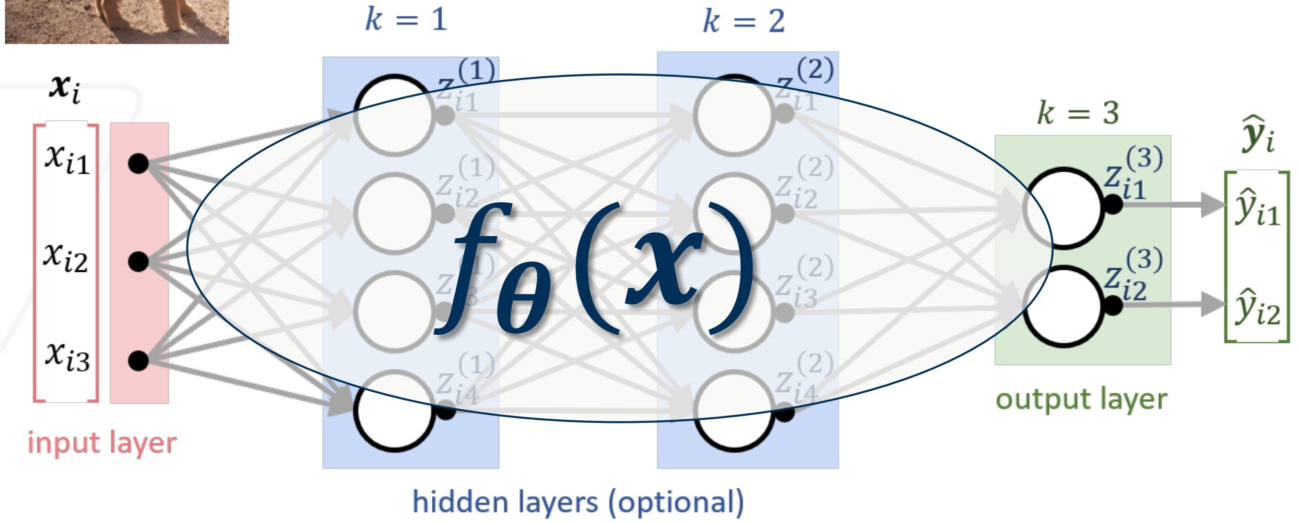
As an Aside..

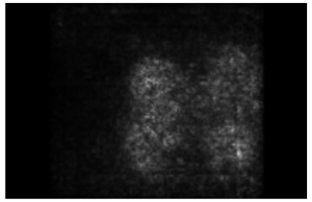
Constructing Explanations vs Evaluating Explanations

Lecture 3 and 4: Construct explanations via gradients; Lecture 6: Evaluate explanations via gradients

Image

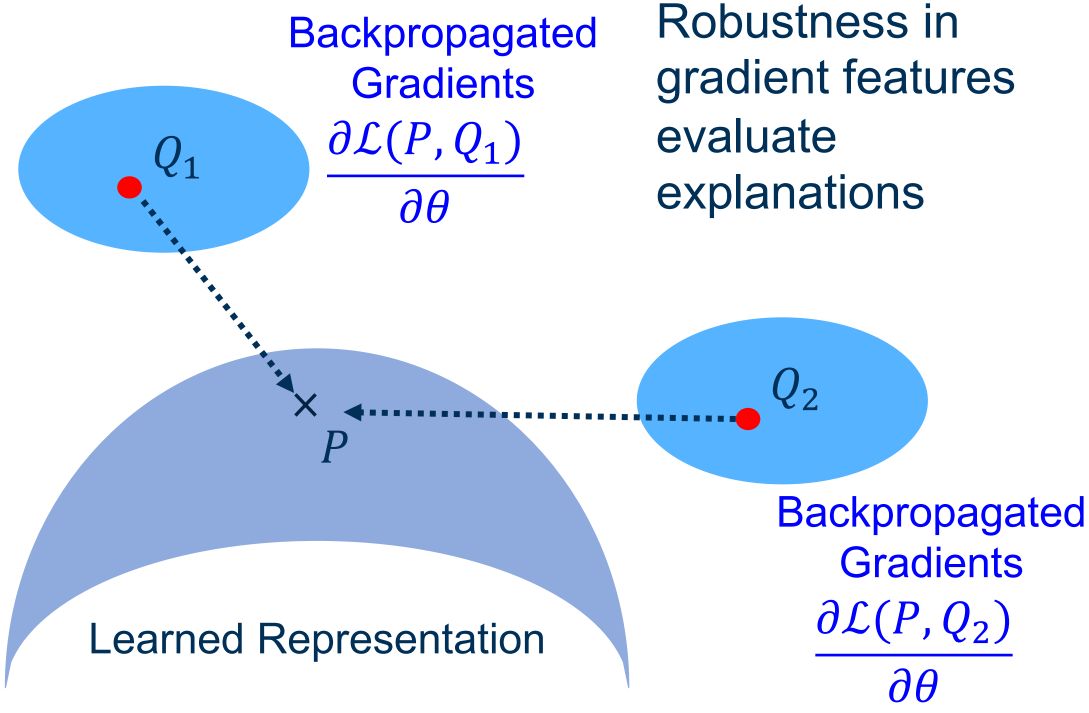
$$f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f'''(a)}{3!}(x-a)^3+\cdots,$$





Gradients are explanations

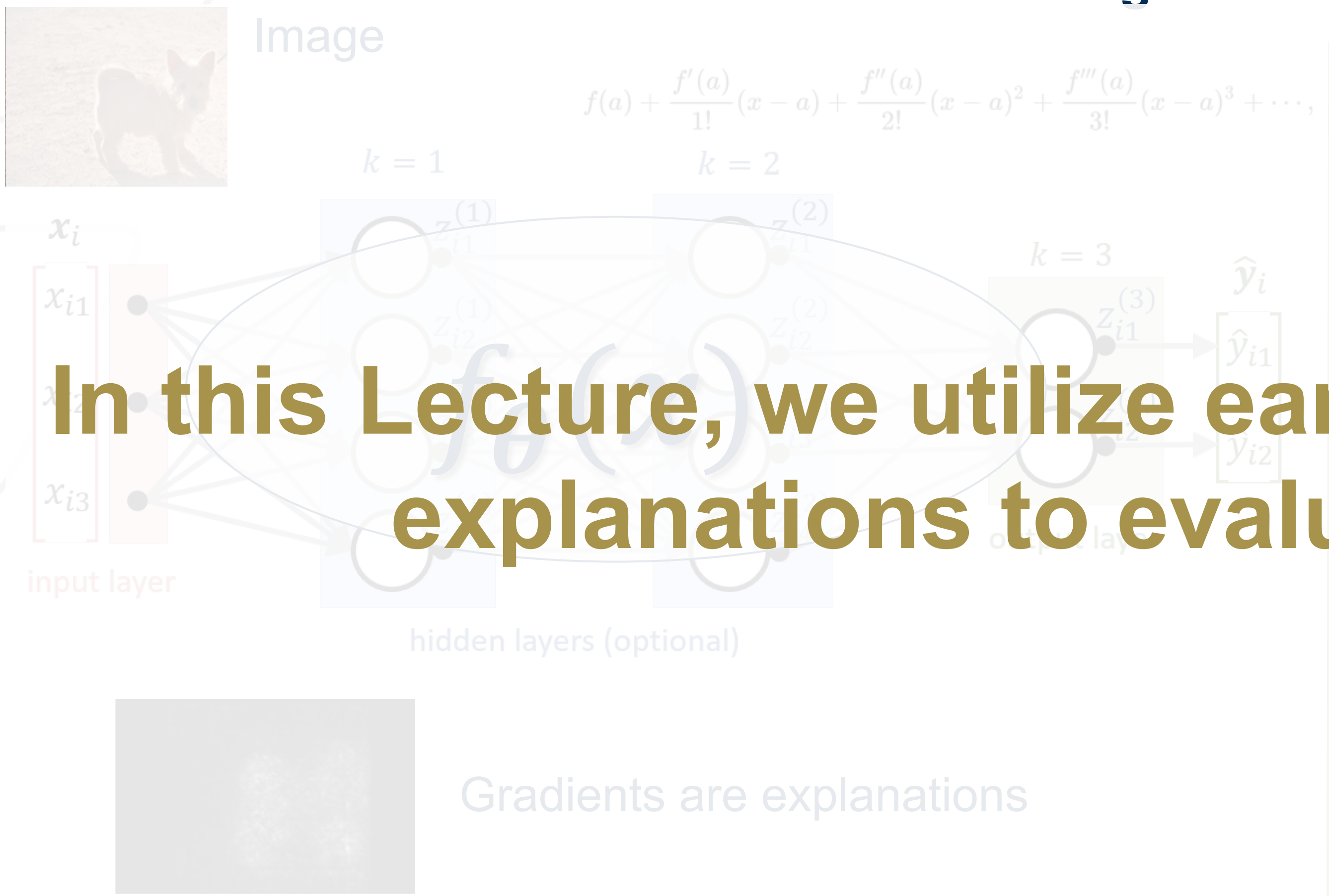
Robustness in gradient features evaluate explanations

