Neural Response Interpretation Through the Lens of Critical Pathways

Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja,
Christian Rupprecht, Seong Tae Kim, Nassir Navab

CVPR VIRTUAL JUNE 19-25

At One Glance

- We identify critical pathways (sparse pathways that encode critical input information).
- We show pruning objective does not identify critical pathways
- We use critical pathways to identify critical input features (feature attribution)

Critical Pathway

Sparse pathway that encodes critical input information (for a single input)

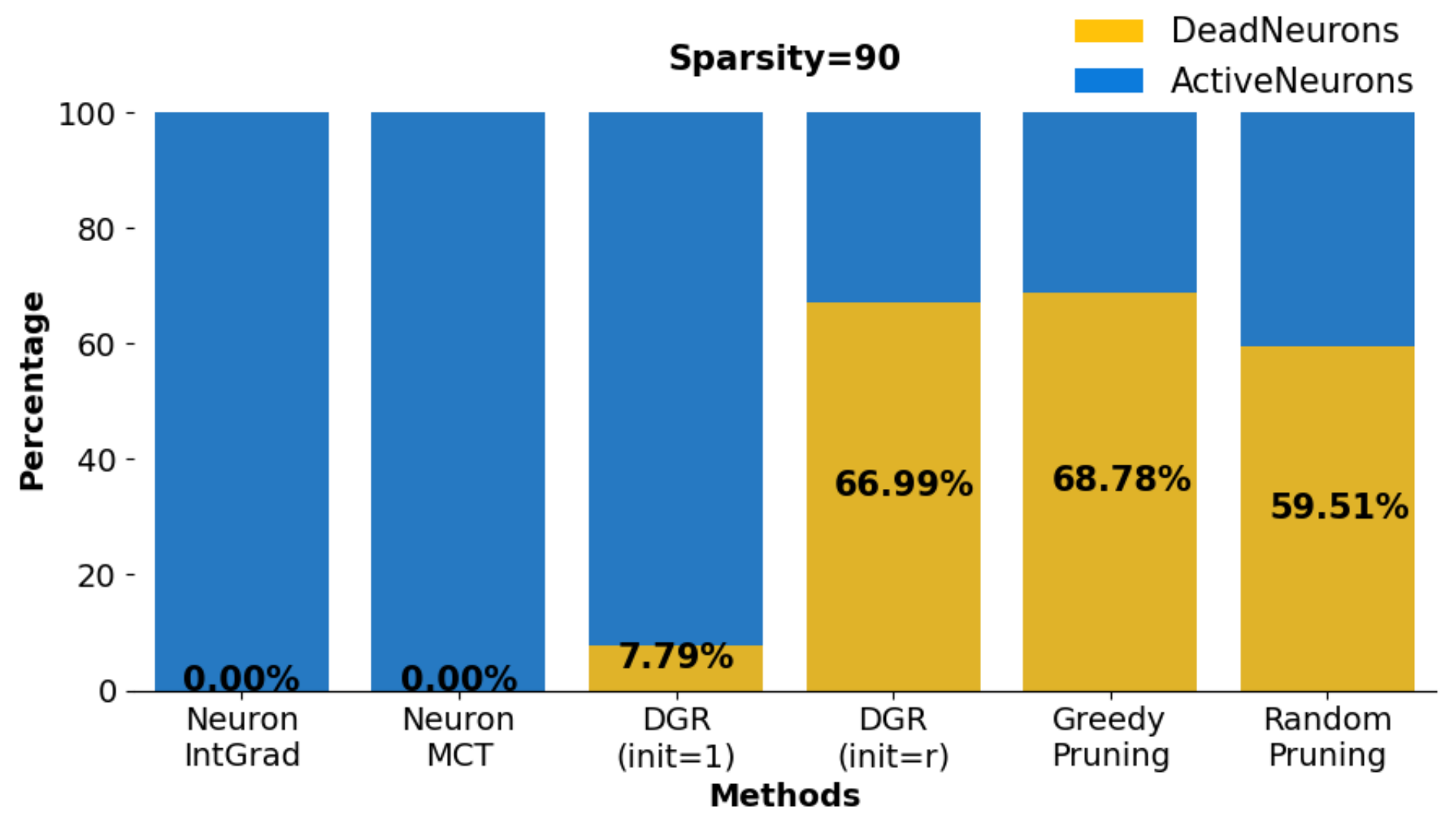
- How to identify critical pathways?

