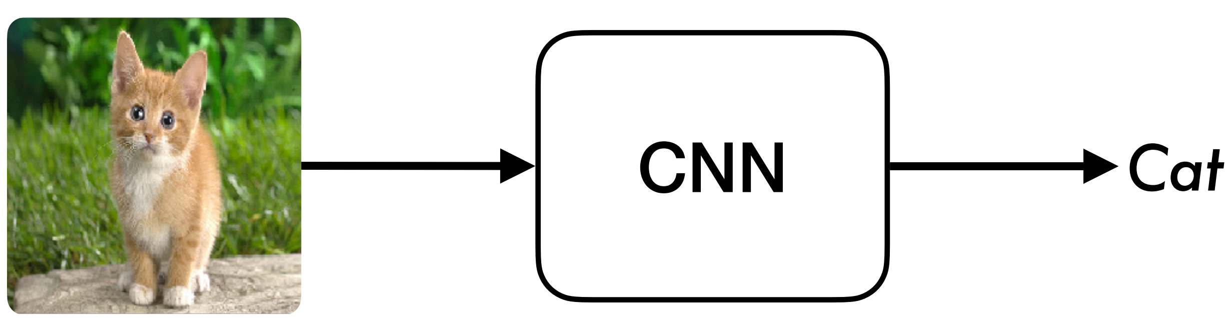
Now You See Me (CME): Concept-based Model Extraction

Dmitry Kazhdan*, Botty Dimanov*, Mateja Jamnik*, Pietro Liò*, Adrian Weller*^
* The University of Cambridge
^ The Alan Turing Institute

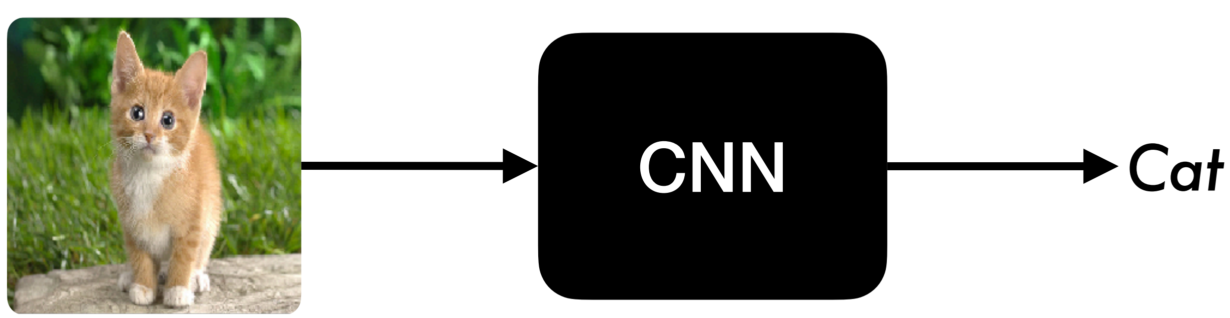
Feature Importance Explanations

- CNNs **applied to many tasks** (e.g. facial recognition, object recognition, VQA...)



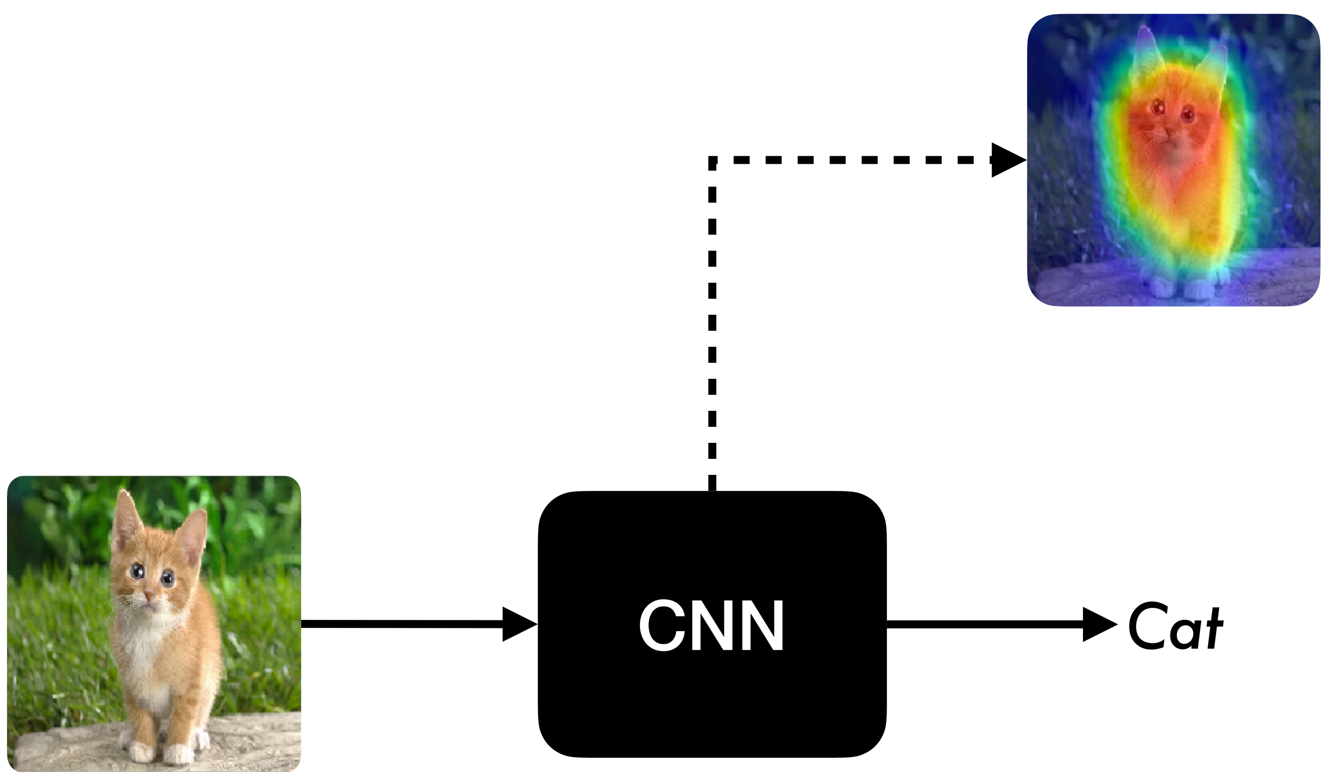
Feature Importance Explanations

- CNNs **applied to many tasks** (e.g. facial recognition, object recognition, VQA…)
- Unfortunately, CNNs are **black-boxes**
- Lots of interest in **XAI**