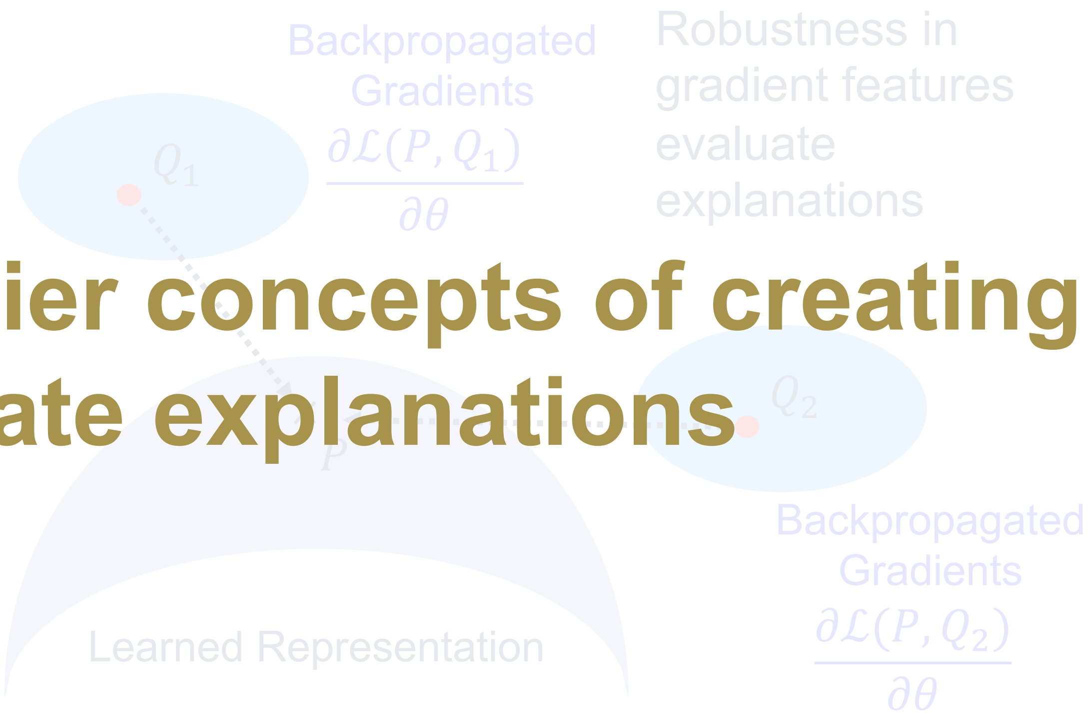


As an Aside..

Constructing Explanations vs Evaluating Explanations

Lecture 3 and 4: Construct explanations via gradients; Lecture 6: Evaluate explanations via gradients



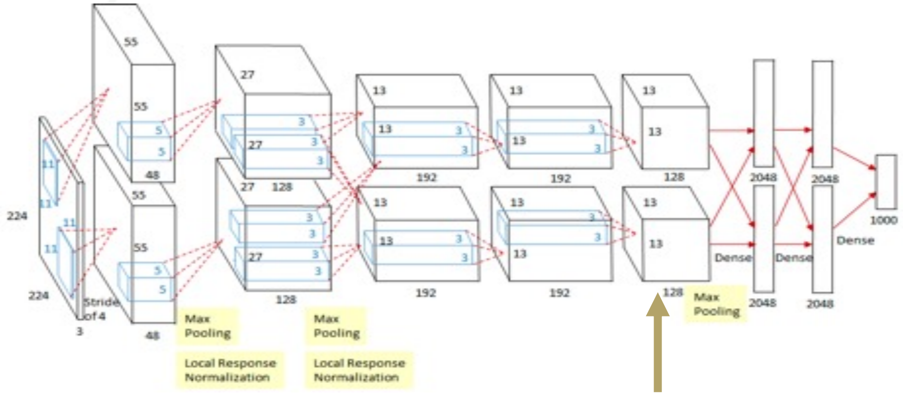


In this Lecture, we utilize earlier concepts of creating explanations to evaluate explanations

Recap: Concept Retrieval

Concept Retrieval as Explanations

In Lecture 3: Maximally Activating Patches were retrieved and acted as explanations



conv5

Maximally Activating Patches: Image patches in the input that cause the <u>maximum activations of certain filters</u>

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Testing with Concept Vectors

What are Concept Activation Vectors (CAV)?

Concept Activation Vectors (CAVs) are the activations from known concepts within data

In Lecture 3: Patches are from data. In Lecture 8: Concepts are *features* within data

Concepts:

Retrieve:

Goal: Given exemplary concept patches, retrieve all relevant concepts in the image

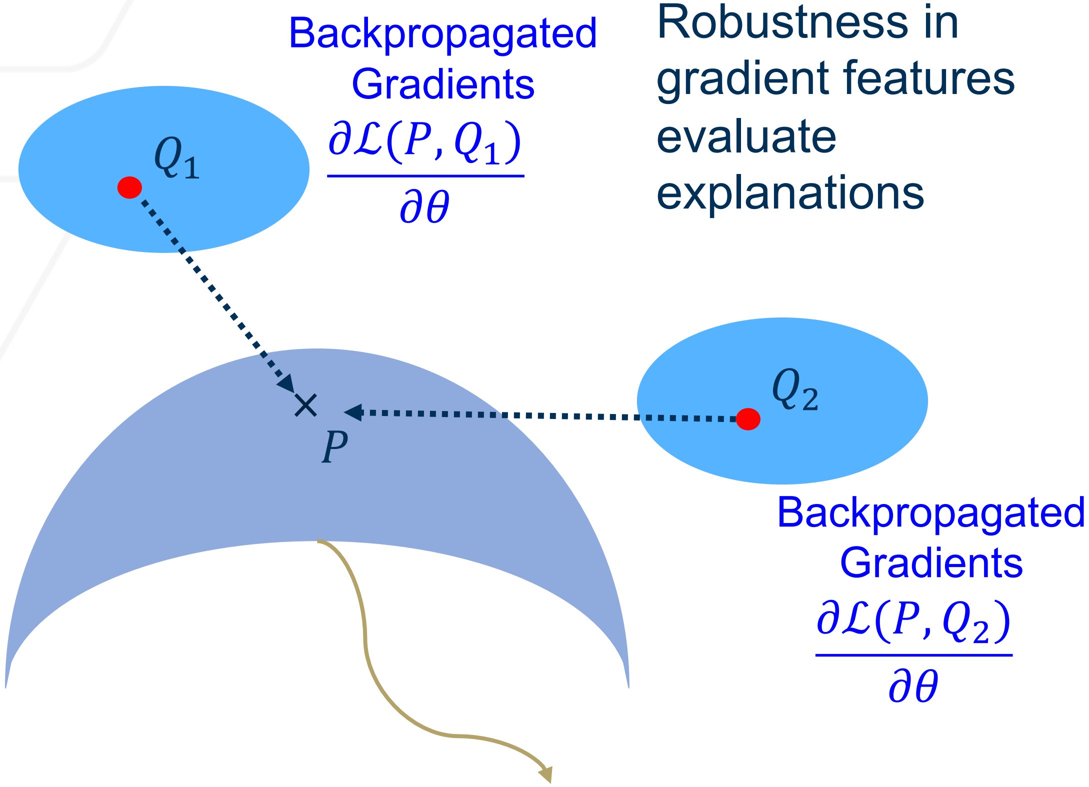
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Testing with Concept Vectors

