JOHANNES KEPLER
UNIVERSITÄT LINZ

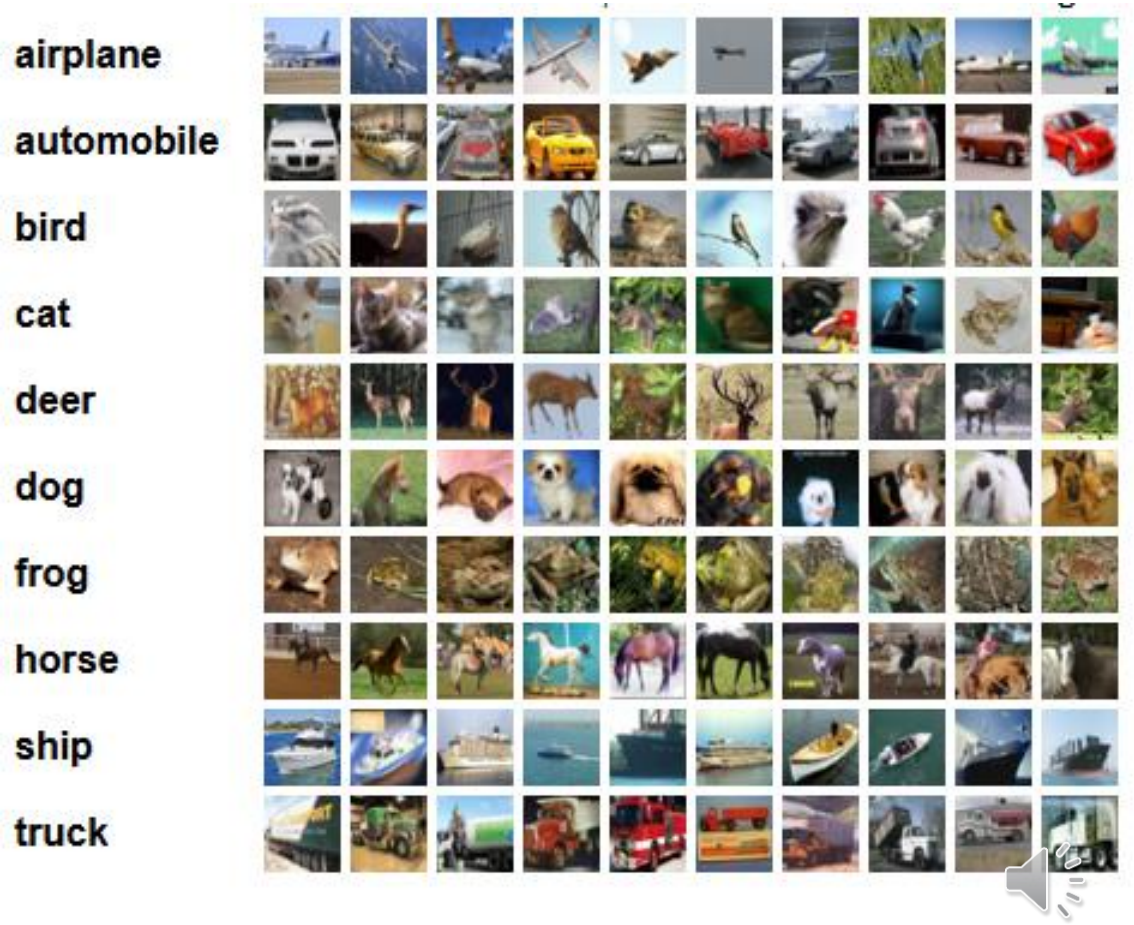
Assignment 2: Model Explanations

Group: Sigma X
Giovanni Filomeno, Moritz Riedl, Verena Szojak & Aaron Zettler

JOHANNES KEPLER
UNIVERSITÄT LINZ
Altenberger Straße 69
4040 Linz, Österreich
jku.at

CIFAR-10 Dataset.

