DEEP LEARNING INSTITUTE

Multiple Data Types (3주차)

Young-Gon Kim
DLI Instructor

DEEP LEARNING INSTITUTE

DLI Mission

Helping people solve challenging problems using AI and deep learning.

- Developers, data scientists and engineers
- Self-driving cars, healthcare and robotics
- Training, optimizing, and deploying deep neural networks

TOPICS

- Week 2 Review
- Image Captioning
- Video Captioning

WEEK 2 REVIEW

IMAGE CAPTIONING

IMAGE CAPTIONING

- Data / Network
    - Microsoft Common Object in Context (MS COCO)
        - Images
        - Five captions for each image
    - VGG16 Network
        - Visual Geometry Group

IMAGE CAPTIONING

- Process
  1. Import libraries
  2. Evaluate data / Pixel to Content
     - Feature vector – FC7
  3. Align captions with images
     - Will work with a subset of the data
  4. Predict next word
     - Parse, tokenize, etc.

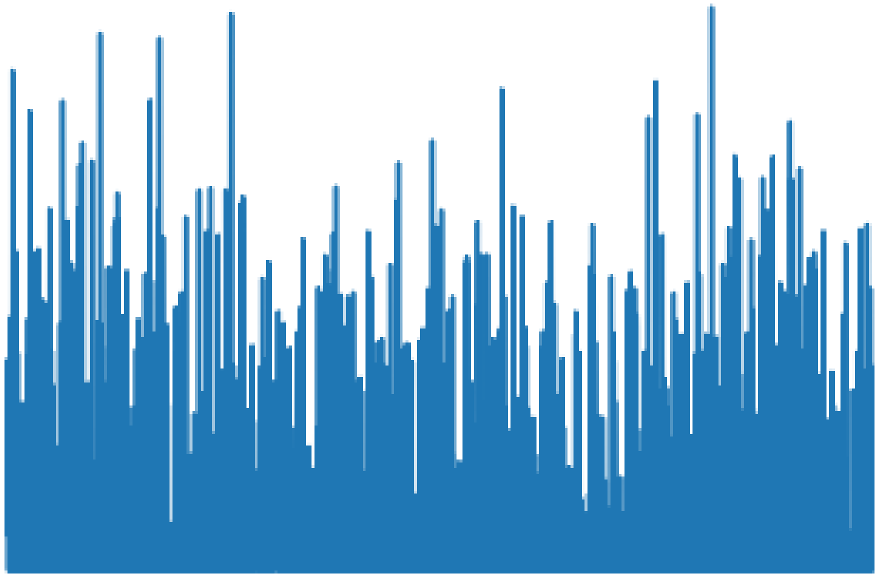


IMAGE CAPTIONING

$$
\begin{array}{l|cccccccccc|}
\text{image} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{a} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{dog} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{with} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{cake} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