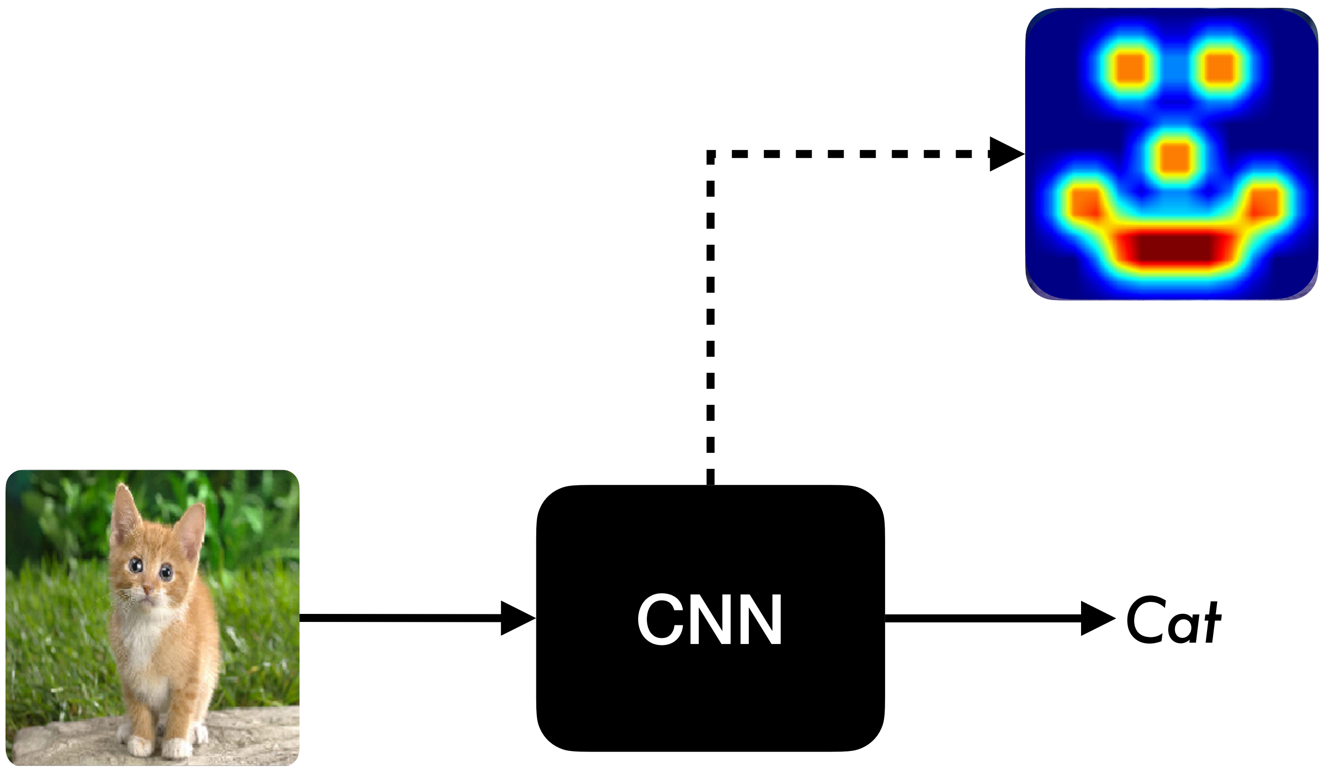
Feature Importance Explanations

- CNNs **applied to many tasks** (e.g. facial recognition, object recognition, VQA…)
- Unfortunately, CNNs are **black-boxes**
- Lots of interest in **XAI**
- Existing approaches are based on **feature importance** (aka heat maps)



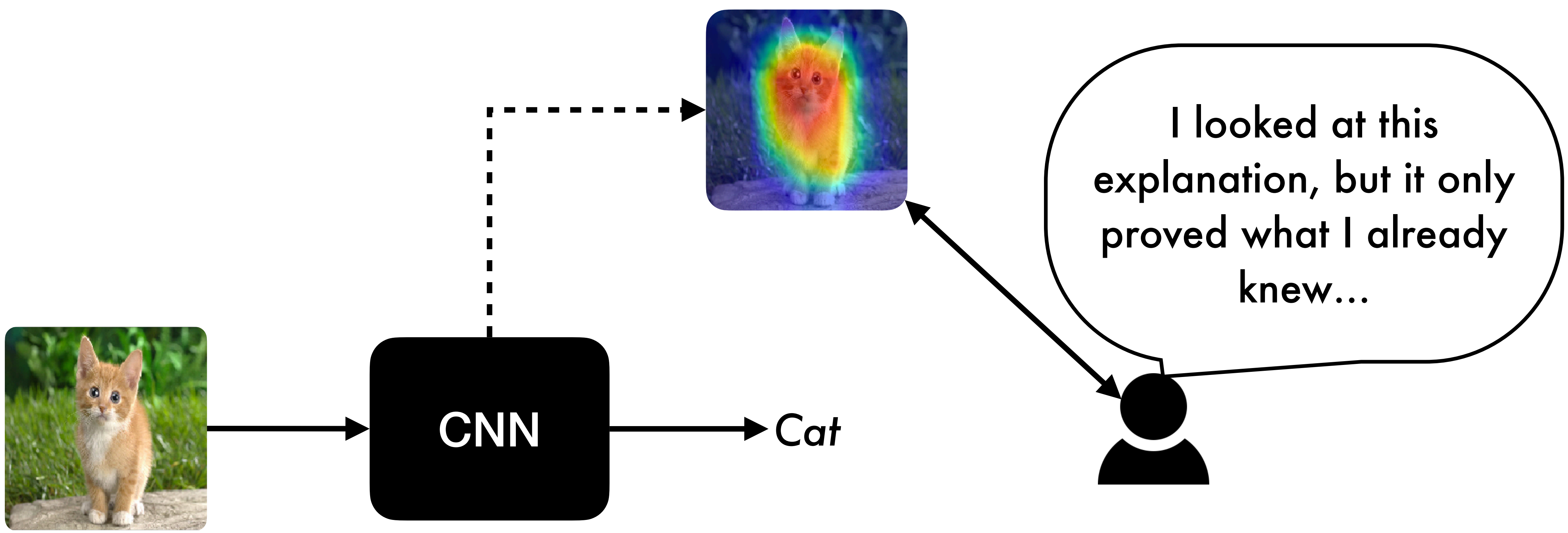
Feature Importance Explanations

- CNNs **applied to many tasks** (e.g. facial recognition, object recognition, VQA…)
- Unfortunately, CNNs are **black-boxes**
- Lots of interest in **XAI**
- Existing approaches are based on **feature importance** (aka heat maps)
- However, shown to be **fragile** and susceptible to **confirmation bias**



