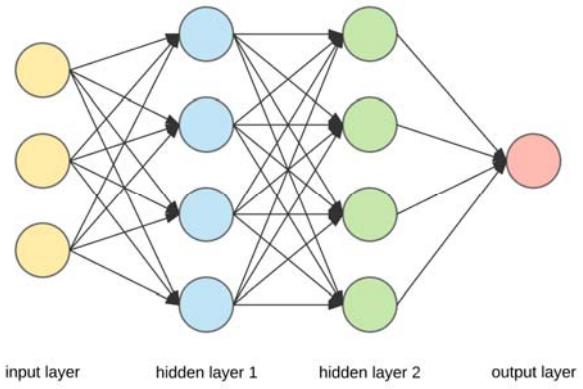
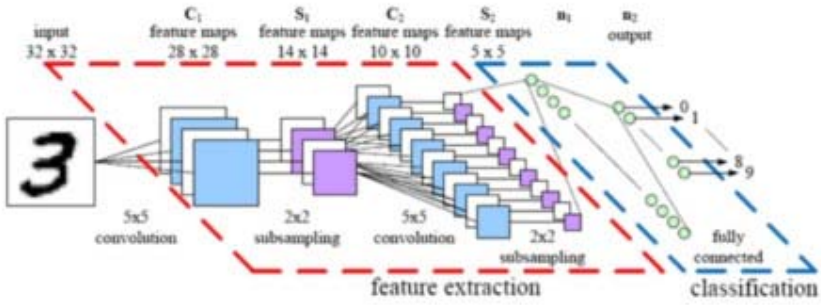
Interpretable Convolutional Neural Networks

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu
University of California, Los Angeles

Neural Network



Convolutional Neural Network



Abstract

- Traditional CNN to interpretable CNN (iCNN) using same training data without additional information.
- Clarify knowledge in higher convolution layers of CNNs.
- In iCNN, each filter represents an object part.
- Filters are more meaningful.
- Helps understand the logic inside a CNN (the patterns memorized by CNN for predictions)

Introduction

- CNN better for object classification/ detection.
- Interpretability as important as discrimination (Bau.)
- iCNN adds interpretability without human supervision.
- Introspect representation at each layer for better interpretation.
- Existing methods for interpretation:
  - Conventional Off line Visualisation of each filter.
  - Diagnosis of pre trained CNN representations.

Introduction

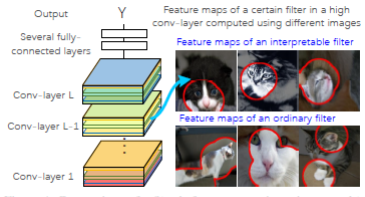


Figure 1. Comparison of a filter's feature maps in an interpretable CNN and those in a traditional CNN.

- Semantics: Shape (Object, Parts), Texture (Scenes, textures, materials, colours).
- Lower layers: Texture
- Higher layers: Parts
- Train each layer in higher layer for a part.
- Traditional CNN: Filter activated for multiple parts (due to learning), but in iCNN only for one part.

Goal: Revise CNN, no annotations, no sample/cost change
Impact: Decreased discrimination

Method

- Add loss for each filter such that it encodes a distinct object contained by single category.
- Filter should be activated by a single part of object rather than repetitions.
- Ex: Right eye and left eye should be triggered separately.
- Assumption: Repetitiveness is a function of texture in lower layers.
- Interpretability and clear semantics important for human trust in prediction.

Figure

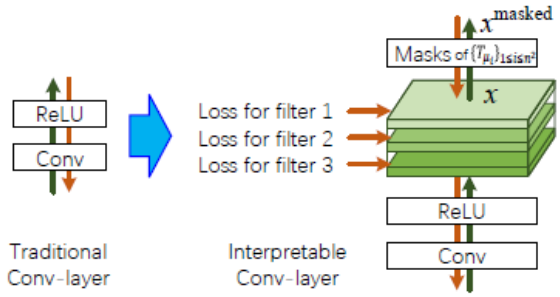


Figure 2. Structures of an ordinary conv-layer and an interpretable conv-layer. Green and red lines indicate the forward and backward propagations, respectively.

Related Work

- Network Visualisation
  - Visualisation of filter in terms of what it is rather than what it is doing.
- Pattern retrievals
  - Extraction of patterns/ features in mid level layers.
  - Discover objects from feature maps of certain filters.
- Model Diagnosis
  - Human intervention for supervision of the model.
  - Ex: LIME Method

Algorithm

- I images; I_c c I belongs to category, c.
- Filter 'f' in a convolution layer.
- Feature map 'x' of 'f' after RELU.
- 'f' object may appear at different locations.
- N^2 template for 'f': T_1, T_2, … $T_N{}^{2.}$
- T_i describes the ideal distribution for 'x' when part triggers ith unit in x.

Figure

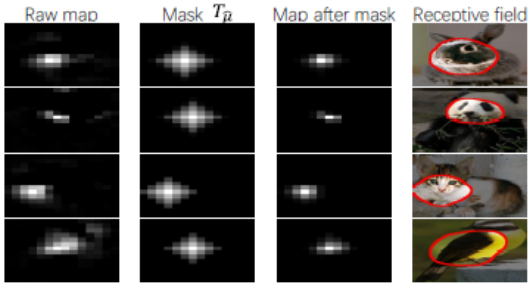


Figure 4. Given an input image I, from the left to the right, we consequently show the feature map of a filter after the ReLU layer x, the assigned mask $T_{\hat{p}}$, the masked feature map x^{masked}, and the image-resolution RF of activations in x^{masked} computed by [38].

Propagation

Forward

- Choose a template from Ti for input image I.
- Mask the image from noise activations.

Backward

- Loss pushes ‘f’ to represent specific part of category ‘c’.
- Use complement of T_i mask for eliminating noises.
- Loss for ‘f’ is the mutual information between T and X.

Model

Learning

- Forward propagation is similar to that in traditional CNN.
- During back propagation, each filter in iCNN receives gradient wrt its feature map x from both final task loss and its local filter loss.

Experiments

- Three datasets:
  - ILSVRC 2013 DET Animal part
  - CUB200-2011 dataset for birds
  - Pascal VOC Part dataset
- Four CNNs to iCNNs:
  - Alex Net, VGG-M, VGG-S, VGG-16

Comparison

Figure 5. Visualization of filters in top conv-layers. We used [38] to estimate the image-resolution receptive field of activations in a feature map to visualize a filter's semantics. The top four rows visualize filters in interpretable CNNs, and the bottom two rows correspond to filters in ordinary CNNs. We found that interpretable CNNs usually encoded head patterns of animals in its top conv-layer for classification.

Evaluation Metrics

- Part interpretability using Related Fields
  - Filter > Feature Map > Activation Scores > Threshold > Top Activations > Scaling from Low resolution to image > Related Field
- Location Stability
  - Distance between two parts across images.
  - Say head and eye.

Results:

- Better metrics than traditional CNNs for interpretation.
- Better accuracy for multi category classification.