이미지 및 영상 자막 생성
Image and Video Captioning

TOPICS

- 실습 구조(Lab Structure)
- 이미지 자막 생성(Image Captioning)
- 비디오 자막 생성(Video Captioning)

LAB STRUCTURE

JUPYTER NOTEBOOKS

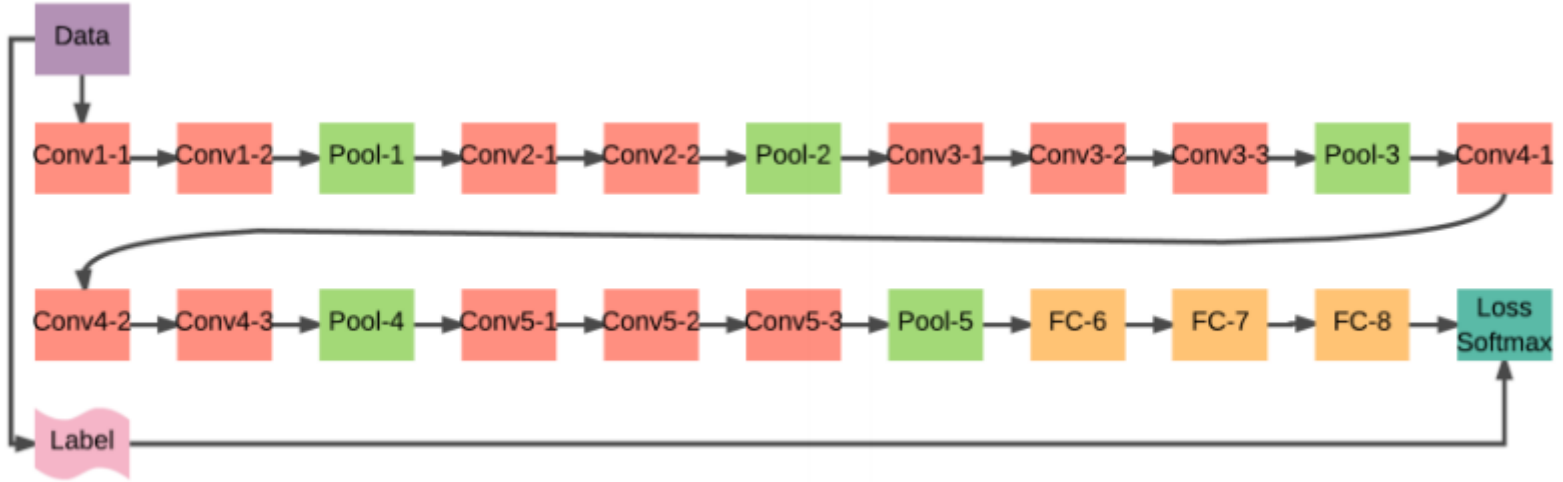
- 실습에 포함된 notebook은 다음과 같다:
  - Image Captioning notebook
  - Video Captioning notebook
  - Reference notebook - from Lab 2

이미지 자막 생성
IMAGE CAPTIONING

학습 데이터/네트워크
TRAINING DATA / NETWORK

이미지 자막 생성

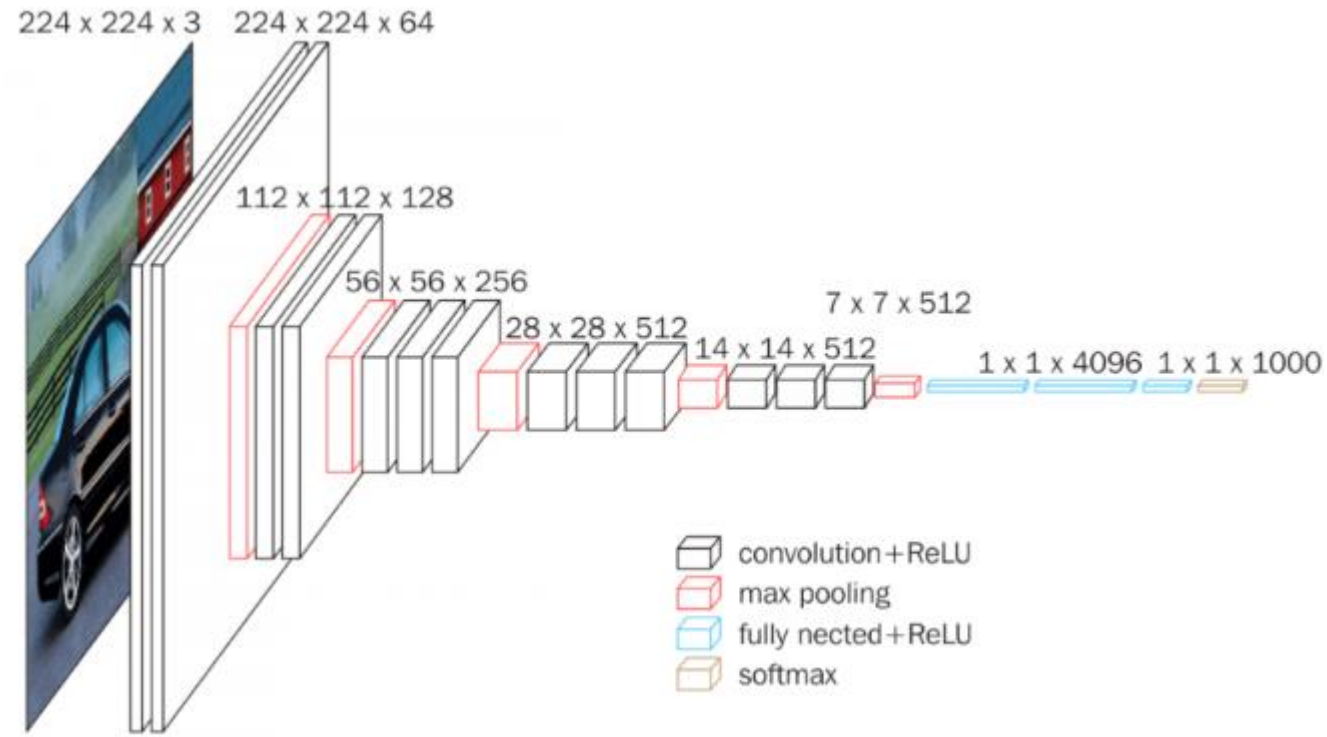
- Microsoft Common Object in Common (MSCOCO)
  - 이미지들
  - 각 이미지 당 최소 5개의 캡션
- VGG 16 네트워크
  - VGG16(Visual Geometry Group)