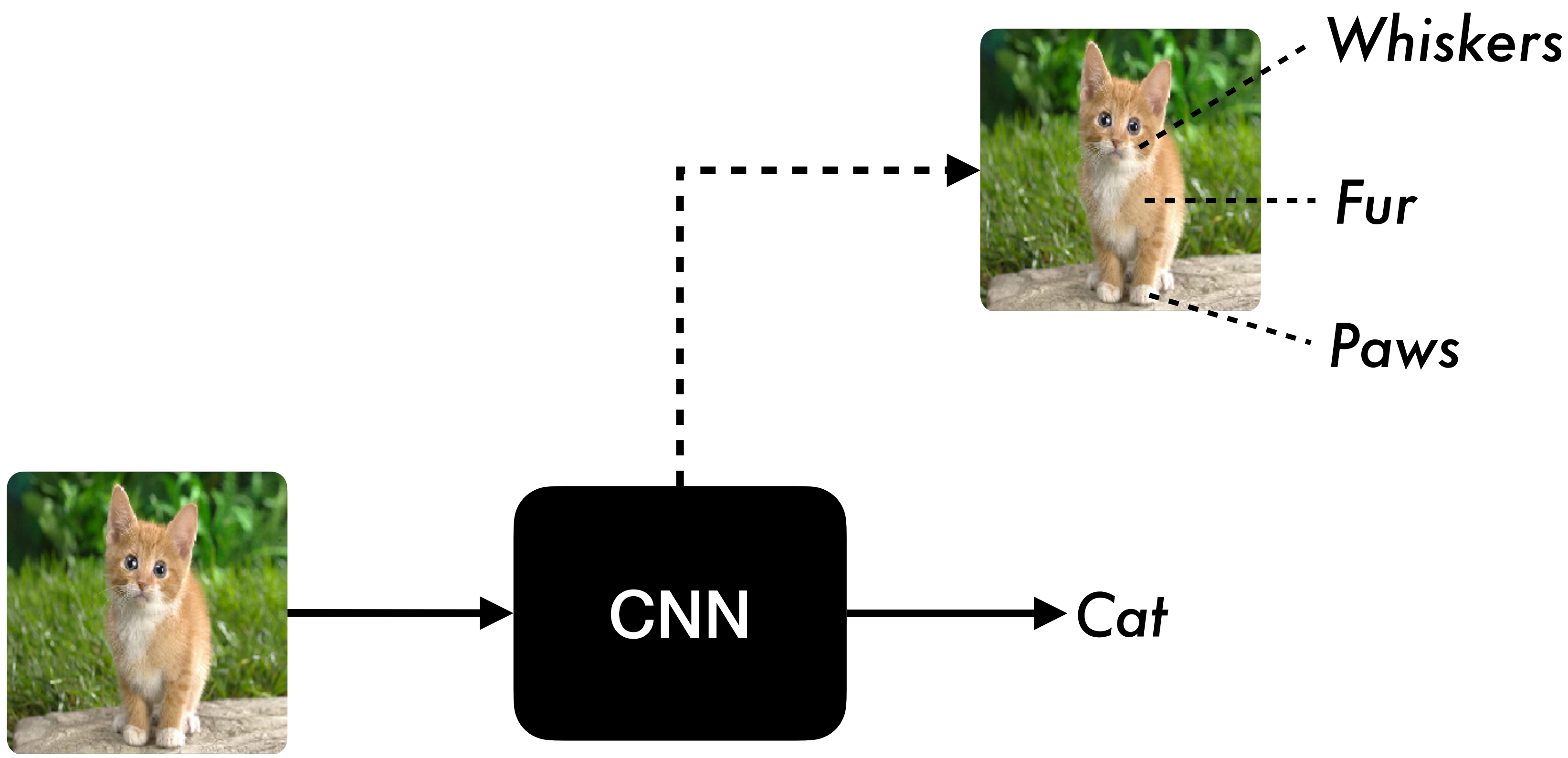
Feature Importance Explanations

- CNNs **applied to many tasks** (e.g. facial recognition, object recognition, VQA...)
- Unfortunately, CNNs are **black-boxes**
- Lots of interest in **XAI**
- Existing approaches are based on **feature importance** (aka heat maps)
- However, shown to be **fragile** and susceptible to **confirmation bias**



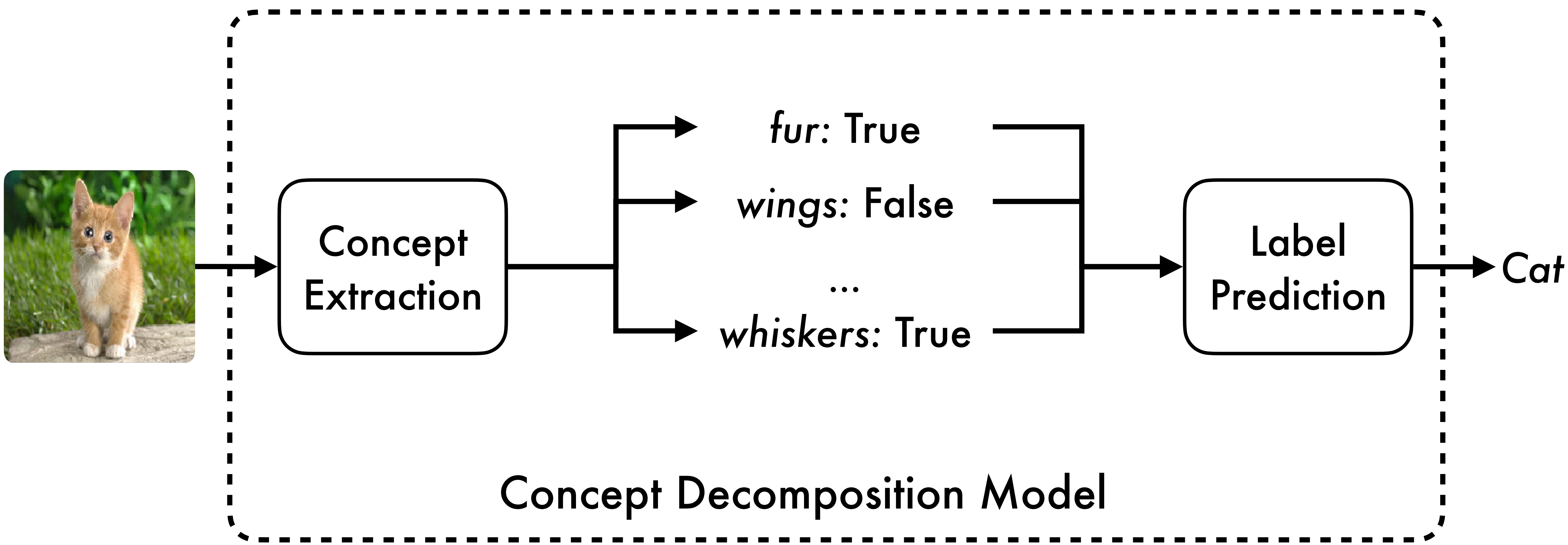
Concept-based Explanations

- Recent work explores **concept-based explanations**
- Explanations provided in terms of high-level concepts (aka attributes)