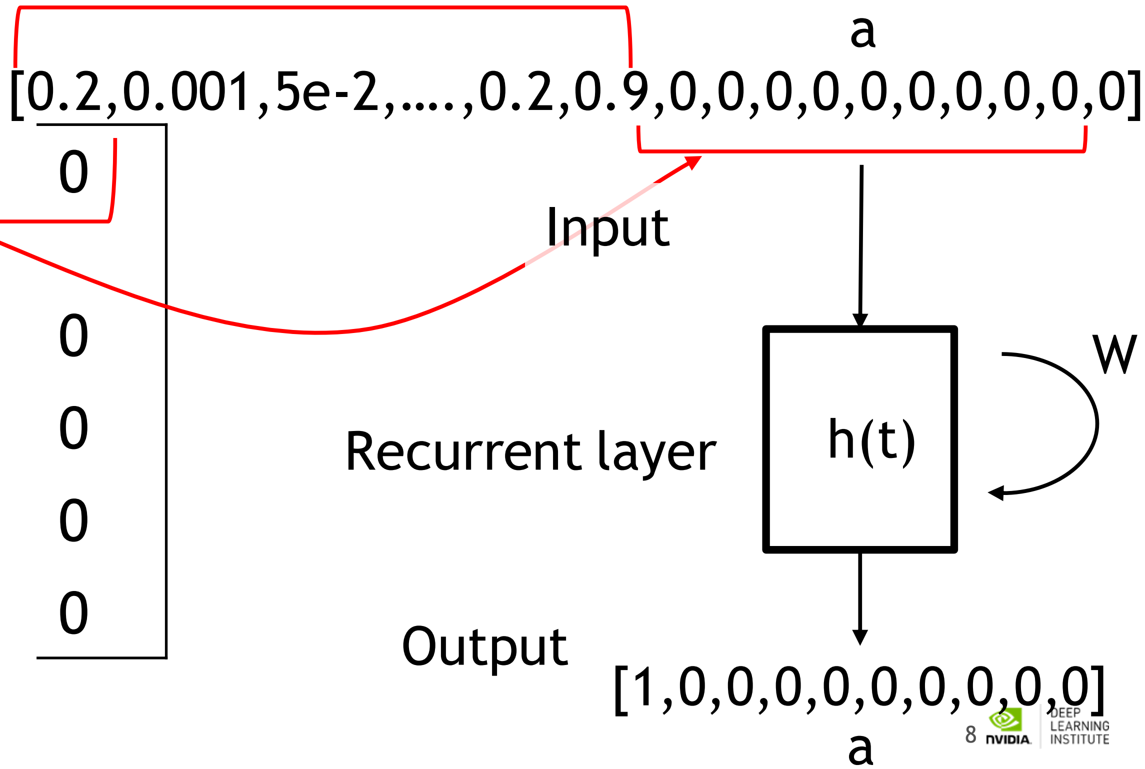
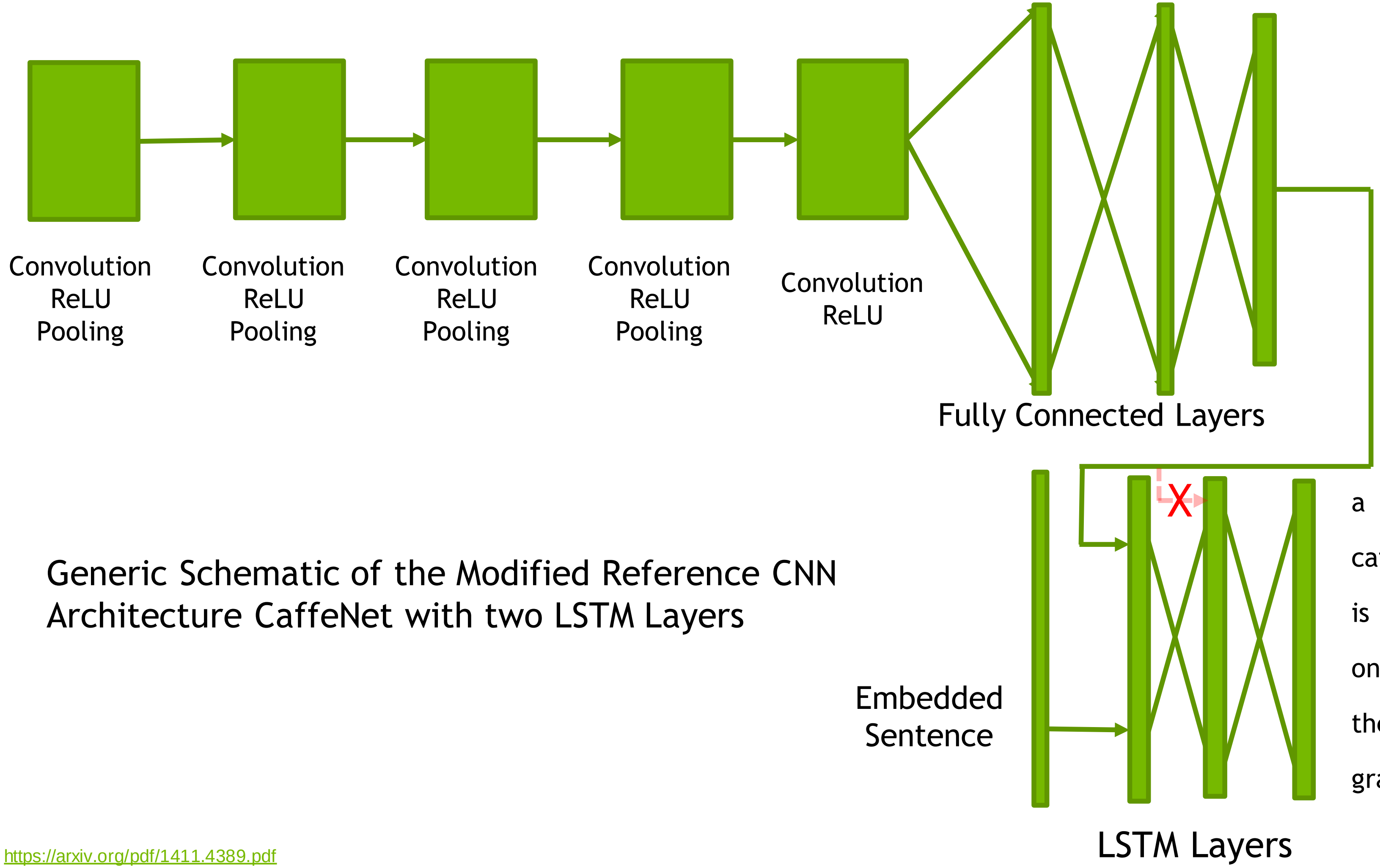