Pathway Identification via Pruning Objective

Pruning objective: Selecting a small subset of neurons for which the response remains close to the original response of the network

The pruning objective does not identify critical pathways

- How does pruning select irrelevant pathways?
We show "how" by devising a pathological pruning algorithm that intentionally selects irrelevant paths (originally dead neurons) but satisfies the objective.

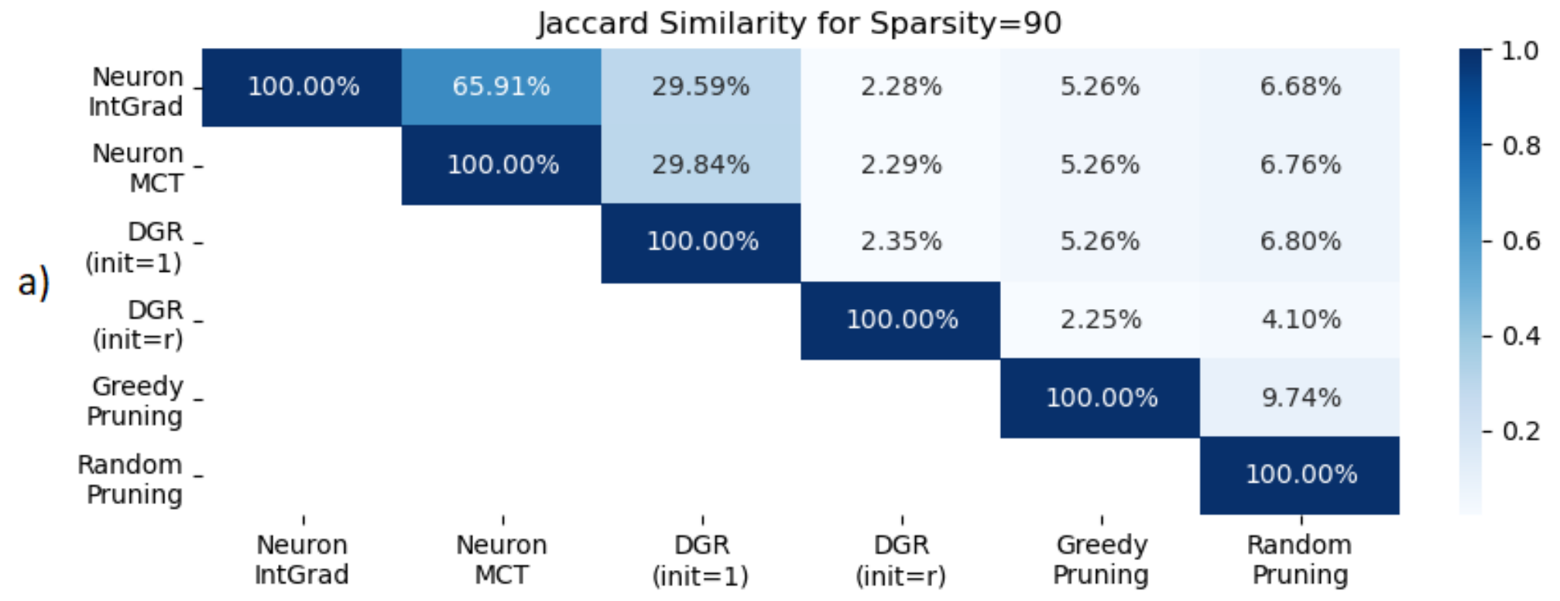


Pathway Identification via Neuron Contribution

Pathway of critical neurons -> To ensure sparse pathways include critical fragments of the encoded input information, we propose pathway selection via neurons' contribution