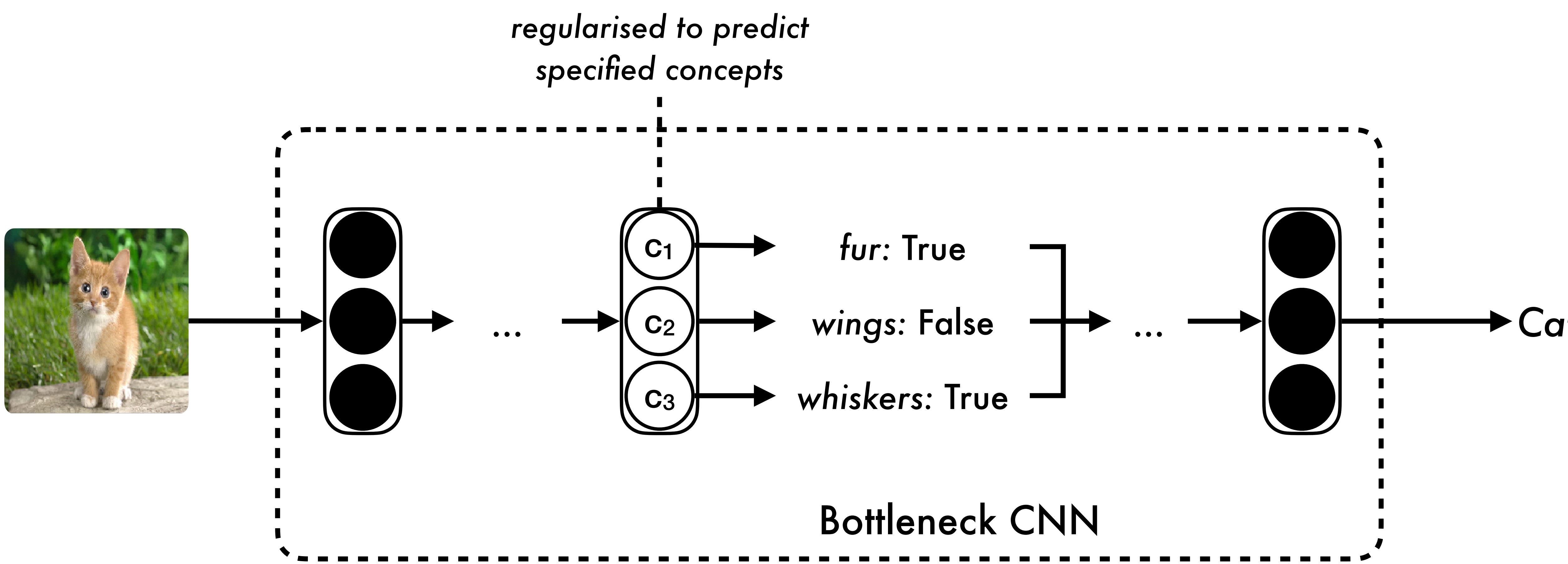
Concept Decomposition

- We introduce notion of **concept decomposition**
- New type of concept-based model
- Separates model processing into:
  - **Concept extraction**: predicting concept information from input
  - **Label prediction**: predicting class labels from concept information
- Concept-decompositional models process inputs hierarchically



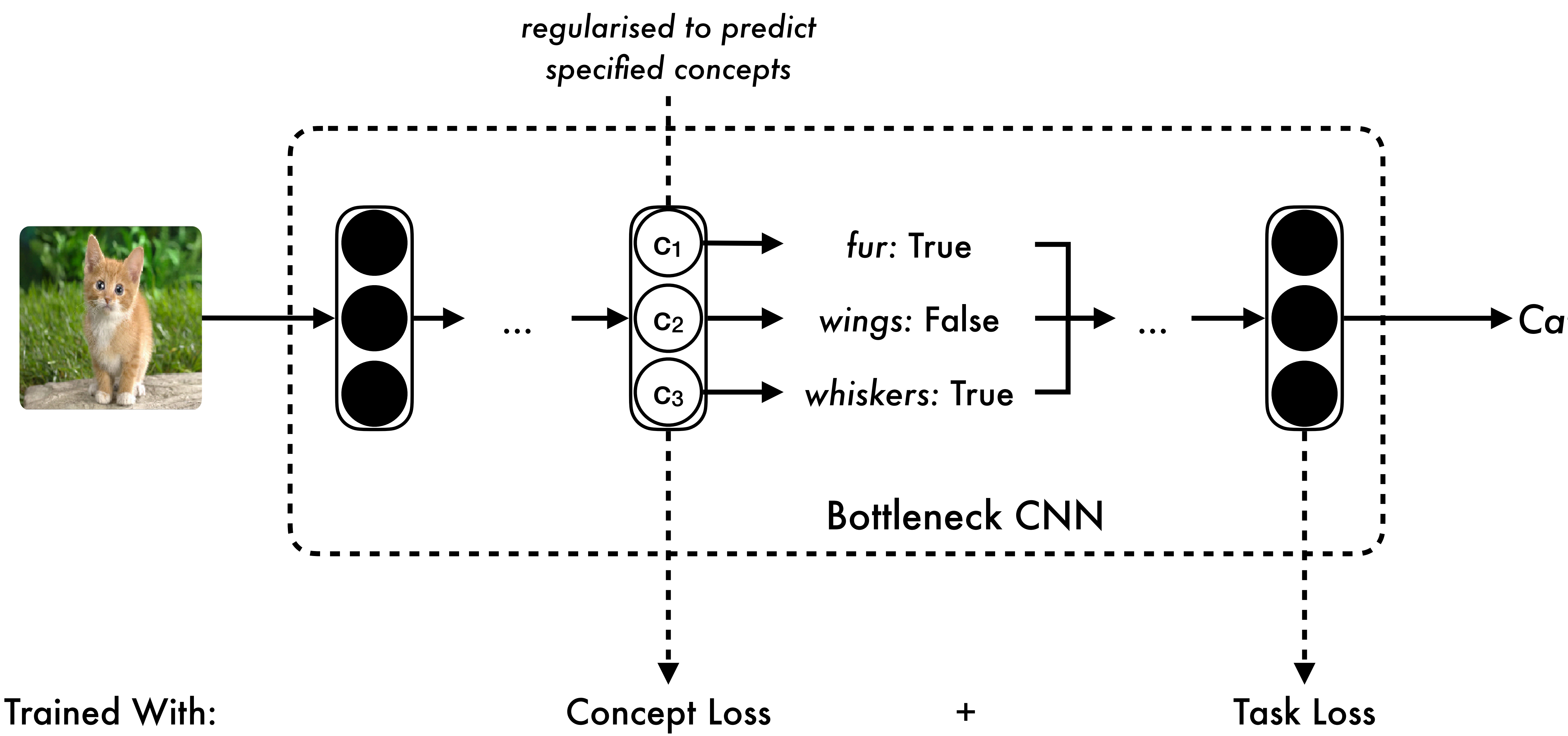
Concept Bottleneck Models

- Assume you have concept labels for every input sample
- Create a CNN with a "bottleneck" layer
- Regularise bottleneck during training, ensuring it predicts provided concepts