- 32x32 RGB images
- 10 classes
- 6000 images per class
- 50000 train & 10000 test images

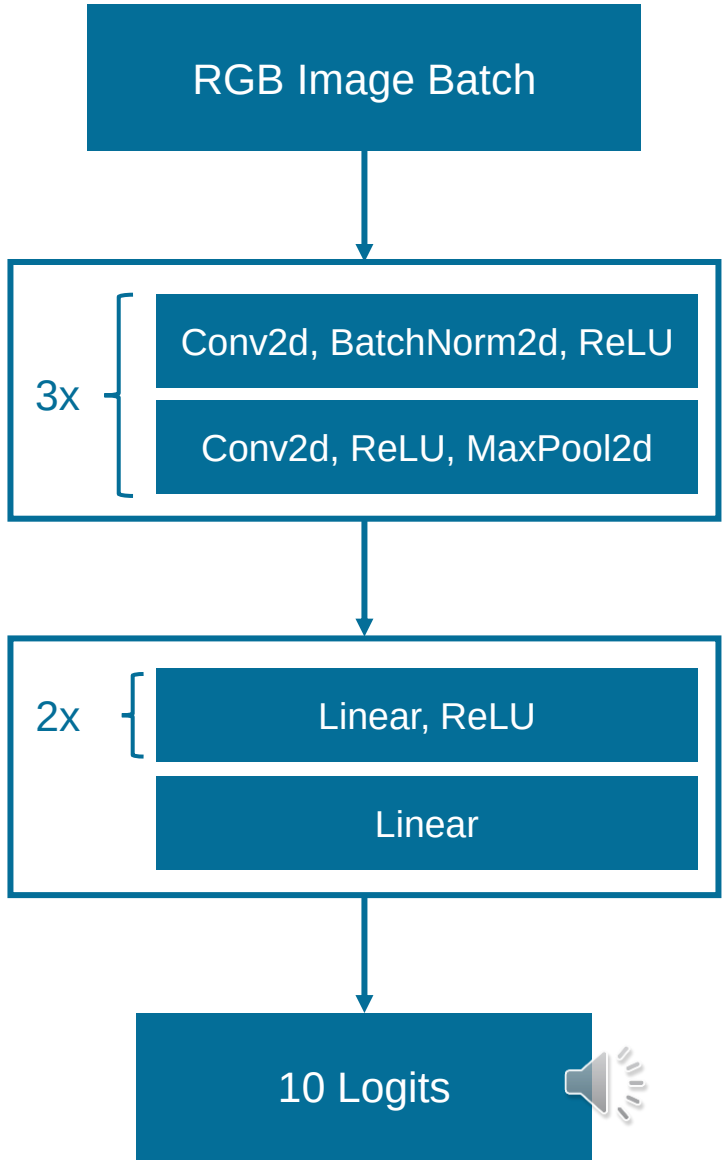


Source:https://www.cs.toronto.edu/~kriz/cifar.html

CNN Model.

- 6 convolutional layers
- 3 linear layers
- ReLU activation function

- Trained for 120 epochs
- Overall test accuracy: 82.59%
  - Highest for means of transportation
    - Automobile 92.00%
  - Lowest for animals
    - Cat 64.30%



Research Questions.

- What image parts are important for classification?
- What patterns/structures has the model learned?
- What difficulties does the model have?

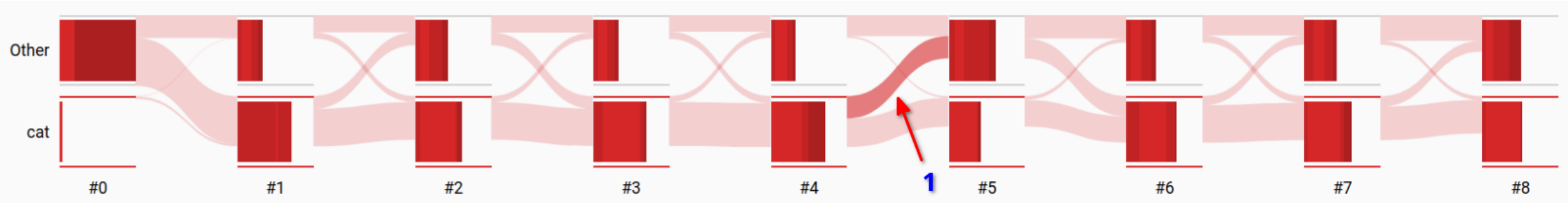
Goal: Use 4 different XAI methods to explore various angles of these questions.