IMAGE CAPTIONING

- Process
   5. Architecture the network (RNN)
   6. Train / build model
   7. Evaluate a training image & captions
   8. Generate a caption for a validation image
   9. RUN LAST CODE BLOCK TO FREE GPU MEMORY

IMAGE CAPTIONING



Generic Schematic of the Modified Reference CNN Architecture CaffeNet with two LSTM Layers

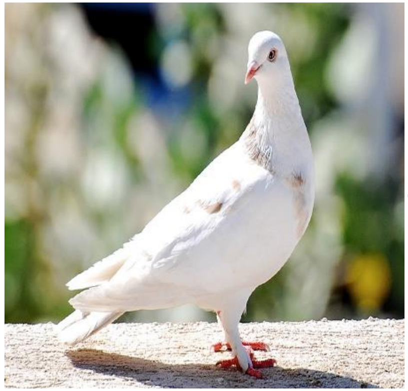
Prediction

$$\begin{array}{l|cccccccccc|}
\text{a} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{cat} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{is} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{on} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{the} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{grass} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}$$

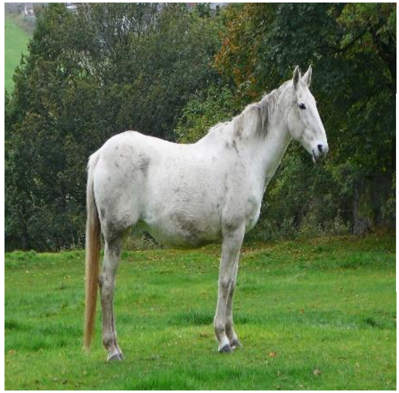
https://arxiv.org/pdf/1411.4389.pdf

IMAGE CAPTIONING

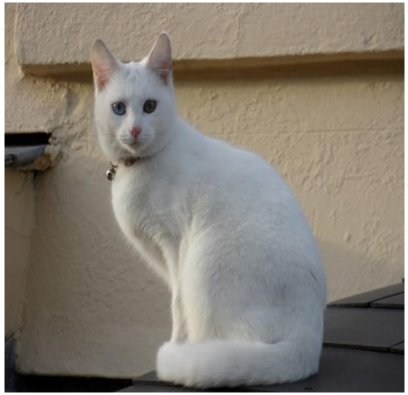
- Results



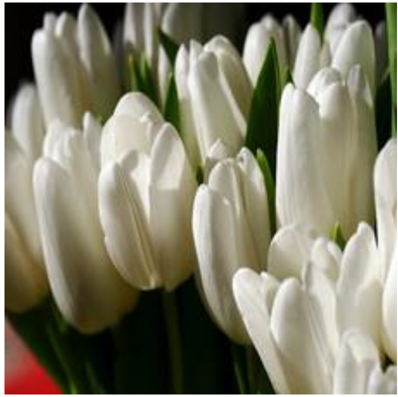
CaffeNet	A white bird standing on top of a sandy beach.
VGG	A small bird standing on the ground.



CaffeNet	A white horse standing in a lush field of grass.
VGG	A white horse standing in a field next to a fence.



CaffeNet	A white cat sitting on a chair.
VGG	A white and white cat laying on a white chair.



CaffeNet	A bunch of bananas that are on a table.
VGG	A close up of a bunch of white flowers.

https://arxiv.org/pdf/1411.4389.pdf

VIDEO CAPTIONING

VIDEO CAPTIONING

- Data / Network
  - Microsoft Research Video Description Corpus (MSVD)
    - About 2,000 video clips
    - Ten captions for each video
  - VGG16 Network
    - Visual Geometry Group

