Explaining Deep Convolutional Neural Networks via Latent Visual-Semantic Filter Attention

Yu Yang[1], Seungbae Kim[2], Jungseock Joo[2]
[1]Department of Computer Science, [2]Department of Communication, University of California, Los Angeles

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

WHY do we need to explain deep CNN?

Linear Regression

FEATURE #1 "Location"
FEATURE #2 "Income"
FEATURE #3 "# Bedrooms"
→ OUTPUT House Value

Neural Networks

FEATURE #1: ?
FEATURE #2: ?
FEATURE #3: ?
FEATURE #4: ?

NOT
*Neural network is all we need! **We also need explanations!***
Fair? *Reliable?*

Explicitly predefined features
✓ Interpretable!

Implicitly learned features
× Black-box!

A **Trained** Deep CNN
Inputs
Outputs of Convolutional Filters (Output Channels)

"If we can know the features learned by each convolutional filter, ..."

FILTER #1 "Face"
FILTER #2 "Suit"
FILTER #3 "Hair"
FILTER #4 "Pet"

"...then the model becomes interpretable! "

OUTPUT Gender

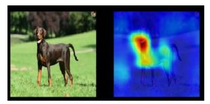
LaViSE -- Latent Visual-Semantic Explainer for CNN

A **Trained** Deep CNN → LaViSE (Our method) → LaViSE output

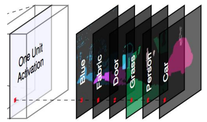
"What concept(s) did this filter learn to identify?"

FILTER #4:
dog, cat, pet, ...

EXISTING APPROACHES



☹ Only show important pixels...
No explicit semantic explanations!



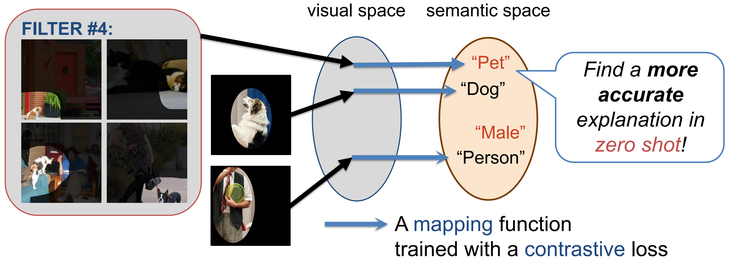
☹ Only use annotated concepts...
Cannot explain unseen concepts!

LaViSE
☺ Provides **visual+semantic** explanations!
☺ Finds the best explanation from **annotated+zero-shot** concepts!

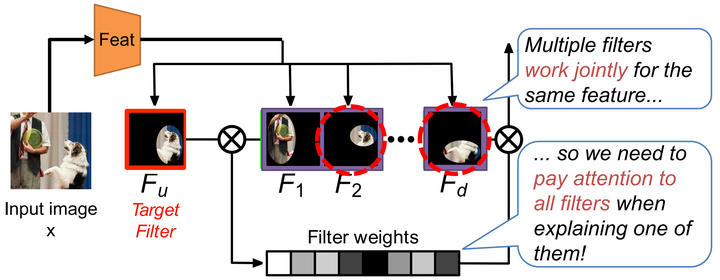
*Plus, the explanations are **more acurate with the filter attention**!*

HOW does LaViSE work?

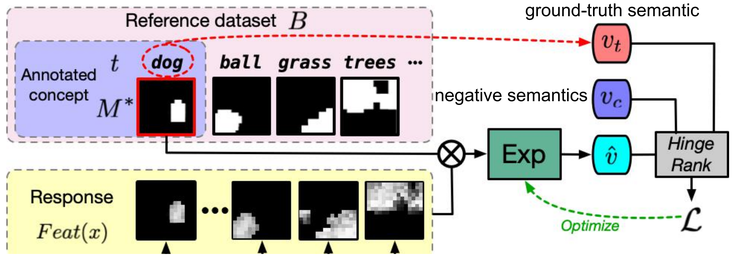
Key component #1: Latent Visual-Semantic Mapping



Key component #2: Filter-level Attention



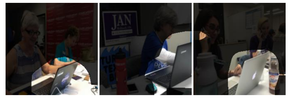
Training LaViSE



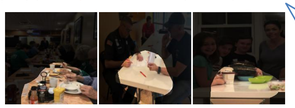
WHAT can LaViSE do?

➢ Explaining Trained Networks with Zero-shot Concepts

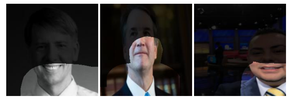

uniform, shirt, suit, jersey, body


laptop, computer, screen, desk, desktop


car, driver, vehicle, bus, truck


table, plate, tray, kitchen, dish

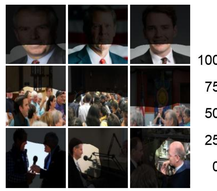
Words in red are *zero-shot* concepts!

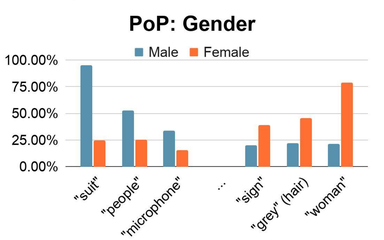

shirt, blazer, hair, suit, black


shirt, team, field, player, letters

➢ Explaining Unlabeled Datasets with Pretrained Networks

LaViSE can *provide insights* of a dataset and help with *bias analysis*!





Summary/Conclusion

➢ We proposed **LaViSE**, a novel framework which can both visually and semantically explain latent representations of a trained CNN.

➢ **LaViSE** enables users to discover concepts that a CNN learned without being explicitly taught.