XAI Methods

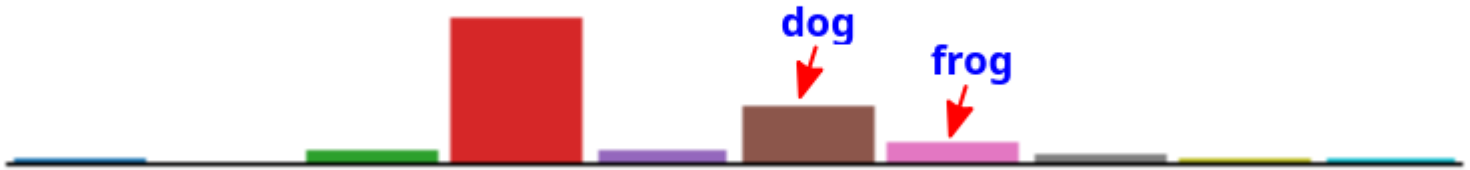
Instance Flow.

- Visualizing the flow of instances over epochs

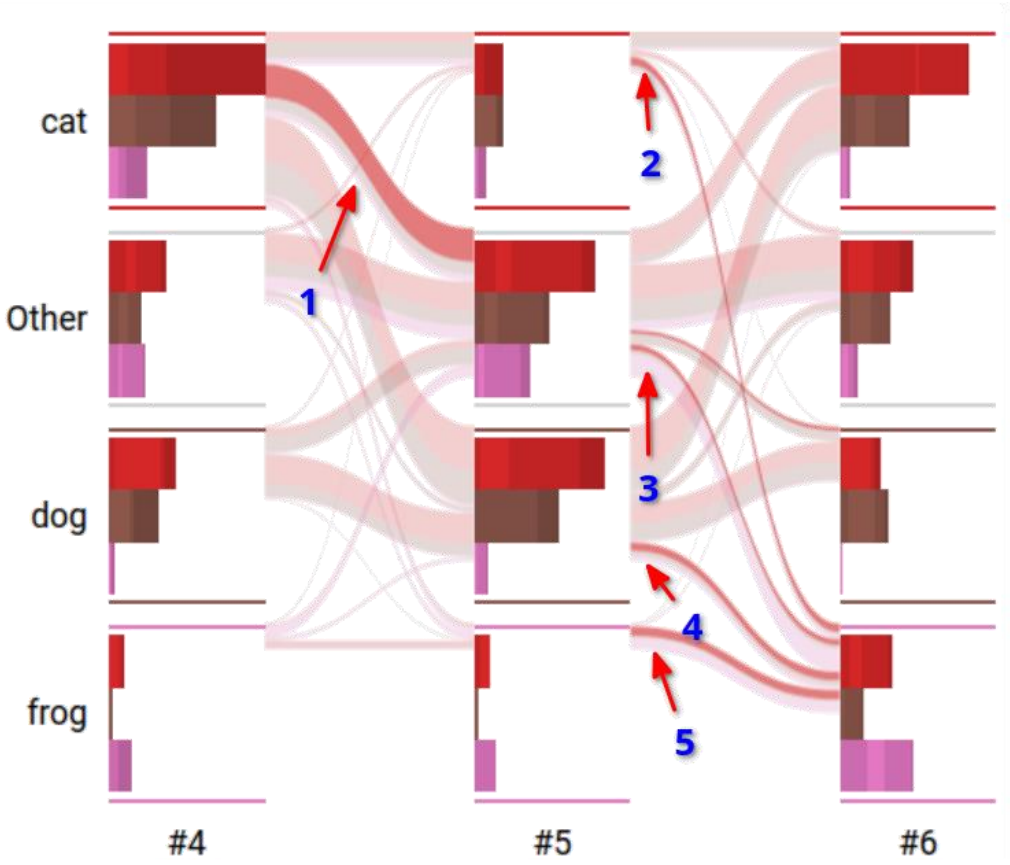
- Sankey diagram for the "cat" class for the first epochs



- Prediction distribution shows bumps for cats misclassified as dogs and frogs



Instance Flow.



- Flow of cats images after epoch #4 to dogs (1)
- Flow of cats images after epoch #5 to frogs (2-5)

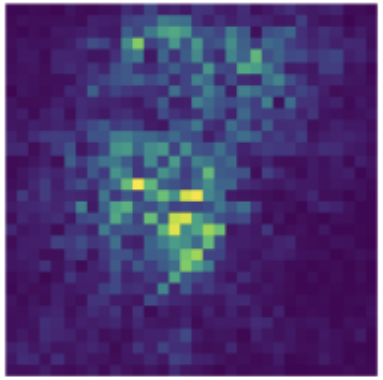
- Cat images might be ambiguous (dataset limitation)

Saliency Maps.

- Highlighting image parts important for model prediction

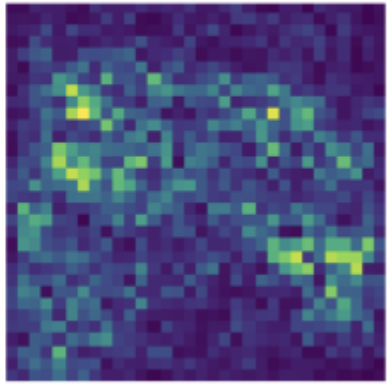
True: deer, Pred: deer

Saliency Map



True: airplane, Pred: airplane

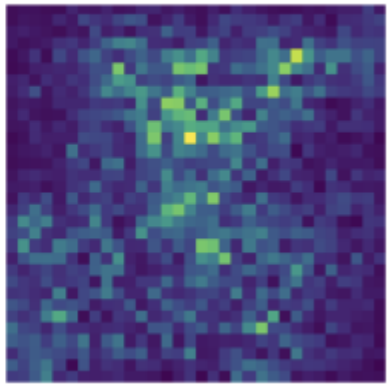
Saliency Map



Saliency Maps.

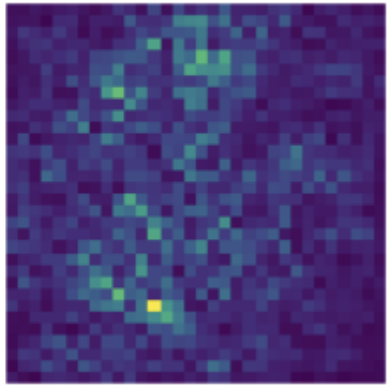
True: dog, Pred: cat

Saliency Map



True: frog, Pred: cat

Saliency Map



- Model focuses on objects itself, not on surroundings
- Attention on color than on shapes

