THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

ECE 599/692 – Deep Learning

Lecture 8 – CNN: Advanced Topics

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

1

Outline

- Lecture 3: Core ideas of CNN
  - Receptive field
  - Pooling
  - Shared weight
  - Derivation of BP in CNN
- Lecture 4: Practical issues
  - The learning slowdown problem
    - Quadratic cost function
    - Cross-entropy + sigmoid
    - Log-likelihood + softmax
  - Overfitting and regularization
    - L2 vs. L1 normalization
    - Dropout
    - Artificial expanding the training set
  - Weight initialization
  - How to choose hyper-parameters
    - Learning rate, early stopping, learning schedule, regularization parameter, mini-batch size,
    - Grid search
  - Others
    - Momentum-based GD
- Lecture 5: The representative power of NN
- Lecture 6: Variants of CNN
  - From LeNet to AlexNet to GoogleNet to VGG to ResNet
- Lecture 7: Implementation on VGGNet
- Lecture 8: Random thoughts

THE UNIVERSITY OF TENNESSEE 2

Random Thoughts

AICIP RESEARCH

- Trend
  - Unsupervised learning
  - Attention is all you need
  - Alternatives to CNN
    - Graph network
    - Capsule network
  - The forever battle between globalness and localness
- How to find a research topic?

THE UNIVERSITY OF TENNESSEE 3

Acknowledgement

- The slides on global vs. local are from a group presentation slide set by Yang Song
- The slides on person re-id are from Alireza Rahimpour

Global vs. Local: The Local Mean Operation

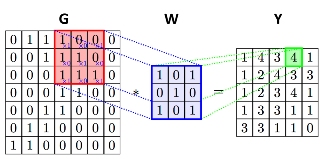


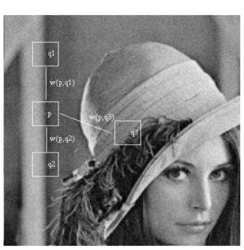
Fig1. Local 2D convolution operator

$$y_i = \sum_{j \in 3\times 3} w_j g(x_j)$$

i --- index of an output position
j --- index of a possible position

Global vs. Local: The Non-Local Mean (NLM) Operator

[Buades and Morel, CVPR 2005]



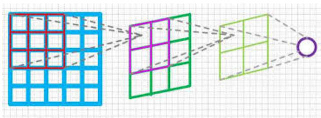
G: noise image

$$y_i = \sum_{j \in I} w(i,j) g(j)$$
$$w(i,j) = \frac{1}{C(i)} e^{-\frac{|\mu(i)-\mu(j)|^2}{h^2}}$$
$$C(i) = \sum_{j \in I} w(i,j)$$

How to do it in deep networks?

AICIP RESEARCH

- In sequential data, long-range dependency is formulated by recurrent operations, e.g., LSTM.
- In image data, long-distance dependency is mainly modeled by repeated Conv. layers progressively.



Some recent developments
1. Dilated convolution

AICIP RESEARCH

[Chen et al. TPAMI 2018]

Illustration of dilated convolution (1D)

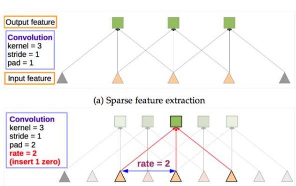
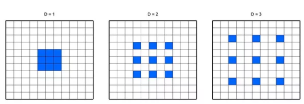


Illustration of dilated convolution (2D)



Pros: the same kernel size with larger receptive field

Some recent developments
2. Deformable CNN

AICIP RESEARCH

[Dai et al. ICCV 2017]

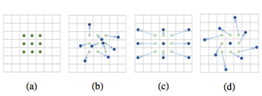


Figure 1: Illustration of the sampling locations in 3 × 3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.

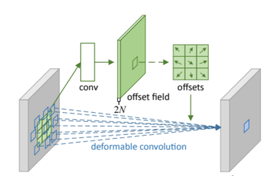
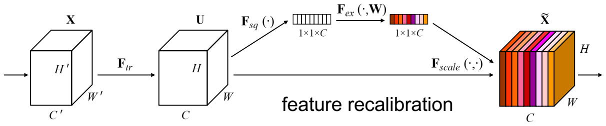


Figure 2: Illustration of 3 × 3 deformable convolution.

Some recent developments
3. Squeeze-and-Excitation Network (SENet)

[SENet:cvpr:2018]



10

The representative and discriminative power – A revisit to variants of CNN models

- AlexNet
- GoogLeNet
- VGGNet
- ResNet
- SENet

11

Random Thoughts

- Trend
  - Unsupervised learning
  - Attention is all you need
  - Alternatives to CNN
    - Graph network
    - Capsule network
  - The forever battle between globalness and localness
- How to find a research topic?

12

AICIP RESEARCH

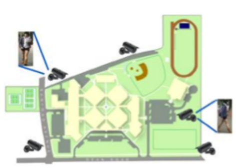
Random Thoughts

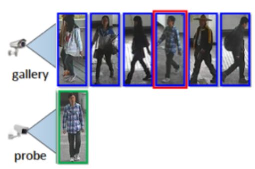
- Trend
  - Unsupervised learning
  - Attention is all you need
    - Person re-identification
    - Attention
    - A new cost function
  - Alternatives to CNN
    - Graph network
    - Capsule network
  - The forever battle between globalness and localness
- How to find a research topic?

