No More Convolutional Neural Networks!

CS 510
Lecture #26
April 7th, 2017

Announcements (repeat)

• PA4 is due Monday, April 17th
• Test #2 will be Wednesday, April 19th
• Test #3 is Monday, May 8th at 8AM
  – Just 1 hour long
  – University schedule says 7:30…

4/7/17 CS 510, Image Computation, ©Ross Beveridge & Bruce Draper

VGG.PY

Don’t need to touch this
This is where you load your images

Invokes VGG net, returns activations (+ example print)

Image Pre-processing Example

Batch Normalization (II)

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]
\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \]

Step 1: Compute Batch Mean
Step 2: Compute Batch St. Dev.

\[ \hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \]
\[ y_i = \alpha \hat{x}_i + \beta \]

Step 3: Normalize Data
Step 4: Rescale Data & Move Mean
Note: \( \alpha \) and \( \beta \) trained by backprop
Resolution

- Resolution is a “by-pass”
- After 2 convolutions, the original source is added back in
- And of course these are stacked

ResNet

- Top Performer (2016)
- Regular architecture
  - 3x3 convolutions
  - Half size → double depth
  - Avg. Pooling*
  - Choice: # of convolutions between pooling layers
- Exceptions
  - 1st layer (7x7)
  - Last layer (7x7)

Depth Improves Performance

- More layers → more weights
- Deeper images → more weights
- For an image depth of 256:
  - Residual (left) → 9x256x256x2 = 1,179,648 weights/layer
  - Bottleneck (right) → 256x64+9x64x64+64x256 = 69,632 weights/layer

Bottleneck Architecture

- K. He, et al., “Deep Residual Learning for Image Recognition”, IEEE Conference on Computer Vision and Pattern Recognition, 2016, Figure 5

ResNet 152

- Bottleneck architecture
- 152 layers deep
  - Note: bottleneck is 3 layers
  - 50 of the units shown here
  - Plus 1st/last layers
- SOA (in 2016)
- Are bottleneck units the equivalent of residual units?

Is vision solved?

- Limits of CNNs
  - Performance not yet at ceiling (I think)
  - VIDEO!
  - Fast adaptation
    - How many samples does it take to learn a class?
    - Standard practice:
      - Keep convolutional layers
      - Refrain fully connected layers
      - Still takes a lot of training samples
  - Performance characterization/guarantees
    - Would you trust your life to a CNN?
      - Think about self-driving cars...
    - Predicting failure, guaranteeing success
      - Similar to what control theory did for auto-pilots
  - Fast & low-power application of CNNs
    - Chip design
Is vision solved? (II)

- Classification ∈ Vision
  - Focus of attention ∈ Vision
  - Tracking ∈ Vision
  - 3D Reconstruction ∈ Vision
  - Attributes ∈ Vision
  - Expertise ∈ Vision
  - Recognition ∈ Vision

What is (Visual) Expertise?

- Example: Faces
  - When you see an image with a face:
    - You detect & categorize the face
      - Whether its relevant to your task of not
      - Faces are always salient
    - You recognize attributes of the face
      - Such as expression, age, and gender
      - Whether its relevant to your task of not
      - Highly sensitive to changes in attributes
    - You know if the face is familiar
      - You know if you have seen this particular face before
      - If it’s someone you know, you identify them
      - Whether relevant to your task or not

What is (Visual) Expertise? (II)

- Expertise identifies instances within classes.
  - e.g. Recognizing your car in the parking lot
- Expertise measures attributes
- Expertise is trained
  - Exposure to many instances
  - Instance differentiation is salient
  - Common important objects, e.g. faces & cars
  - Result of training
    - dog show judges
    - ornithologists

What is Recognition?

- Recognition is multi-modal
- Example: when you recognize a cat
  - It triggers the word ‘cat’
  - It anticipates the sound of cat
    - Meow, caterwall
  - It anticipates the tactile feel of a cat
    - Soft, furry
  - It anticipates episodic memories of cats
    - I was scratched once…
- Vision isn’t solved until its integrates with a larger AI.