Methodology

Concept Activation Vectors (CAVs) are explicit evaluation techniques that require a new classifier

Robustness in gradient features evaluate explanations



Step 1: Forward pass all images through a trained network

Step 2: Record all activations for all training data

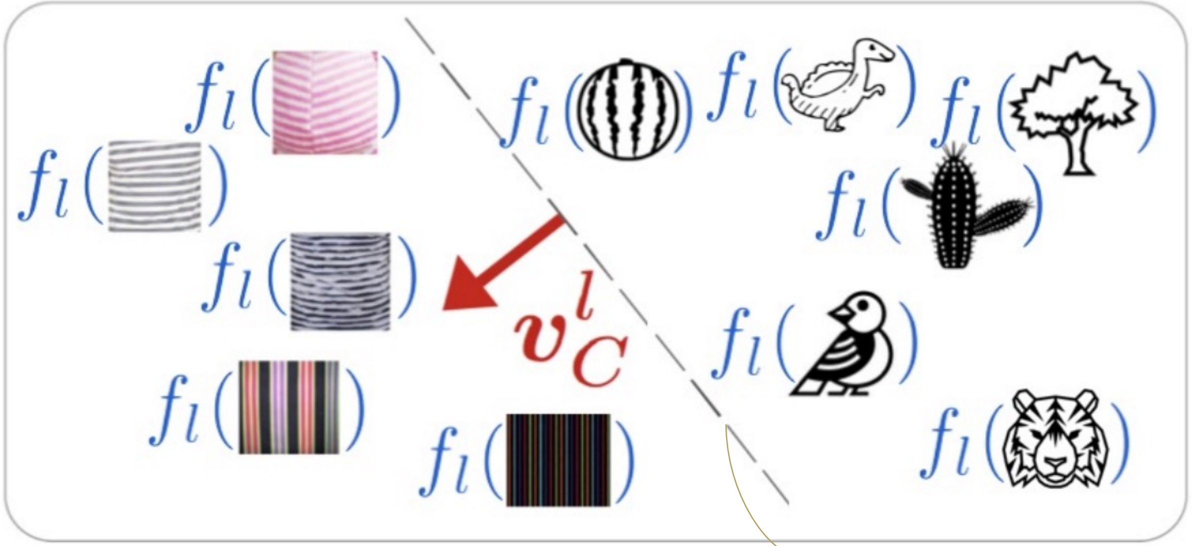
Step 3: Construct a linear classifier on all (labeled) activation concepts (from any layer)

In Lecture 6, we backpropagate using the base network: CNNs, Transformers etc.

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Testing with Concept Vectors
Methodology

Gradients are used as directional information using simple linear classifiers and gradients



Step 1: Forward pass all images through a trained network

Step 2: Record all activations for all training data

Step 3: Construct a linear classifier on all (labeled) activation concepts (from any layer)

Step 4: To obtain explanations, find sign of gradient from test image against the trained classifier. Positive sign indicates the concepts influence the decision, while negative sign indicates no influence

Trained classifier from Step 3

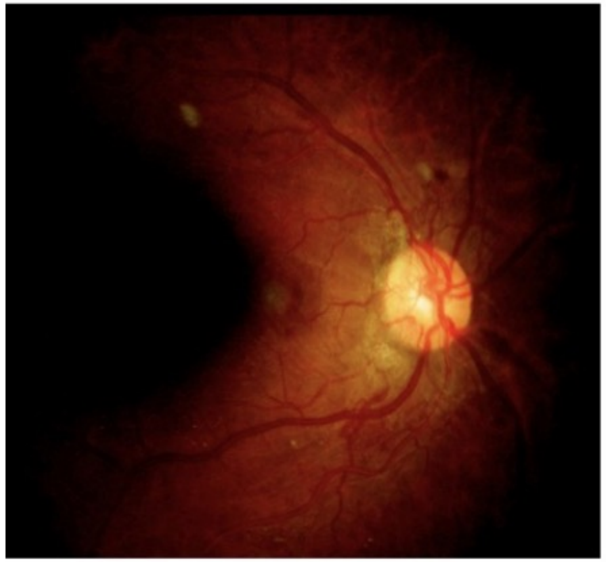
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

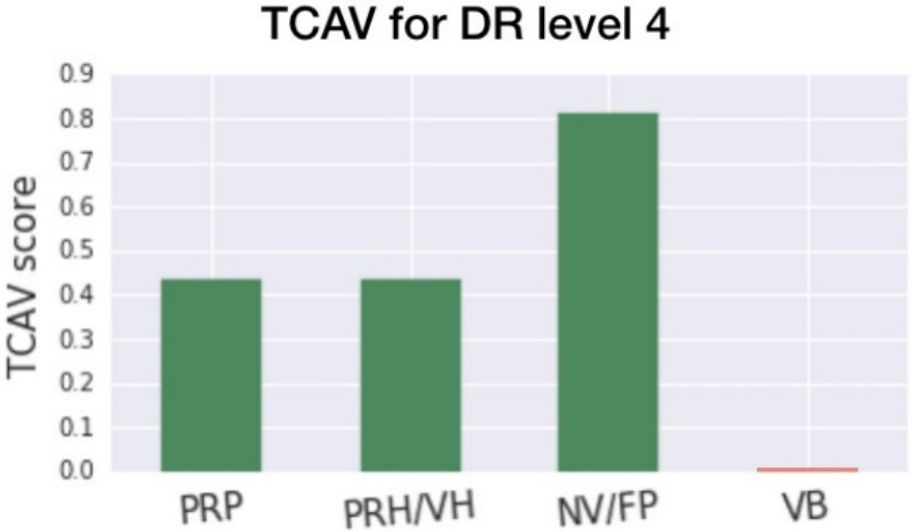
Testing with Concept Vectors

Results

Gradients are used as directional information using simple linear classifiers and gradients

DR level 4 Retina





Green = Important concept

Red = Irrelevant concept

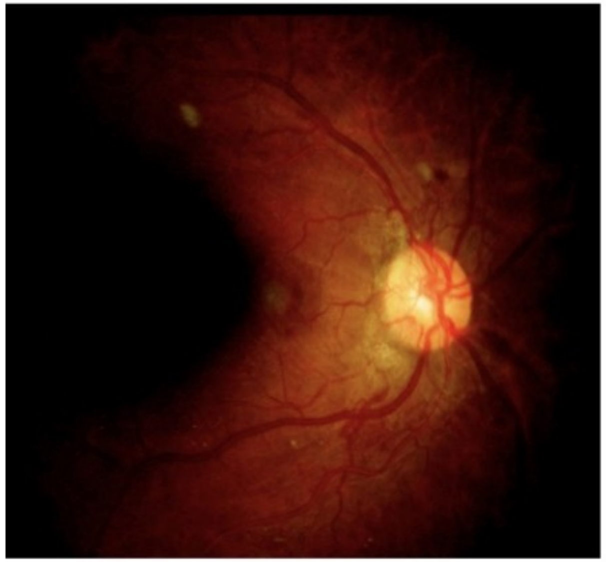
Given biomarkers, TCAV attributes the severity level of Diabetic Retinopathy (DR) to the biomarkers

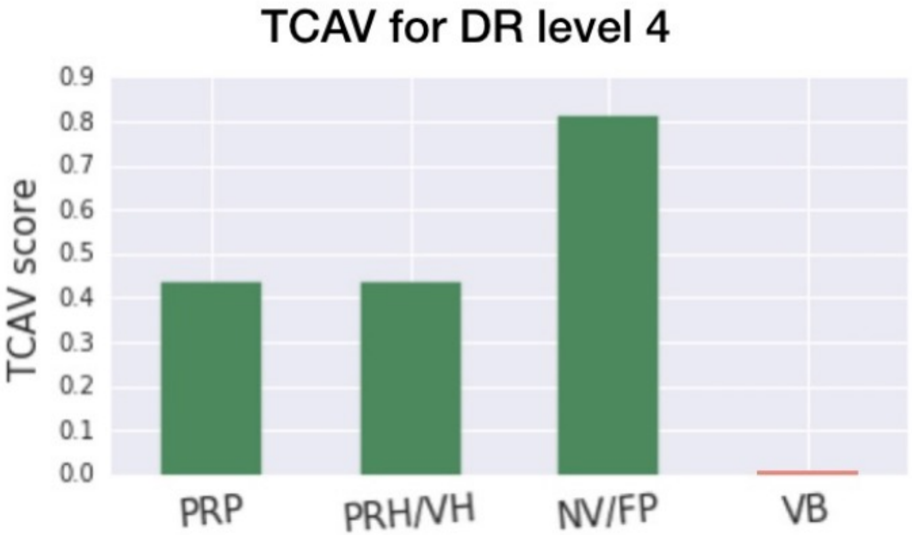
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Testing with Concept Vectors
Summary