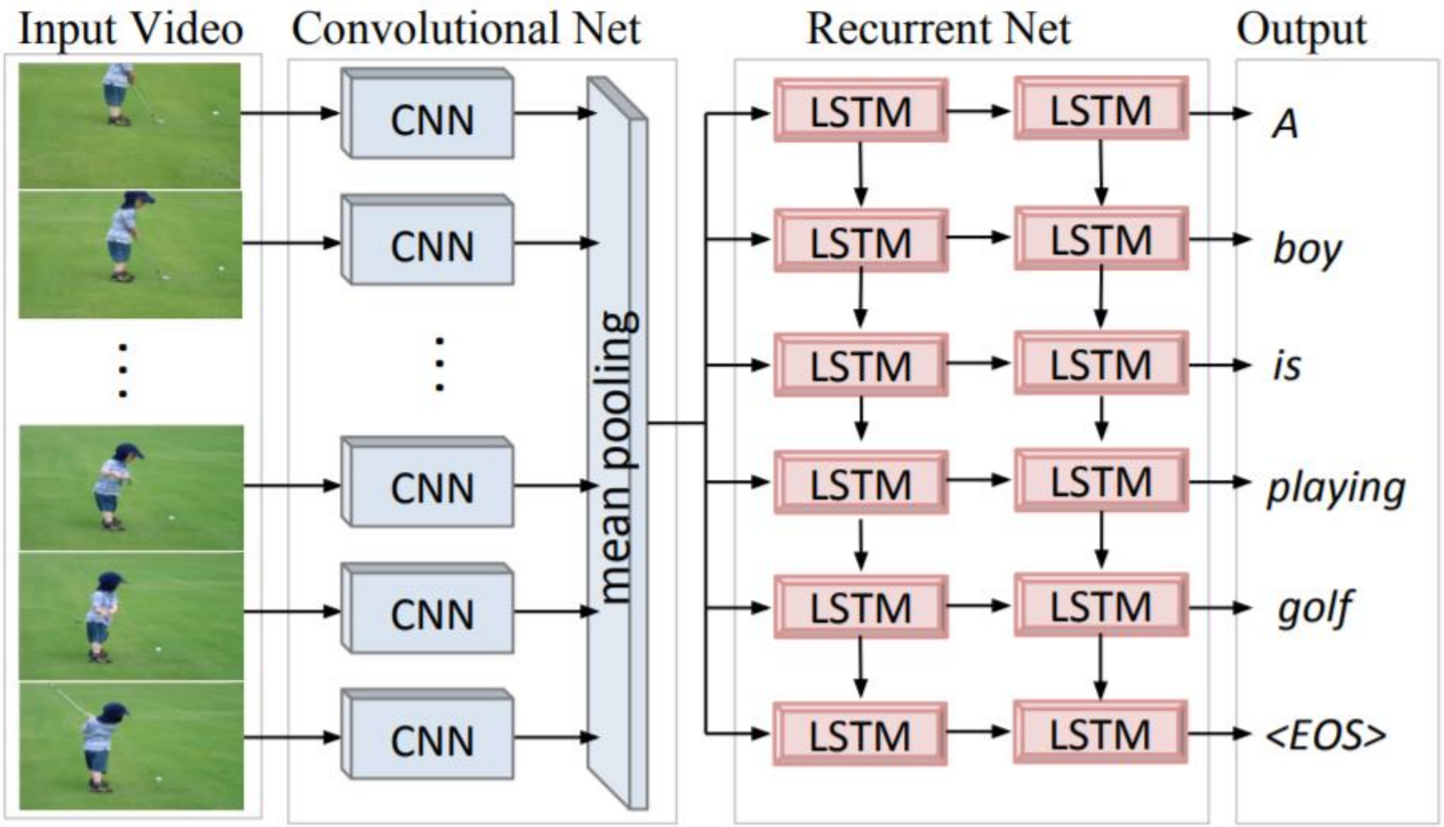
VIDEO CAPTIONING

- Process
  1. Import libraries
  2. Evaluate videos and captions
     - Create a mean vector of a single clip
     - This will generate a high-level representation of each frame from layer fc7
  3. Align captions with feature maps
     - Will work with a subset of the data
  4. Predict next word for captions
     - Parse, tokenize, etc.

VIDEO CAPTIONING

- Process
   5. Architect the network (RNN)
      - NOTE: Troubleshooting wording before running code block
   6. Train / build model
      - NOTE: Troubleshooting wording before running code block
   7. Evaluate a training image & captions
   8. Generate a caption for a validation image

VIDEO CAPTIONING



Prediction

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

https://arxiv.org/pdf/1412.4729.pdf

VIDEO CAPTIONING

- Results

A man is riding a horse

A animal is eating

A dog is standing

https://arxiv.org/pdf/1412.4729.pdf
https://www.aclweb.org/anthology/P11-1020.pdf

Reference

- https://arxiv.org/pdf/1411.4389.pdf
- https://arxiv.org/pdf/1412.4729.pdf
- https://www.aclweb.org/anthology/P11-1020.pdf

NVIDIA

DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli