ConvNets & Multi-modal Log-bilinear Language Model

Renjie Liao
2017.Feb.14

*Some materials are credited to Jamie Kiros*
Motivation – ConvNets are everywhere!

(Krizhevsky et al, 2012)
Motivation – ConvNets are everywhere!

Face recognition (Taigman et al, 2014)

Action recognition from video (Simonyan et al, 2014)
Motivation – ConvNets are everywhere!

- Street sign recognition (Sermanet et al, 2011)
- Galaxy classification (Dieleman et al, 2014)
- Mitosis detection (Ciresan et al, 2013)
Motivation – ConvNets are everywhere!

Playing Atari games (Mnih et al, 2013)

- Many, many more applications (and not only vision):
  - Object detection
  - Image segmentation
  - Pose estimation
  - Image captioning
  - Pedestrian detection
  - Semantic image search
  - Extractive summarization
  - Sentiment analysis of text
A brief review

Fully connected:
(unique weights across all pairs of neurons)

Main operation:
Matrix Multiply

Convnet:
(neurons are volumes, weights are shared)

Main operation:
Convolution
Some terminology

Channels (e.g. 3 for RGB image)

Kernel (or filter)
5 in this example

Each “slice” across depth is a feature map

(think of this just like an image, but with 5 channels instead)
1D forward pass, strides, padding

- Weight sharing: the kernel is scanned across the input (as opposed to fully connected networks)

- Larger strides reduce computation cost, but usually at the expense of accuracy

- In this example, each side is “padded” with an extra 0
Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.

\[
\begin{array}{c}
\begin{array}{c}
(4 \times 0) \\
(0 \times 0) \\
(0 \times 0) \\
(0 \times 0) \\
(0 \times 0) \\
(0 \times 1) \\
(0 \times 1) \\
(0 \times 1) \\
(0 \times 1) \\
+ \ (-4 \times 2) \\
-8
\end{array}
\end{array}
\]
2D Convolution Example
Example #1

- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 1
- No padding

- What is the output volume?
- How many parameters are there?
Example #1

- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 1
- No padding

- What is the output volume?  
  28 x 28 x 5
- How many parameters are there?  
  ((5 x 5) x 3) x 5 = 375
Example #2

- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 3
- No padding

- What is the output volume?
- How many parameters are there?
Example #2

- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 3
- No padding

- What is the output volume? 10 x 10 x 5
- How many parameters are there? \((5 \times 5) \times 3 \times 5 = 375\)
GENERATING TEXT CONDITIONED ON IMAGES

In this picture there is another grey pavement on the right; three grey clouds and a blue sky in the background; the houses and on the left before it; a dark green, wooded slopes behind it; grey clouds in a light blue sky in the background; snow covered mountains.

This product contains a slip resistant and mesh upper is fully designed for breathable durability. The detachable leather footbed is the high, they feature a lady - like footbed that light sophistication and flirty tear silhouette to glam up your feet, style to help your thing. With traditional support.
The Log-Bilinear Language Model (LBL)

- Word representations $r_{w_i}$, context matrices $C_i$
- Predicted next word representation $\hat{r} = \sum_{i=1}^{n-1} C_i r_{w_i}$
- R: matrix where each row is a word feature from the vocabulary
- Score $\hat{r}$ with each word and normalize:

$$P(w_n = w | w_1:n-1) = \frac{\exp(\hat{r}^T r_w + b_w)}{\sum_j \exp(\hat{r}^T r_j + b_j)}$$

- Backprop through both parameters and word embeddings
Suppose we have image features $\mathbf{x}$

- Simplest approach: Bias the predicted next word representation:

$$\hat{\mathbf{r}} = \left( \sum_{i=1}^{n-1} C_i r_{w_i} \right) + C_m \mathbf{x}$$

- This turns out to be a surprisingly effective model (given good image features)