Gradients are used as directional information using simple linear classifiers and gradients

DR level 4 Retina





Green = Important concept

Red = Irrelevant concept

- **Provides feature-based explanations**: Combines low-level features with high-level semantics
- **Labeled features** (or concepts) are **not always available** in visual data
- Requires an additional classifier

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Outline

Lecture 8: Concept Vectors: Utility in Training and Testing

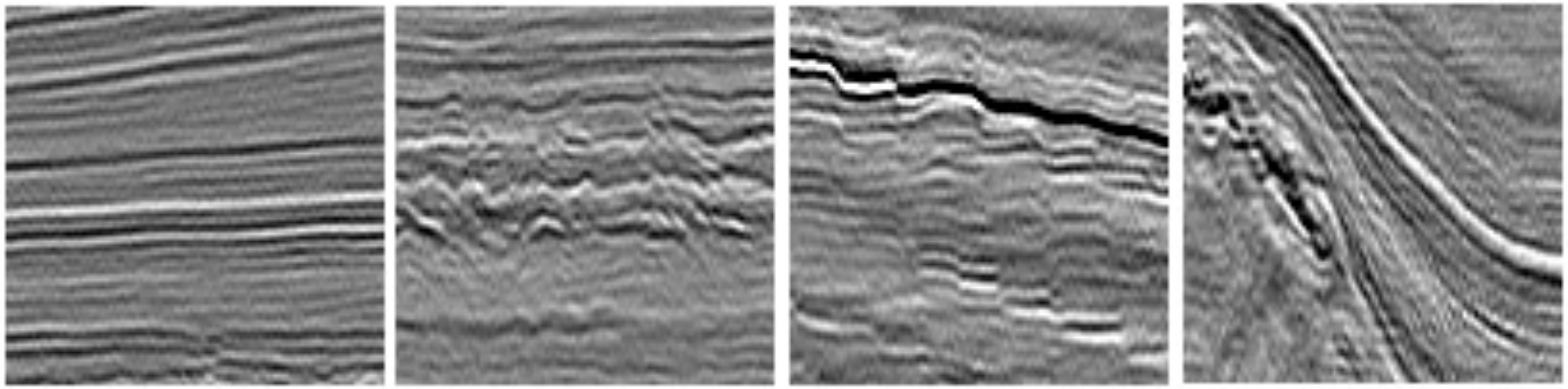
- Concept Vectors for Explainability
- Testing with Concept Activation Vectors
- Concept Retrieval
  - Case study in seismic interpretability
  - Training for concept retrieval
- Concept Weights
  - Regularization-based concepts
  - Preprocessing-based concepts
  - Sparsity-based concepts
  - Color space-based concepts
  - Texture-based concepts
- Takeaways

Case Study: Concept Retrieval in Seismic Images

Case Study: Seismic Interpretability

Given an exemplary seismic structure, retrieve concepts of the same structure

Exemplar concepts



Horizon | Chaotic | Fault | Salt dome

Not given: Large training concepts (structures), that can be used to predict concepts

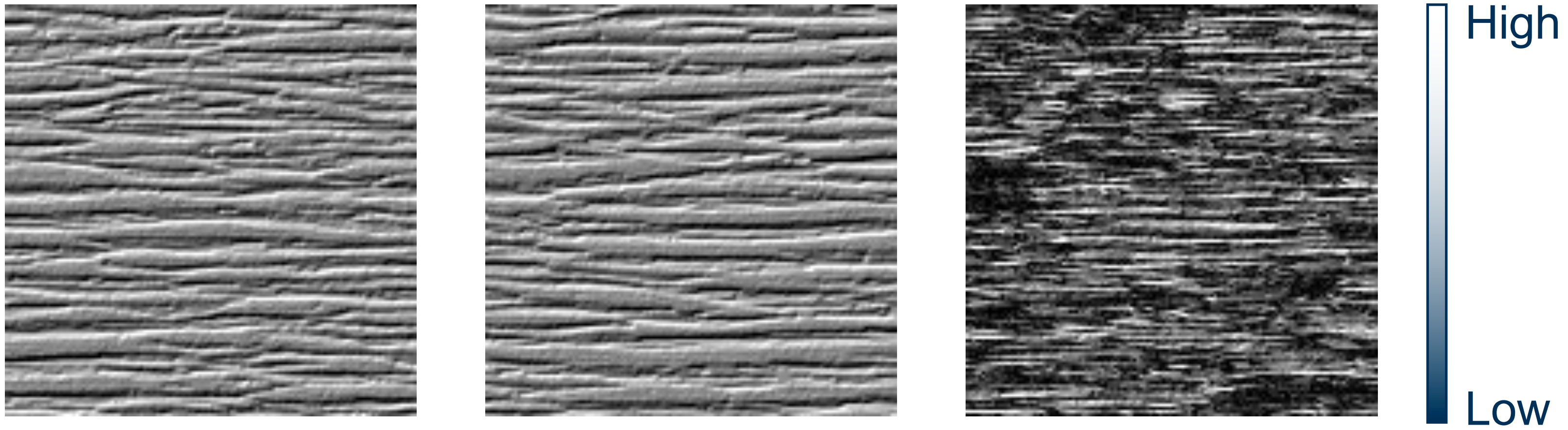
Known: The concepts of horizons, chaos, fault, and salt dome do not occur within the same pixels, i.e. they are orthogonal

We utilize this knowledge to train a neural network for pixel-wise segmentation based on concepts

Case Study: Concept Retrieval in Seismic Images

Case Study: Texture Retrieval

Challenge 1: Typical measures of Pixel-Pixel Correspondence does not apply



(a) Texture image 1

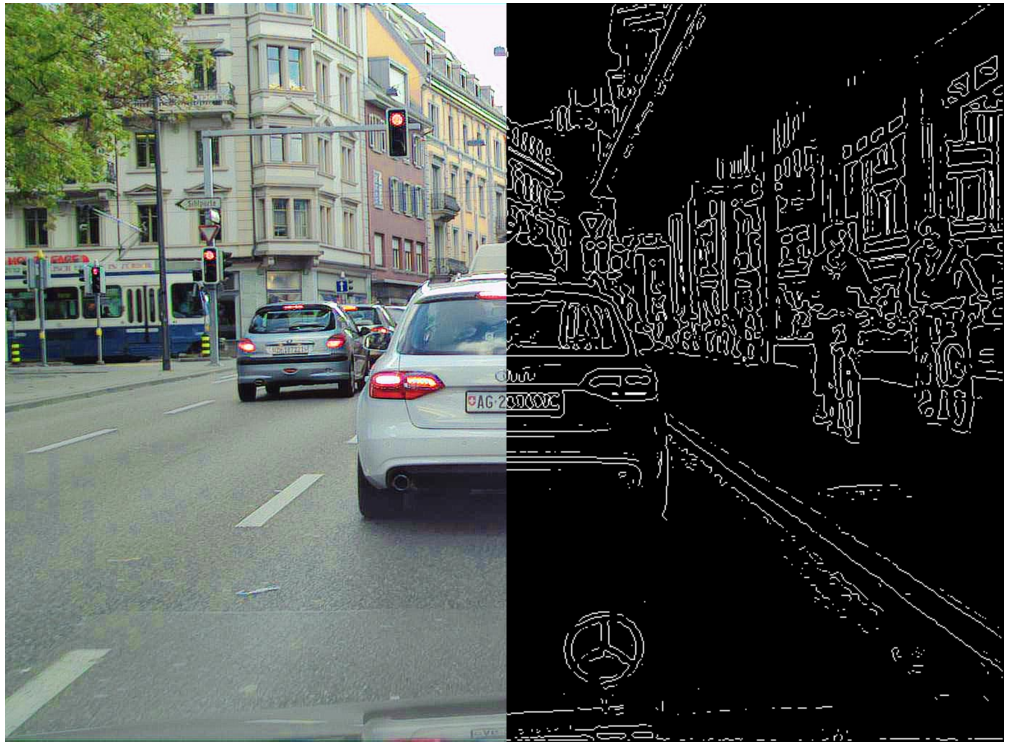
(b) Texture image 2

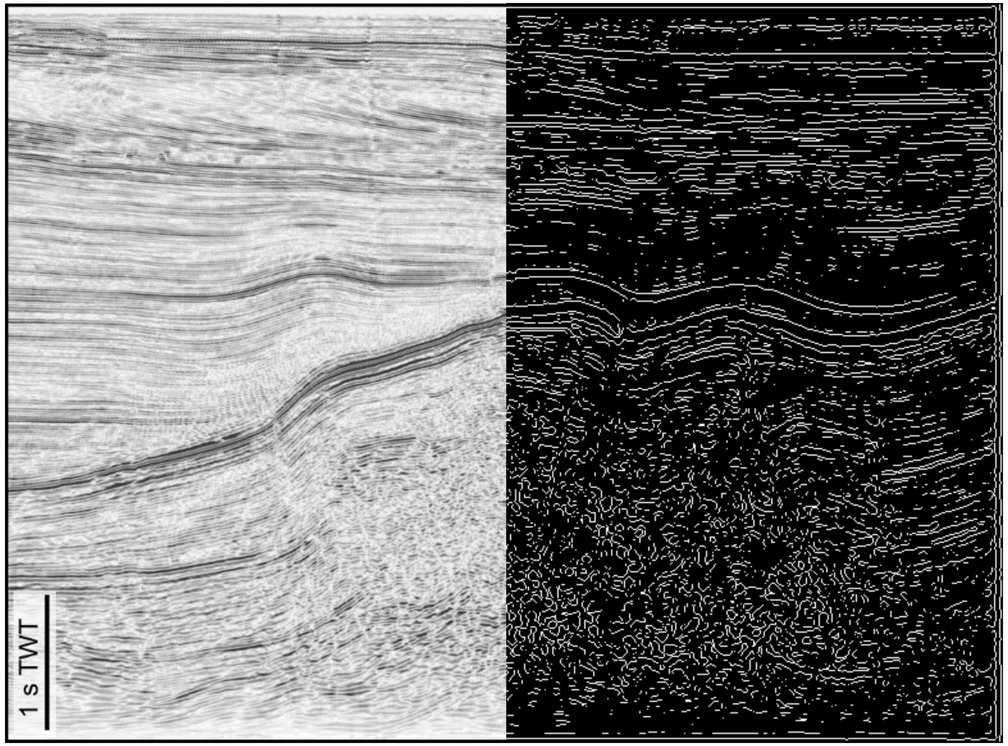
(c) Absolute difference between (a) and (b)

Case Study: Concept Retrieval in Seismic Images

Case Study: Texture Retrieval

Challenge 2: Boundaries between objects are not well defined





Case Study: Concept Retrieval in Seismic Images

Case Study: Texture Retrieval

Challenge 3: No color information

Case Study: Concept Retrieval in Seismic Images

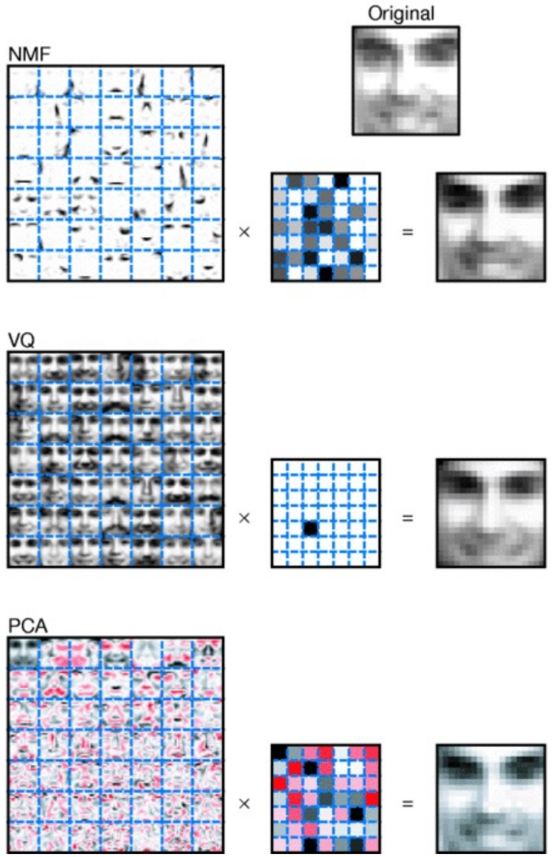
Traditional Non-negative Matrix Factorization

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

Traditional NMF formulation:

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \qquad \text{s.t.}\, \mathbf{W}, \mathbf{H} \geq 0$$

$$\boxed{\underset{\mathbf{W},\mathbf{H}}{\arg\min} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2, s.t. \mathbf{W}, \mathbf{H} \geq 0}$$



Lee, D. D., and H. S. Seung, 1999, Learning the parts of objects by non-negative matrix factorization.: Nature, 401, 788–91

Case Study: Concept Retrieval in Seismic Images

Solution: Orthogonal Concept Activation Vectors

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

$$\arg\min_{\mathbf{W},\mathbf{H}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 + \lambda_1||\mathbf{W}||_F^2 + \lambda_2||\mathbf{H}||_F^2 + \gamma_1||\mathbf{H}\mathbf{H}^T - \mathbf{I}||_F^2$$
$$\text{s.t.} \mathbf{W}, \mathbf{H} \geq 0 \quad \text{and} \quad \rho(\mathbf{w}_i) = \rho_w$$

- $\mathbf{X} \in \mathbb{R}_+^{N_p \times N_s}$: data matrix containing seismic images
- $\mathbf{W} \in \mathbb{R}_+^{N_p \times N_f}$: feature matrix
- $\mathbf{H} \in \mathbb{R}_+^{N_f \times N_s}$: coefficients matrix
- $\rho(\cdot)$: sparisty of a vector

$$\rho(\mathbf{w}_i) = \frac{\sqrt{N_p} - \frac{||\mathbf{w}_i||_1}{||\mathbf{w}_i||_2}}{\sqrt{N_p} - 1}$$

Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

Solution: Orthogonal Concept Activation Vectors

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

- $\mathbf{X}$: contains seismic images as columns
- $\mathbf{W}$: initialized with sparse features extracted from $\mathbf{X}$

$\mathbf{X}$ → [*k*-means clustering on each class] → [apply sparsity constraint*] → $\mathbf{W}, \mathbf{Q}$

- $\mathbf{Q} \in \{0,1\}^{N_f \times N_l}$: is a binary cluster membership matrix used to extract the output labels
- $\mathbf{H}$: is initialized with uniform random values in $[0,1]$