Concept Bottleneck Models

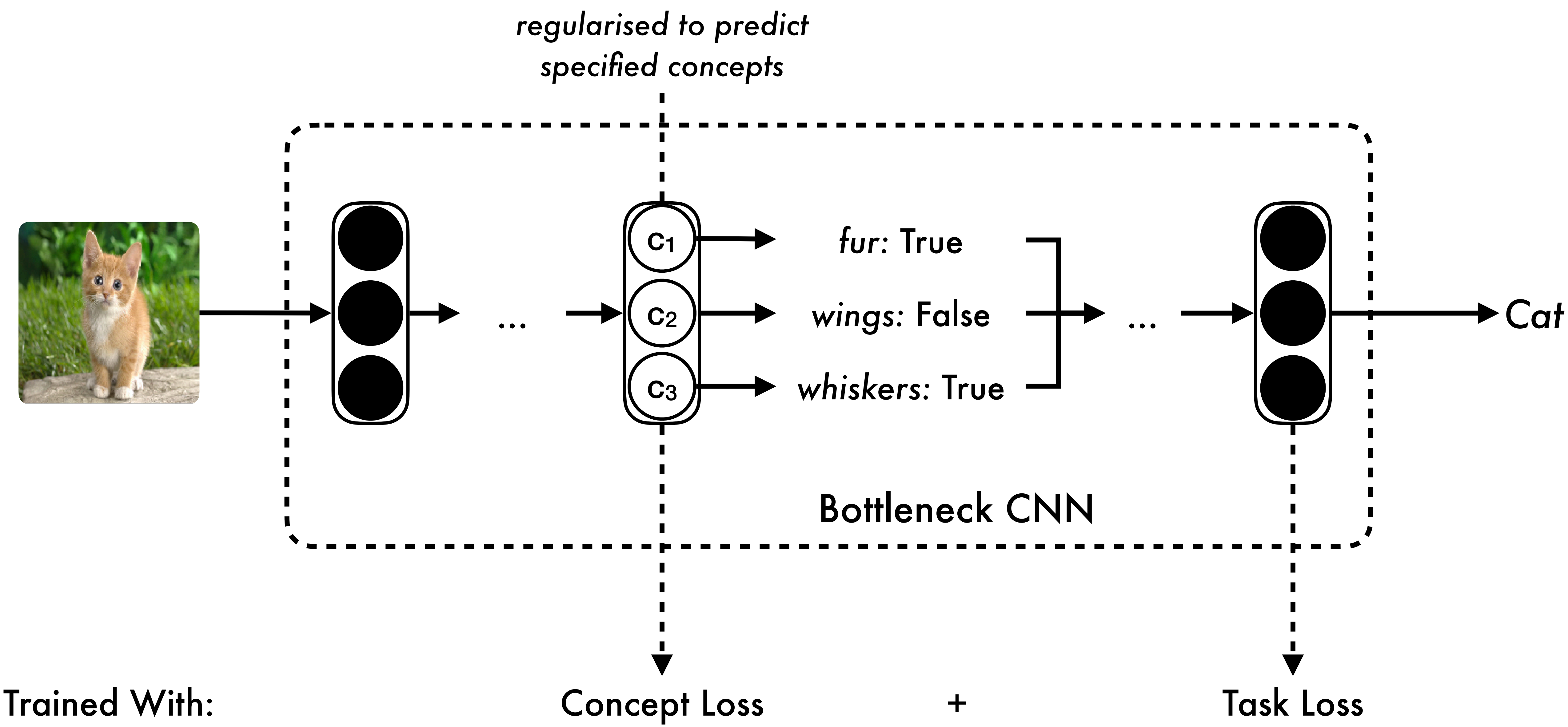
- Assume you have concept labels for every input sample
- Create a CNN with a “bottleneck” layer
- Regularise bottleneck during training, ensuring it predicts provided concepts



Concept Bottleneck Models

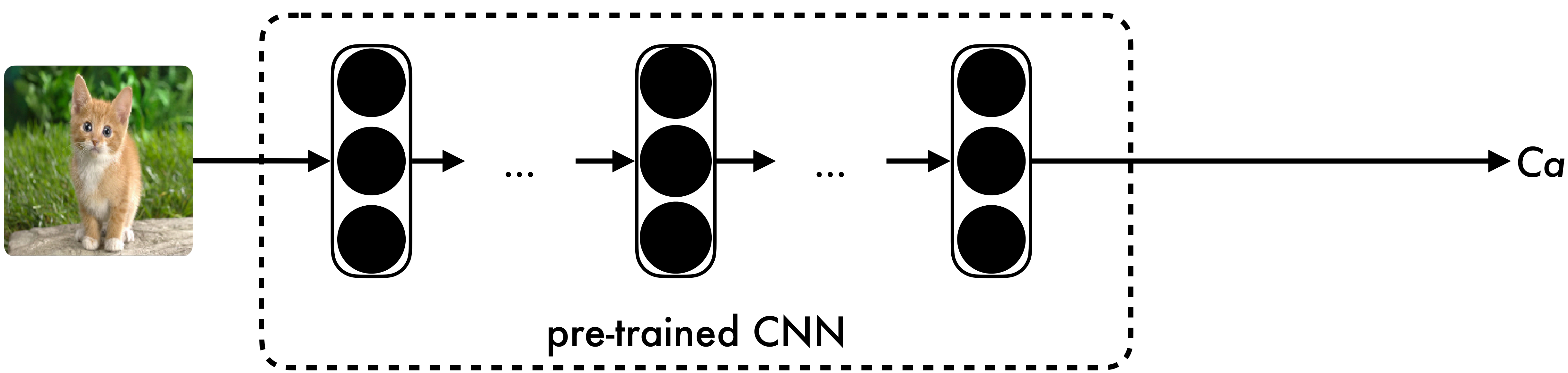
- CBMs:
  - Assumes **all** relevant concepts are known
  - Assumes **every** input point has associated concept labels available

- However:
  - Unregularised CNNs still learn the **relevant concepts**
  - Can therefore **extract** this knowledge from CNNs