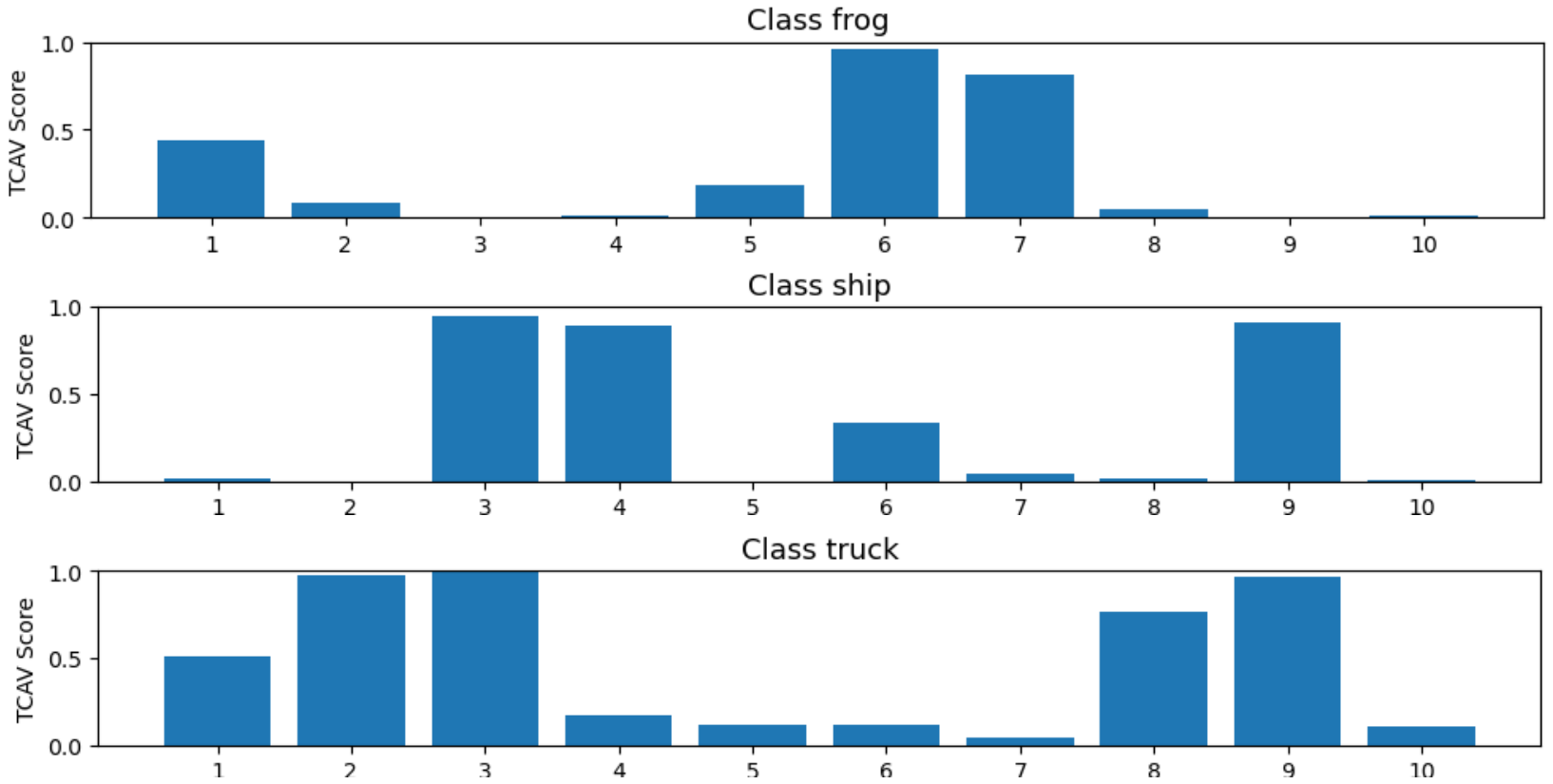
SHAP - SHapley Additive exPlanations.

- Highlighting pixel-level contributions for each prediction

- Specific object parts contribute positively to prediction
- Surrounding also important

Invertible Concept-based Explanations.

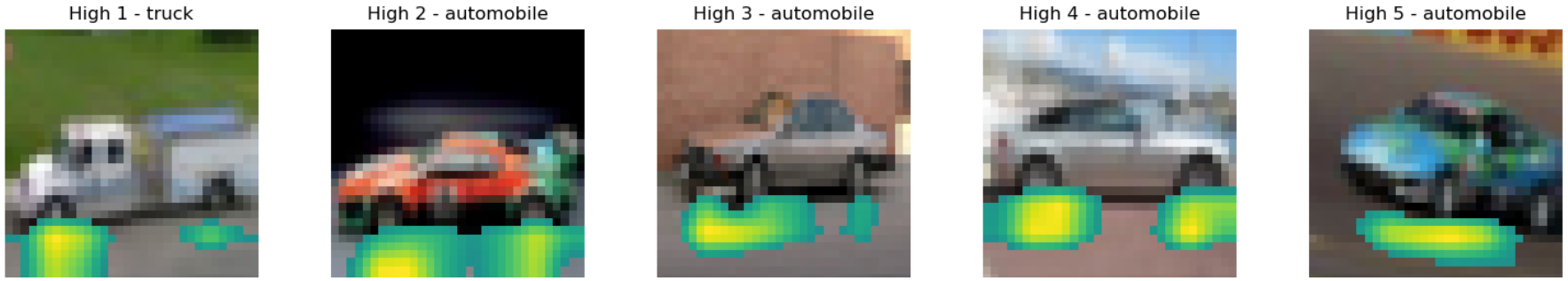
- Using activation maps for unsupervised extraction of concepts (CAVs)