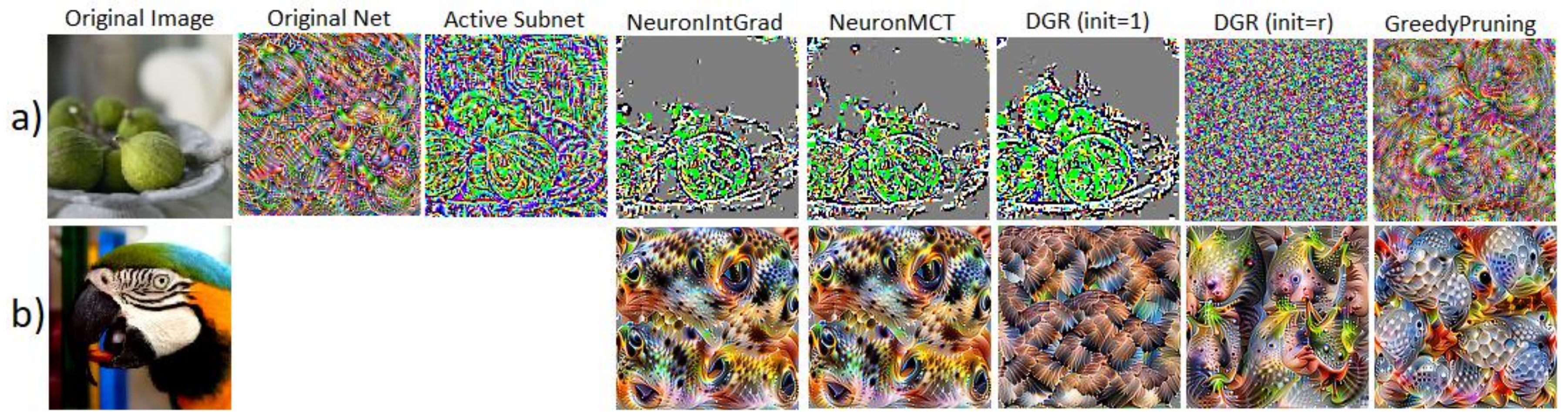
Pathway Analysis

The paths selected by different pruning methods do not overlap
-> many paths satisfy the pruning objective



Pathway Decoding

- What features are associated with the pathways?
a) What pattern corresponds to each path?
b) What pattern corresponds to most important neuron in each path?

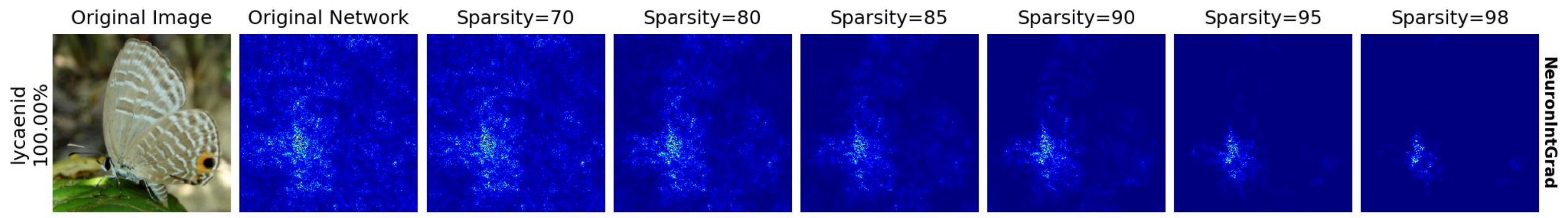


Feature Attribution via Pathway Gradient

Critical pathways selected via neuron contributions are locally linear in a neighborhood (Original network is not locally linear)