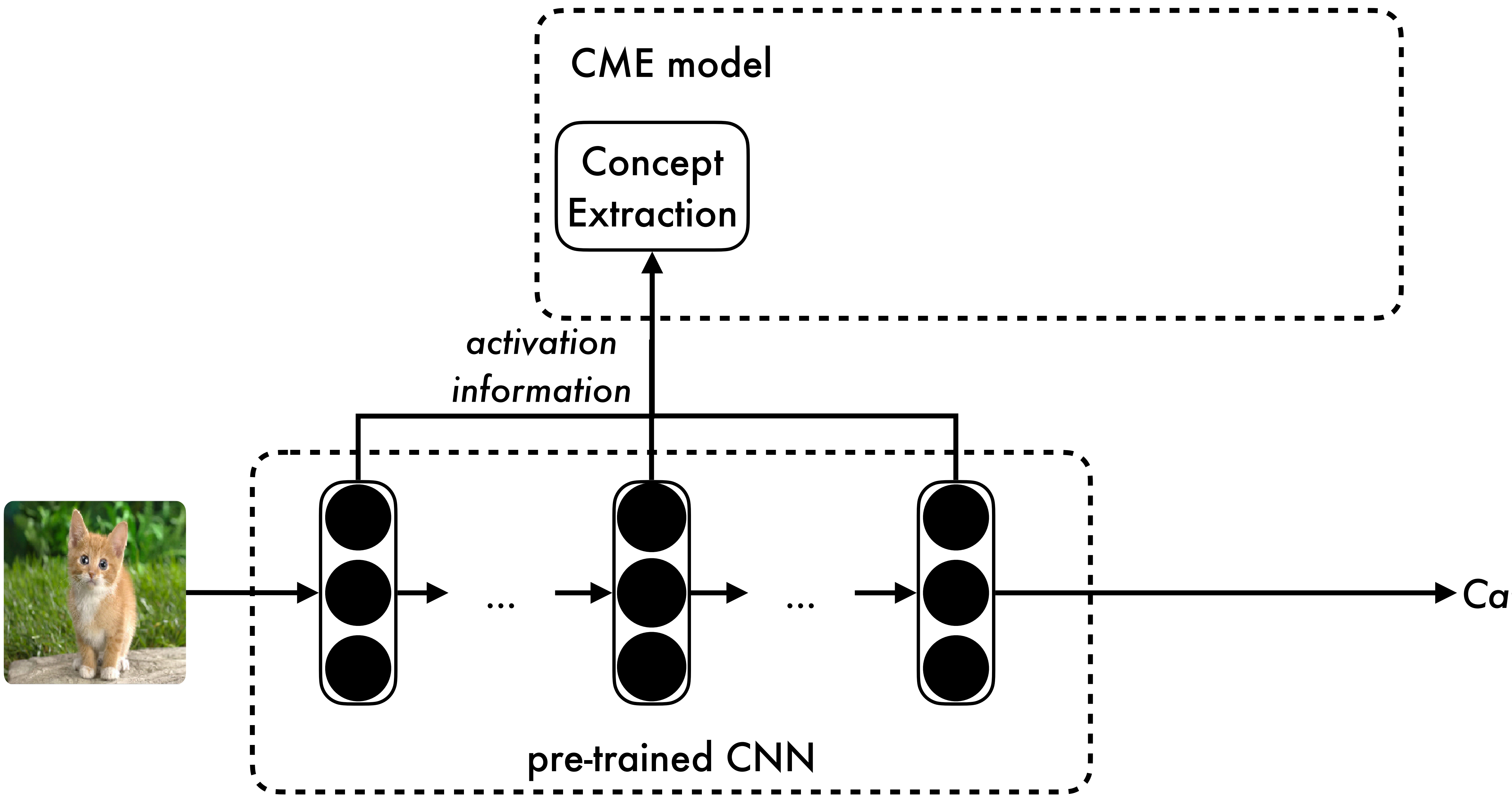
CME: (C)oncept-based (M)odel (E)xtraction

- Take a pre-trained CNN
- Train a semi-supervised concept predictor model on top of layer activations
- Train label predictor on top of concept predictor
- Leverage CNNs for performing concept information extraction automatically



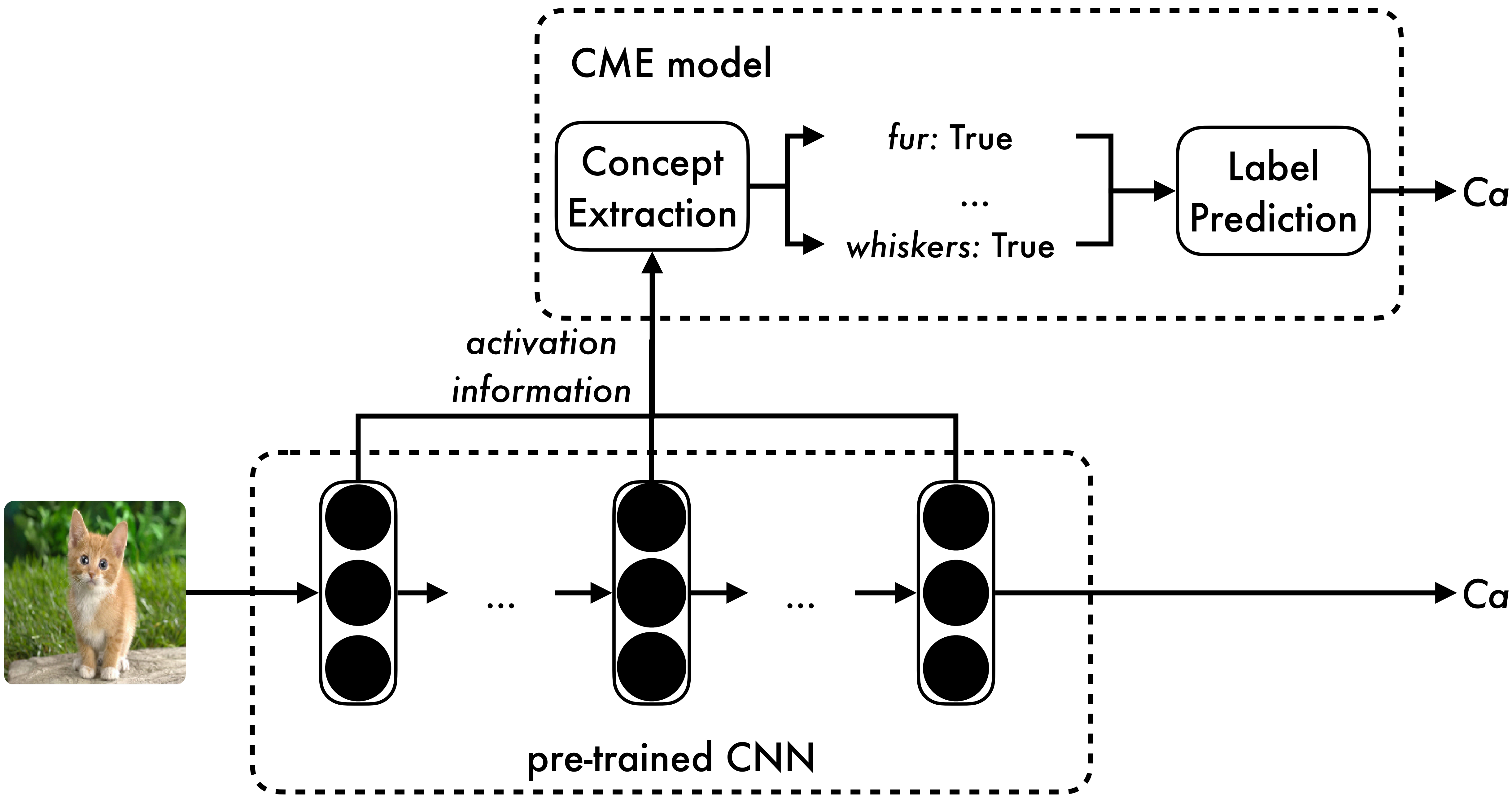
CME: (C)oncept-based (M)odel (E)xtraction

- Take a pre-trained CNN
- Train a semi-supervised concept predictor model on top of layer activations
- Train label predictor on top of concept predictor
- Leverage CNNs for performing concept information extraction automatically