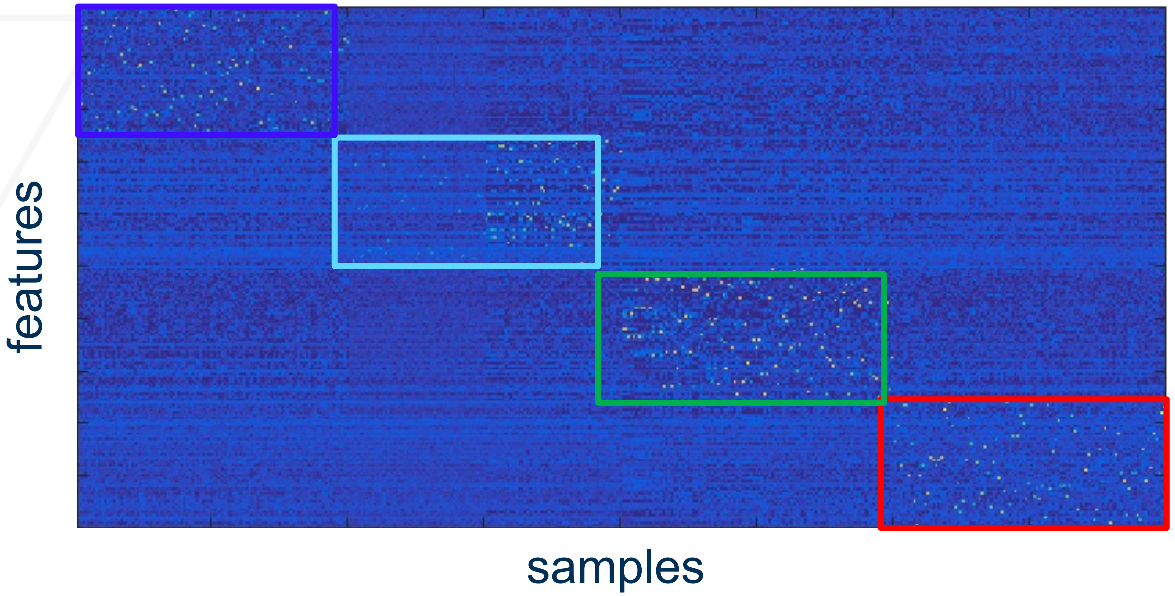
Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

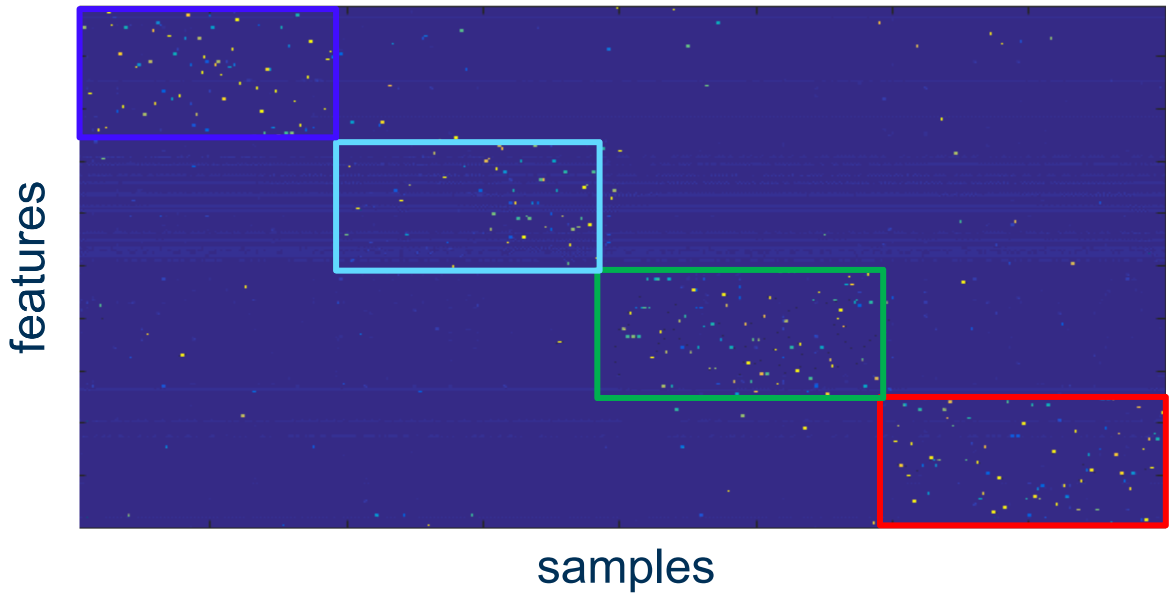
Solution: Orthogonal Concept Activation Vectors

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

H without the orthogonality term:



H with the orthogonality term:



Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

Solution: Orthogonal Concept Activation Vectors

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

$$\arg\min_{\mathbf{W},\mathbf{H}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 + \lambda_1||\mathbf{W}||_F^2 + \lambda_2||\mathbf{H}||_F^2 + \gamma_1||\mathbf{H}\mathbf{H}^T - \mathbf{I}||_F^2$$
$$\text{s.t.}\, \mathbf{W}, \mathbf{H} \geq 0 \quad \text{and} \quad \rho(\mathbf{w}_i) = \rho_w$$

$$\arg\min_{\mathbf{W}} \; ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 + \lambda_1||\mathbf{W}||_F^2 \;\; \text{s.t.}\, \mathbf{W} \geq 0, \rho(\mathbf{w}_i) = \rho_w$$

$$\arg\min_{\mathbf{H}} \; ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 + \gamma_1||\mathbf{H}\mathbf{H}^T - \mathbf{I}||_F^2 + \lambda_2||\mathbf{H}||_F^2 \;\; \text{s.t.}\, \mathbf{H} \geq 0$$

Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

Solution: Orthogonal Concept Activation Vectors

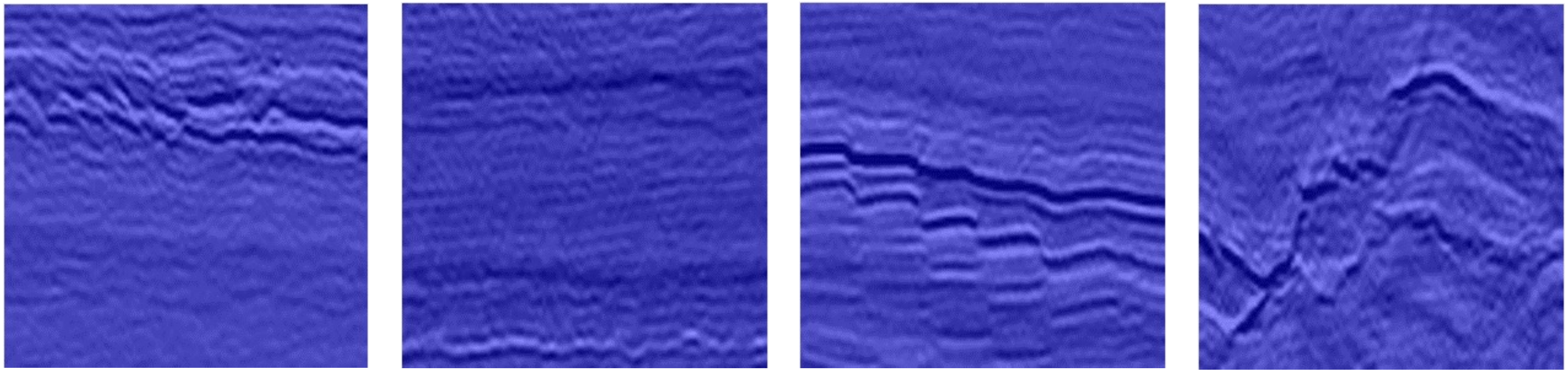
Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

$$\mathbf{W}^t = \frac{(\mathbf{W}^{t-1} \odot \mathbf{X}\mathbf{H}^{t-1^T} + \epsilon)_{ij}}{\mathbf{W}^{t-1}\mathbf{H}^{t-1}\mathbf{H}^{t-1^T} + \lambda_1 \mathbf{W}^{t-1} + \epsilon)_{ij}}$$

$$\mathbf{H}^t = \frac{\mathbf{H}^{t-1} \odot (\mathbf{W}^{t^T}\mathbf{X} + \gamma_1 \mathbf{H}^{t-1} + \epsilon)_{ij}}{\mathbf{W}^{t^T}\mathbf{W}^t\mathbf{H}^{t-1} + \gamma_1(\mathbf{H}^{t-1}\mathbf{H}^{t-1^T}\mathbf{H}^{t-1}) + \lambda_2 \mathbf{H}^{t-1} + \epsilon)_{ij}}$$

Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Train a deconvolution network to produce orthogonal activation vectors using non-negative matrix factorization