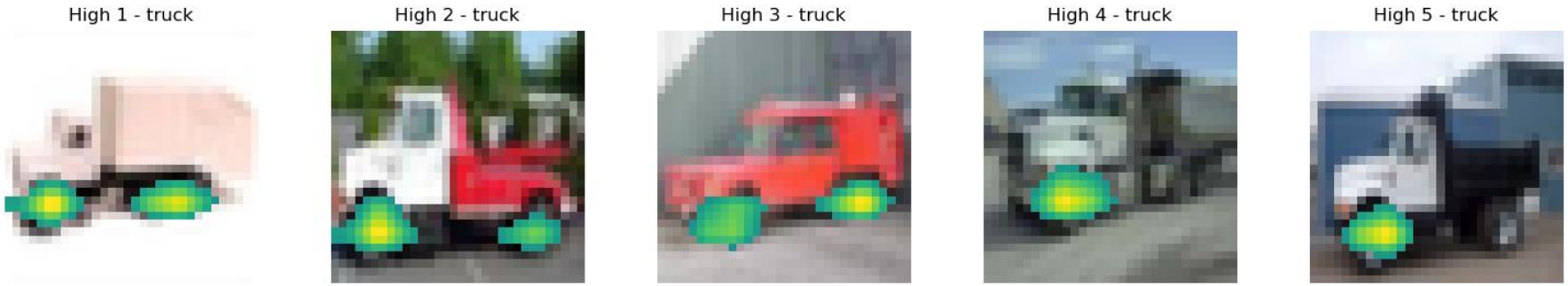
Invertible Concept-based Explanations.

- Model learned concepts for locations, shapes, colors
- Fine-grained concepts for higher-accuracy classes

Images for Concept 1

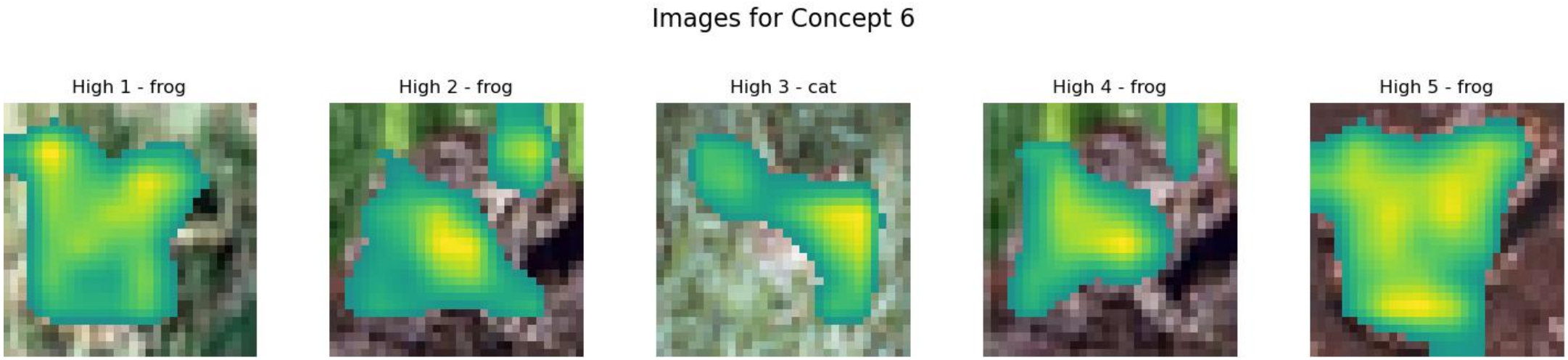


Images for Concept 2



Invertible Concept-based Explanations.

- Concepts too coarse for lower-accuracy classes
- Model has not learned discriminative concepts



Findings

Answers to Research Questions.

- What image parts are important for classification?
  - Object itself & its surroundings
- What patterns/structures has the model learned?
  - Mostly colors, also shapes
- What difficulties does the model have?
  - Ambiguous images lead to confusion (cats)
  - Learned too simple concepts

Questions