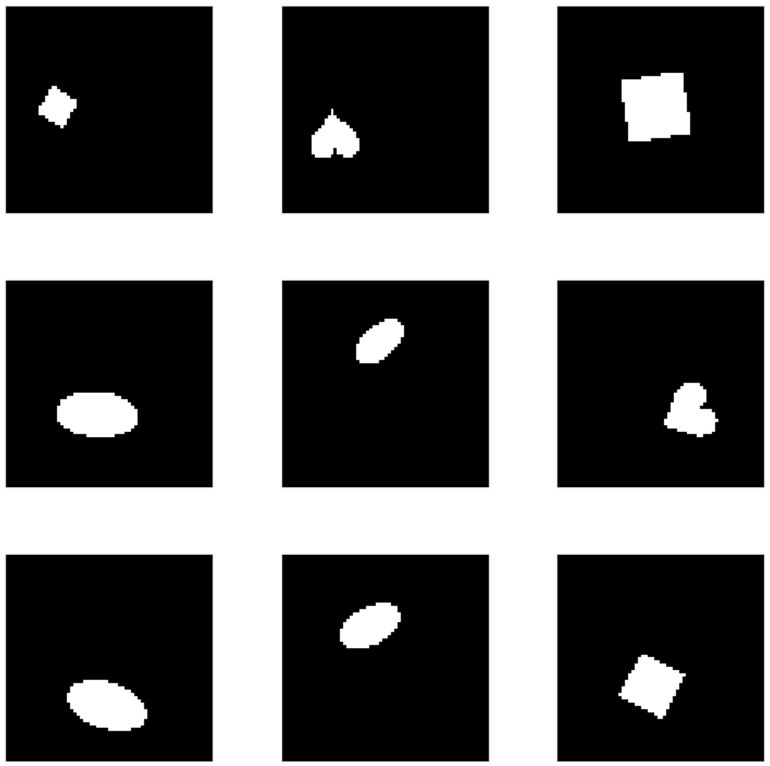
CME: (C)oncept-based (M)odel (E)xtraction

- Take a pre-trained CNN
- Train a semi-supervised concept predictor model on top of layer activations
- Train label predictor on top of concept predictor
- Leverage CNNs for performing concept information extraction automatically



dSprites

- 2D 64x64 black-and-white images
- Generated from all possible combinations of concepts:
  - Shape (square, ellipse, heart)
  - Scale (6 values linearly spaced in [0.5, 1])
  - Orientation (40 values in [0, 2 pi])
  - Position X (32 values in [0, 1])
  - Position Y (32 values in [0, 1])