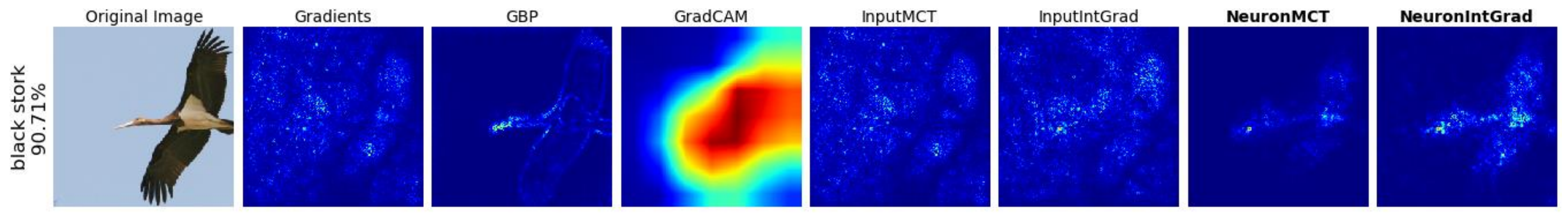
Pathway locally linear -> gradient reflects the local critical input features

We leverage local linearity to identify what features in the input are contributing to the response –
> We propose "**Pathway Gradient**" feature attribution method

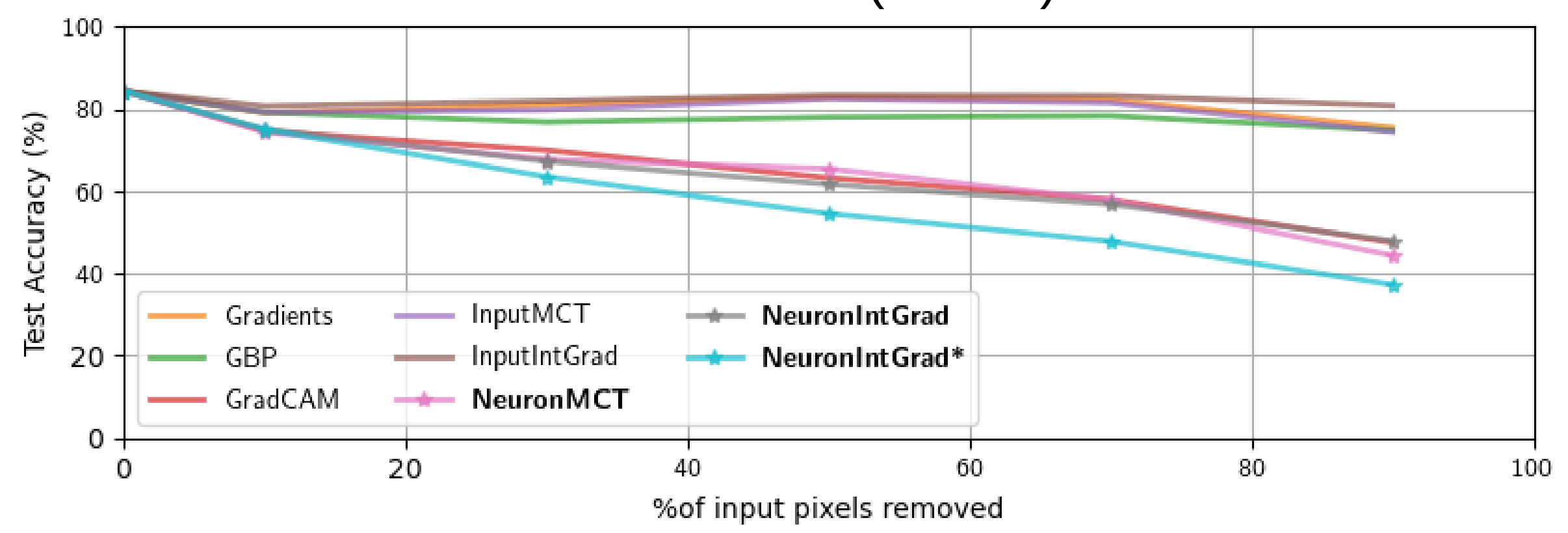


As we select sparser critical pathways, Pathway Gradient reveals input features that are more critical.

Feature attribution experiments also confirm that selected pathways using neuron contributions correspond to critical input features



Remove-and-Retrain (ROAR) evaluation



Acknowledgements: The authors acknowledge the support of the Munich Center for Machine Learning (MCML) and partial support of Siemens Healthineers. C. Rupprecht is supported by Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI) and by the European Research Council (ERC) IDIU-638009.