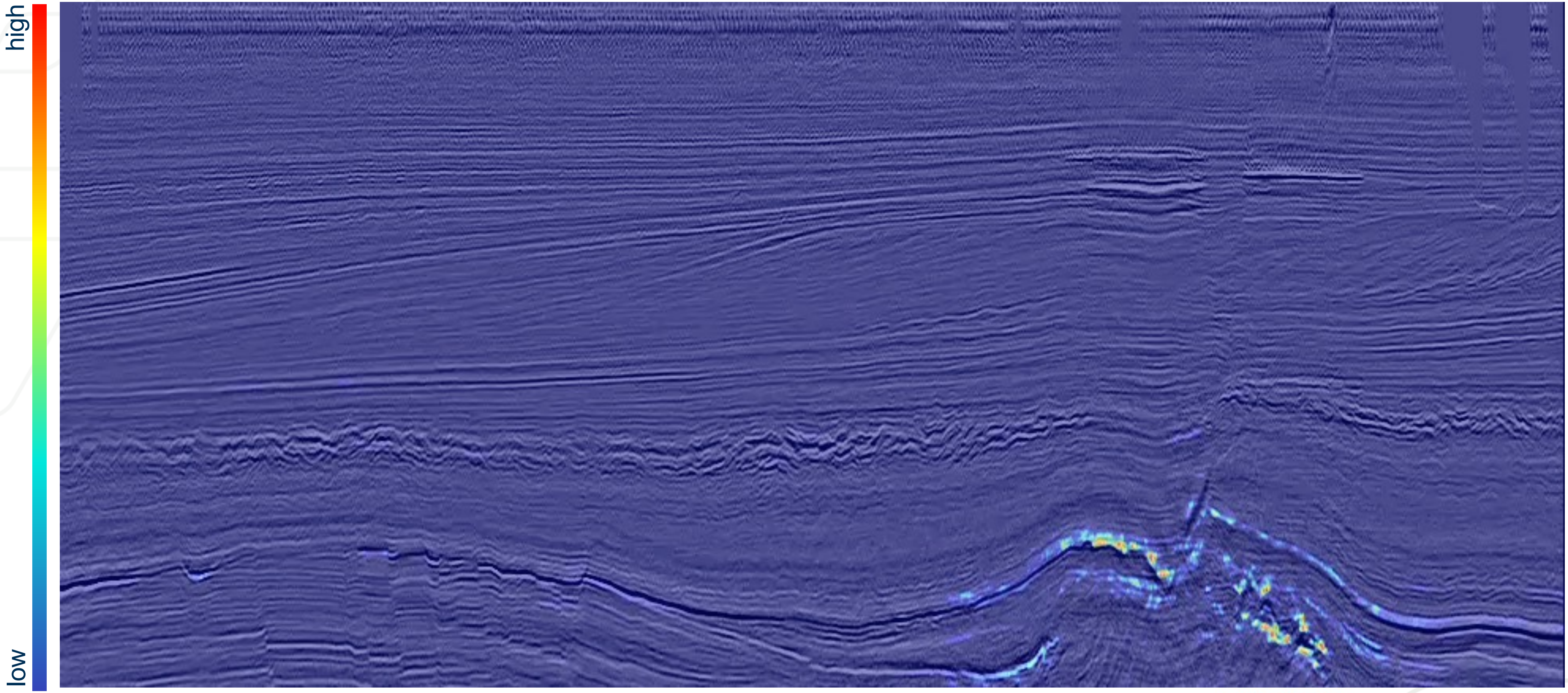


Horizon Chaotic Fault Salt dome

Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

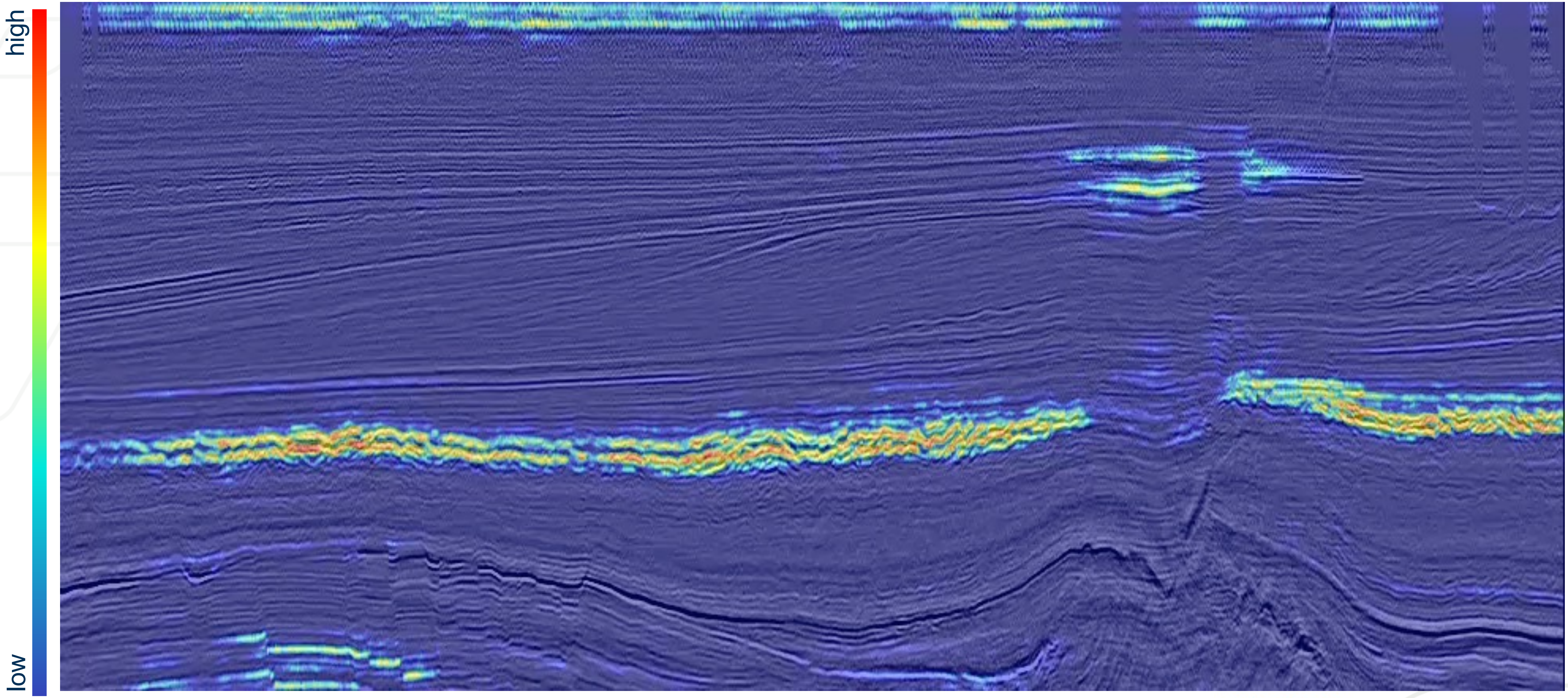
Results for Salt Dome



Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

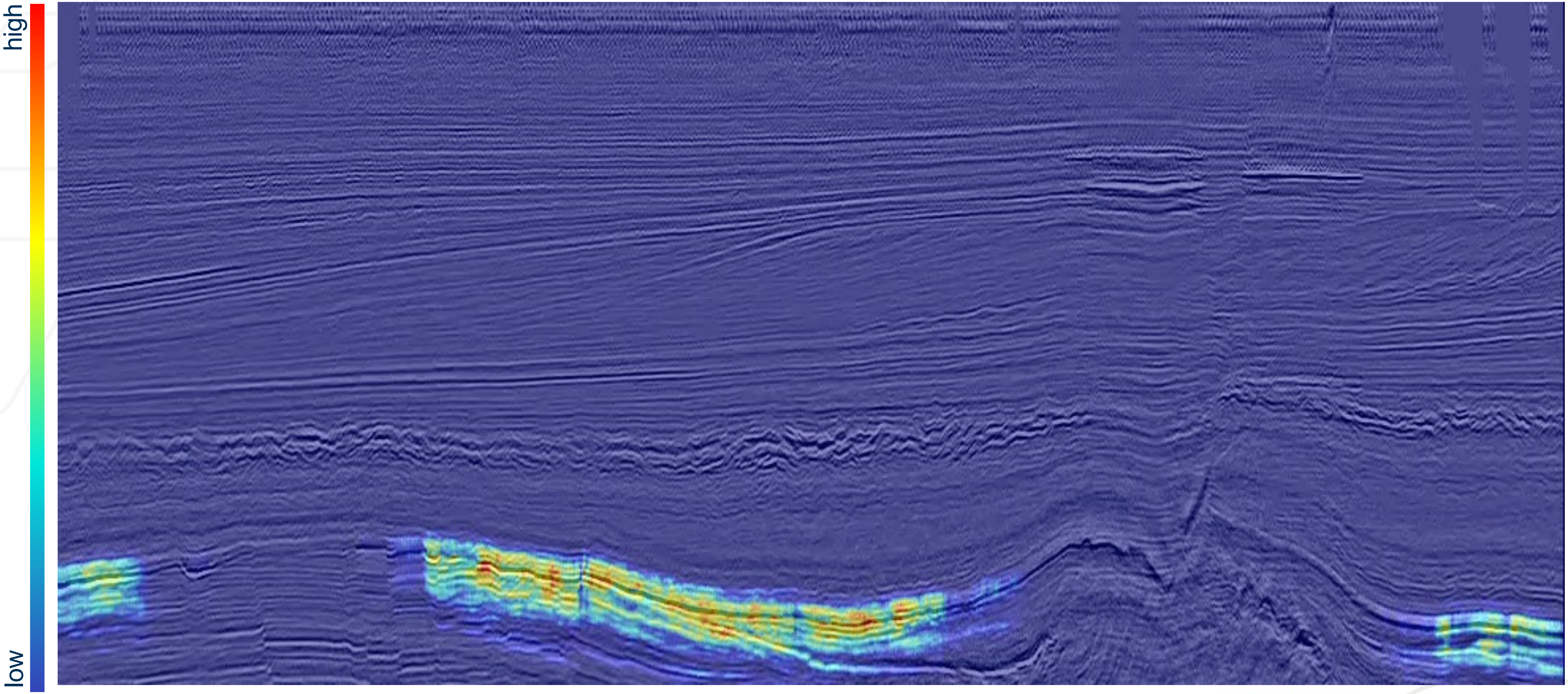
Results for Chaotic Regions



Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

Results for Fault Regions

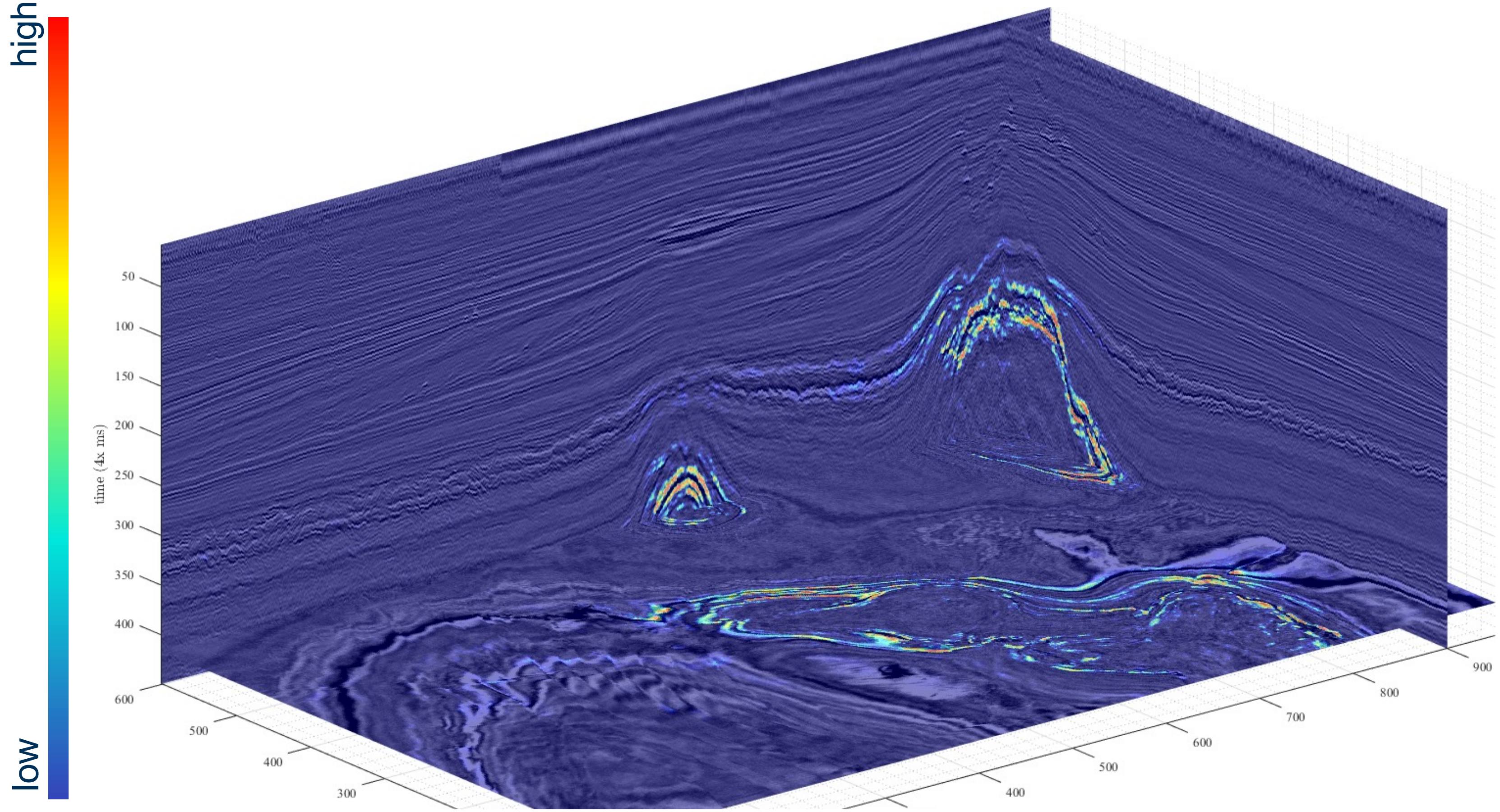


Y. Alaudah and G. AlRegib, "A weakly-supervised approach to seismic structure labeling," *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Case Study: Concept Retrieval in Seismic Images

Visual Explanations for computed images

3D view of computed seismic imaging explanations



Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.

Takeaways

Takeaways from Lecture 8

- Concepts are interpretable semantic features that can be represented mathematically
- They include low-level features like edges, texture and color as well as high level features including classes, and objects
  - Concept activation vectors provide a connection between the two sets of features
- Concept-based testing provides importance explanation to explanations
  - However, training concepts are not always available.
  - Moreover, the advantage of deep learning is in removing the dependence on handcrafted features. This advantage is nullified
- Given some property of concepts within data (for instance orthogonality), the network maybe trained to predict and explain the concepts in a weakly supervised fashion