AICIP RESEARCH

Person Re-identification

- Person re-identification is the problem of matching the same individuals across multiple cameras, or across time within a single camera.





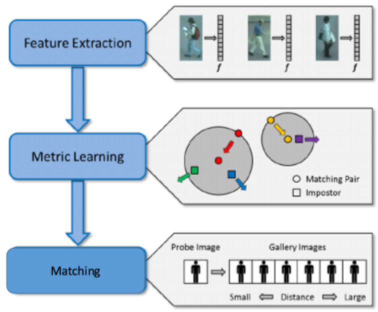
AICIP RESEARCH

Challenges:

(a) Cross-view lighting variations (b) Camera viewpoint changes

(c) Clothing similarity (d) Background clutter & occlusions

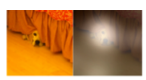
AICIP RESEARCH



AICIP RESEARCH

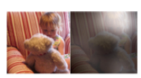
Motivation:

- **Attention**: Focusing on specific parts of the input.
- Inspired by neuroscience.



- To reduce the computational burden of processing high dimensional inputs by selecting to only process subsets of the input.
  e.g., Large Medical Images.

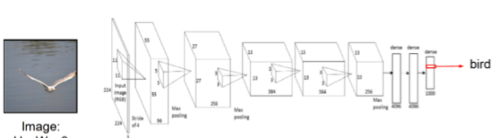
- To allow the system to focus on distinct aspects of the input and thus improve the generated output.



AICIP RESEARCH

Motivation:

Attention models in Deep Neural Networks

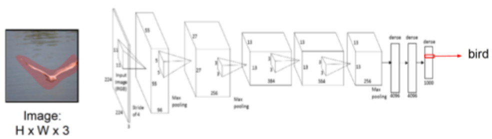


The whole input volume is used to predict the output...

Motivation:

AICIP RESEARCH

Attention models in Deep Neural Networks



The whole input volume is used to predict the output...

...despite the fact that not all pixels are equally important

Person Re-ID using Attention Mechanism

AICIP RESEARCH

Motivation:

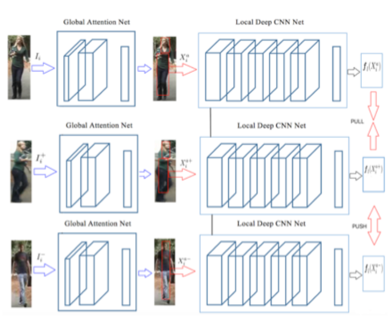
Humans do not focus their attention on an entire scene at once when they want to identify another person.
Instead, they pay attention to different parts of the scene (e.g., the person's face) to extract the most discriminative information.

Our proposed model objective:

Being able to focus on a certain region of an image with **high resolution** while perceiving the surrounding image in **low resolution**.

Model Architecture

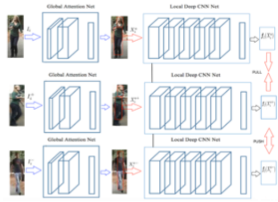
AICIP RESEARCH



Model Architecture-Triplet loss

AICIP RESEARCH

- We want the distance of the learned features of the same person to be less than the distance between the images from different persons by a defined margin.
- Cost function for N triplet images $J = \frac{1}{N}\sum_{i=1}^{N}[\|f_l(X_i^a) - f_l(X_i^{a+})\|_2^2 - \|f_l(X_i^a) - f_l(X_i^{a-})\|_2^2 + \alpha]_+$



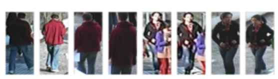
where the term $[z]_+ = \max(z, 0)$ denotes the standard hinge loss.

Datasets:

AICIP RESEARCH

- The CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Each person has four images with two from each camera.

- The CUHK03 dataset contains 13164 images of 1360 identities. All pedestrians are captured by six cameras, and each person's image is only taken from two camera views.



Preliminary Results:

AICIP RESEARCH

Comparison of performance of the proposed GAN to the state-of-the-art on CUHK01 dataset.

Method	Rank1	Rank5	Rank10	Rank20
FPNN (li2014,CVPR)	22.87	58.20	73.46	86.31
SDALF (farenzena,2010,CVPR)	9.90	41.21	56.00	66.37
eSDC (zhao,2013,CVPR)	22.84	43.89	57.67	69.84
KISSME (kostinger,2012,CVPR)	29.40	57.67	72.42	86.07
Partb-reid (cheng,2016,CVPR)	53.7	84.3	**91.0**	96.3
GAN-L	54.6	83.6	89.4	90.2
GAN	**64.2**	**86.4**	90.6	**96.9**

Random Thoughts

AICIP RESEARCH

- Trend
  - Unsupervised learning
  - Attention is all you need
  - Alternatives to CNN
    - Graph network
    - Capsule network
  - The forever battle between globalness and localness
- How to find a research topic? (from Ng)

THE UNIVERSITY OF TENNESSEE KNOXVILLE

25