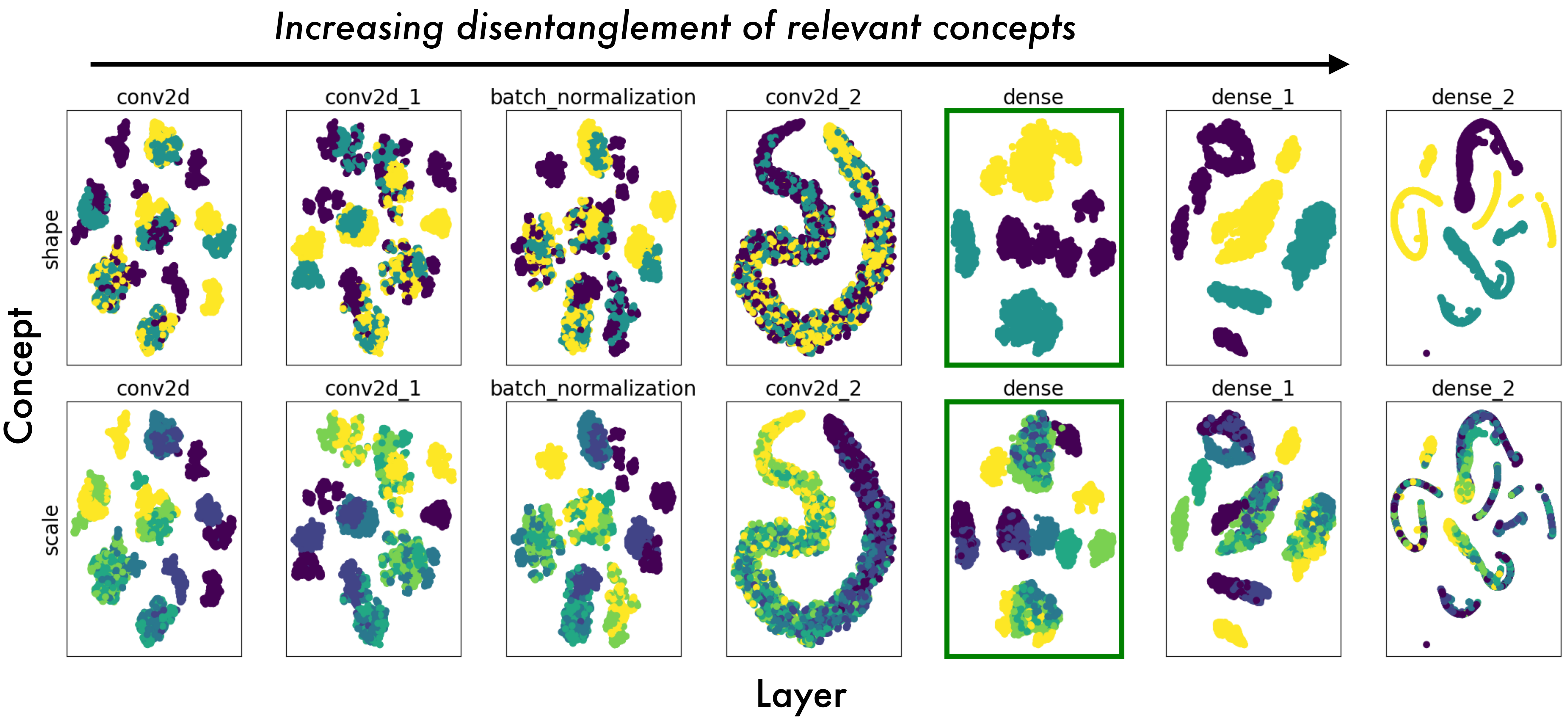


dSprites Highlights

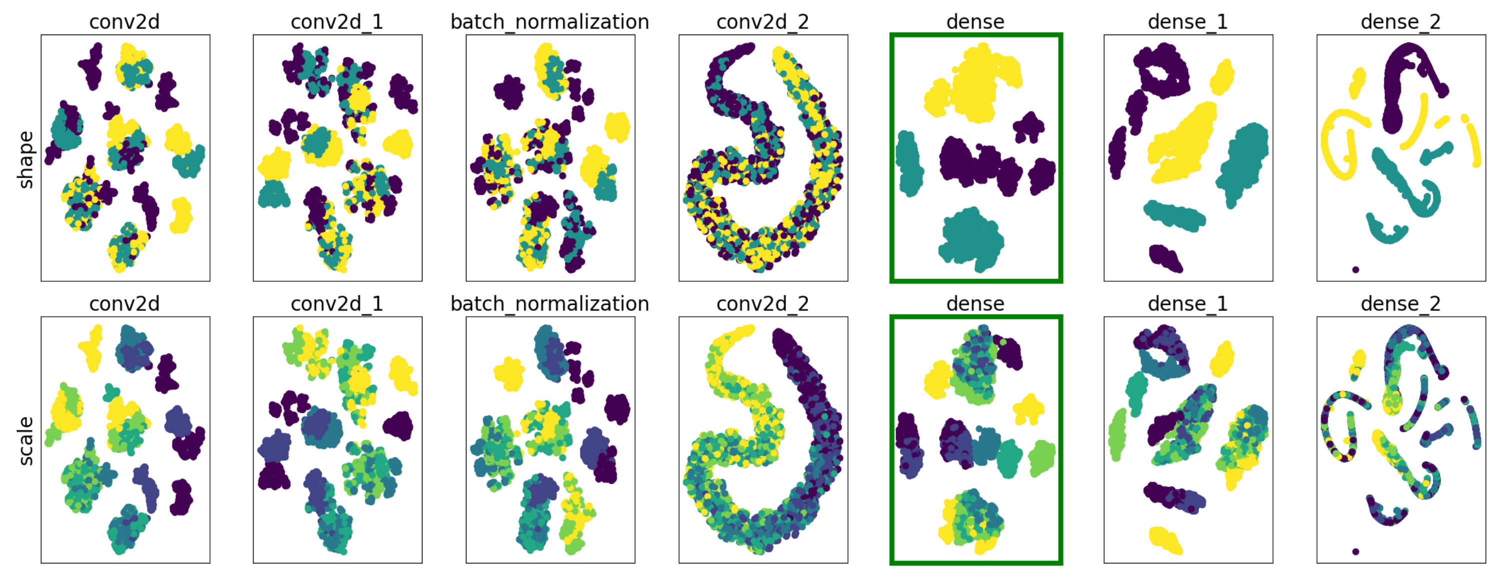
- Task: (shape, scale) ==> unique class ID
- CNN trained to predict these class IDs from images
- Benchmarked against Net2Vec for concept extraction
- Used tSNE to explore model latent space wrt concepts

dSprites Highlights

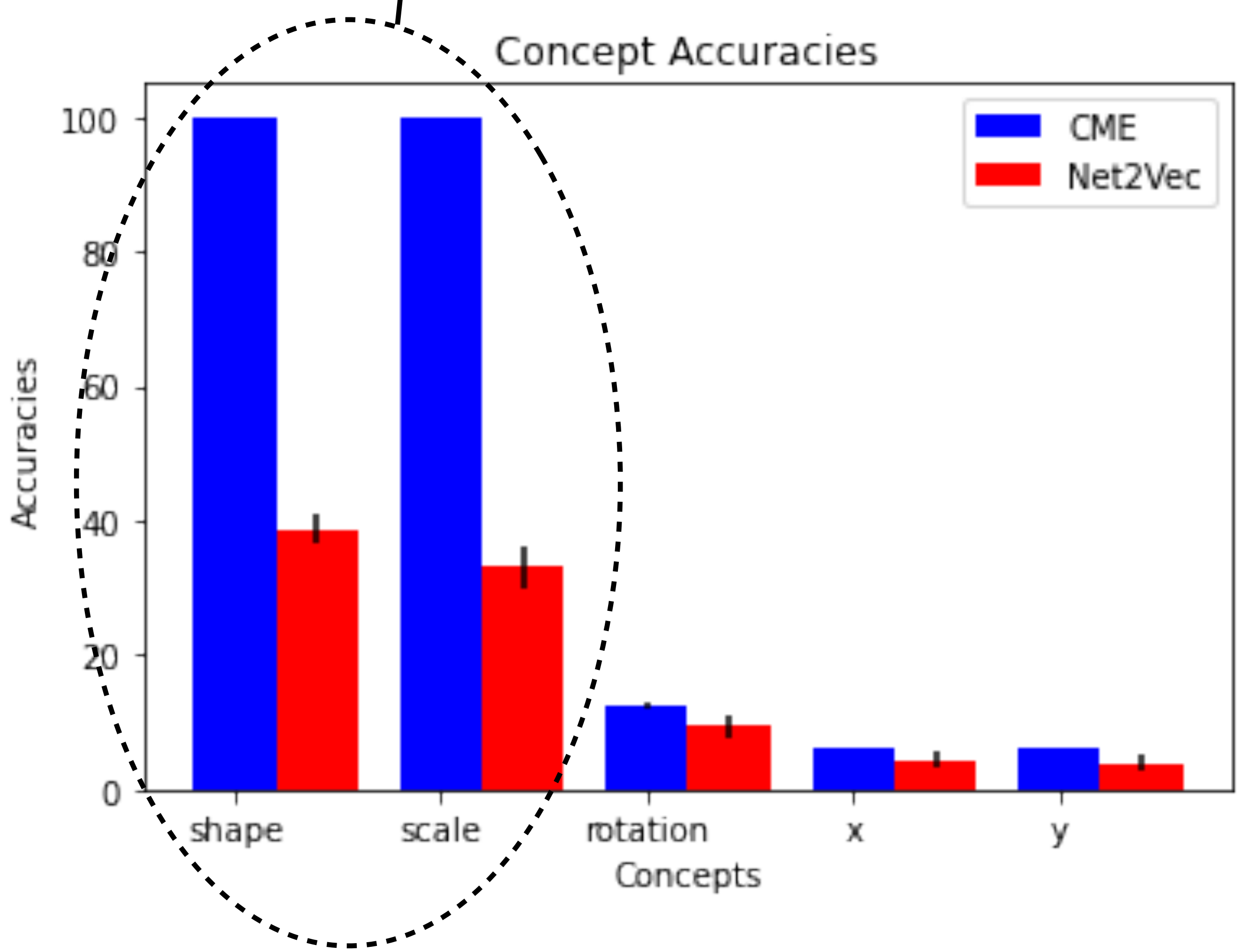
- Task: (shape, scale) ==> unique class ID
- CNN trained to predict these class IDs from images
- Benchmarked against Net2Vec for concept extraction
- Used tSNE to explore model latent space wrt concepts



dSprites Highlights



Relevant concepts (shape & scale) predicted by CME with high accuracy



Task performances:
Original CNN: 100.0 +/- 0.3%
CME model: 99.3 +/- 0.5%

Only required 100 concept-labelled samples

Caltech-UCSD Birds

- 11,788 images of 200 bird species
- 112 binary concepts, such as:
  - Beak colour
  - Wing colour
  - Beak shape
  - etc…
- Task: Predicting the correct bird species
- Compared CME with CBM approaches
- Demonstrated how CME can be used to filter out irrelevant concepts
- See paper for more details

Future Directions

- Human-in-the-Loop extensions:
  - CME: can't fine-tune/correct the model
  - Explore interactive methodologies for extracting *and* injecting concept information

- Further applications:
  - In imaging tasks, “concepts” are often not rigorously-defined
  - In other areas (e.g. physics, or drug discovery), there are tasks with more well-defined domain-specific concepts

Conclusions

- Concept-based explanations gaining traction
- Concept Decomposition (CD): new type of deep concept-based model
- CME leverages power/knowledge of pre-trained CNNs to extract CD models
- Showcased results
- Discussed future work
- Link: http://ceur-ws.org/Vol-2699/paper02.pdf