References

Lecture 8: Concept Vectors: Utility in Training and Testing

- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.
- Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.
- Lee, D. D., and H. S. Seung, 1999, Learning the parts of objects by non-negative matrix factorization.: Nature, 401, 788–91
- Y. Alaudah and G. AlRegib, “A weakly-supervised approach to seismic structure labeling,” *87th Annual SEG Meeting Extended Abstracts*, Houston, Texas, 2017.
- M. Prabhushankar, G. Kwon, D. Temel and G. AlRegib, ”Semantically Interpretable and Controllable Filter Sets,” 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 1053-1057.
- M. Prabhushankar, D. Temel and G. AlRegib, ”Generating adaptive and robust filter sets using an unsupervised learning framework,” 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 3041-3045.
- D. Temel, M. Prabhushankar, and G. AlRegib, “UNIQUE: Unsupervised Image Quality Estimation,” the *IEEE Signal Processing Letters*, vol.23, no.10, pp.1414-1418.
- Prabhushankar, Mohit, Dogancan Temel, and Ghassan AlRegib. "Ms-unique: Multi-model and sharpness-weighted unsupervised image quality estimation." *arXiv preprint arXiv:1811.08947* (2018).
- M. A. Shafiq, M. Prabhushankar, H. Di, and G. AlRegib, “Towards Understanding Common Features Between Natural and Seismic Images,” *Expanded Abstracts of the SEG Annual Meeting*, Anaheim, CA, Oct. 14-19, 2018.