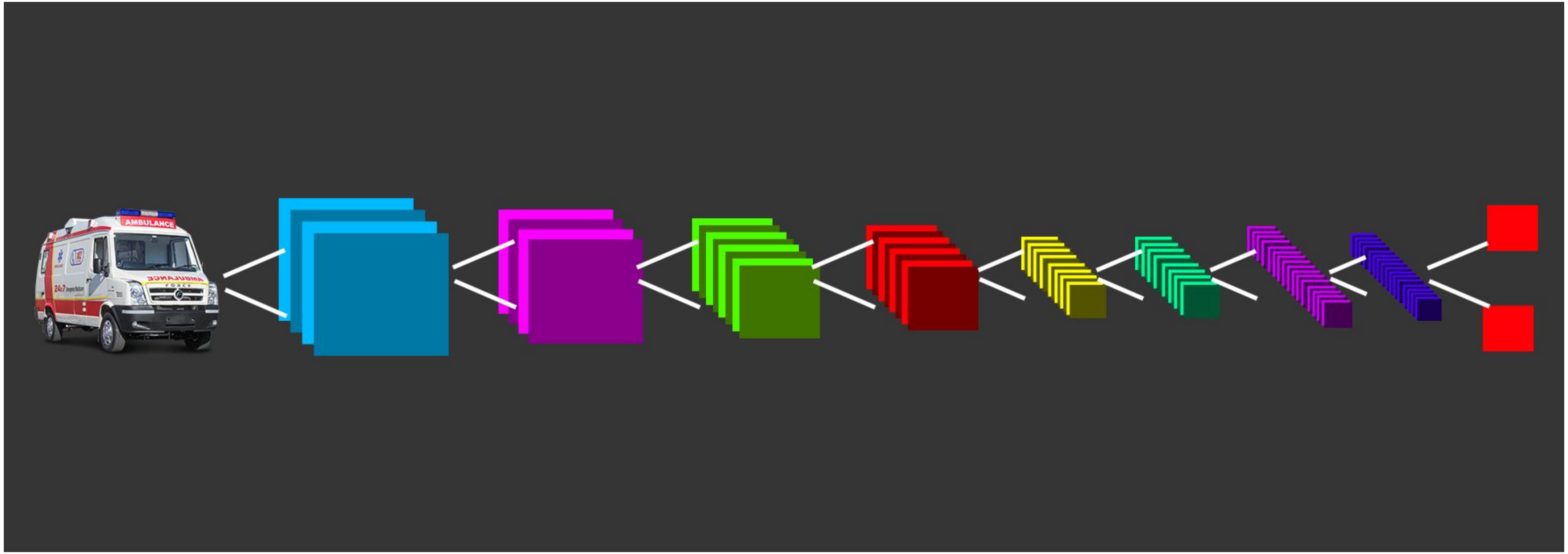
How can we interpret a Neural Network?

Are Neural Networks really incomprehensible?



Attempts to Interpret a Neural Network

Attempts to Interpret a Neural Network

We can

- Understand the model architecture

Attempts to Interpret a Neural Network

We can

- Understand the model architecture
- Visualize the filters / weights

Attempts to Interpret a Neural Network

We can

- Understand the model architecture
- Visualize the filters / weights
- Extract the output of intermediate neurons / layers

Attempts to Interpret a Neural Network

We can

- Understand the model architecture
- Visualize the filters / weights
- Extract the output of intermediate neurons / layers
- Locate important parts of the image according to the model

Attempts to Interpret a Neural Network

Suggestions?

Thank You