VGG16

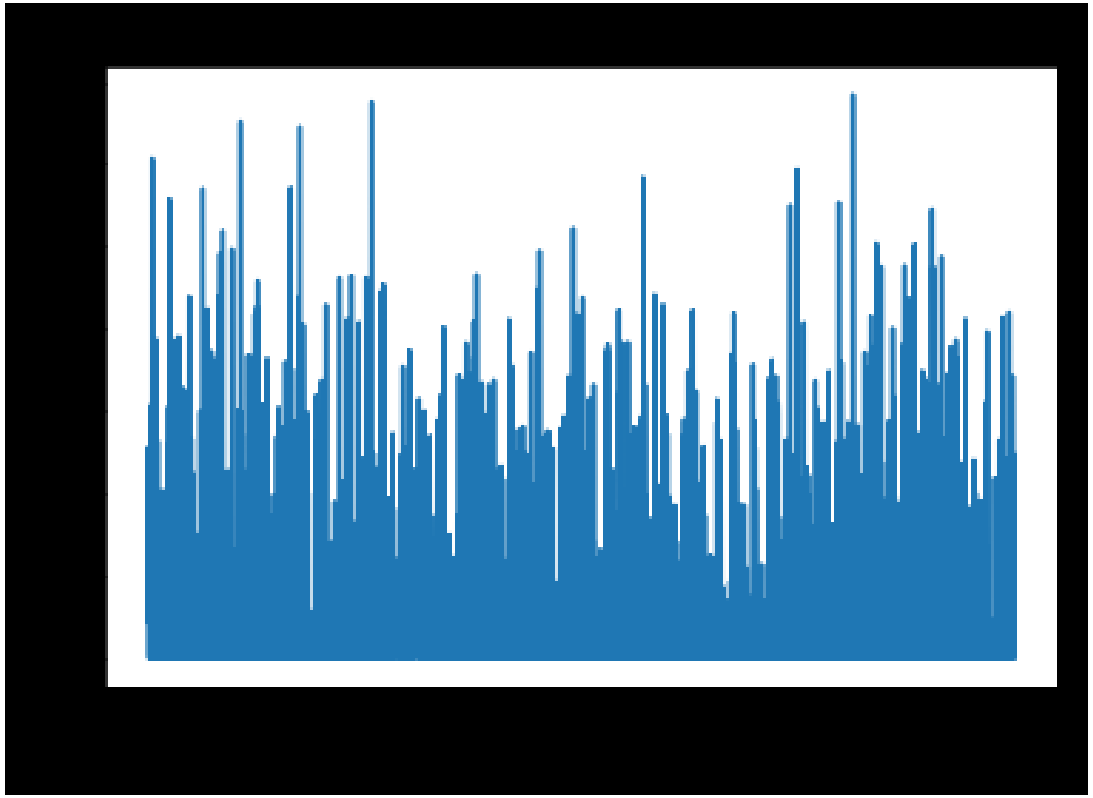
A	A-LRN	B	C	D	E
11개의 레이어	11개의 레이어	13개의 레이어	16개의 레이어	16개의 레이어	19개의 레이어
conv 3, 64	conv 3, 64	conv 3, 64	conv 3, 64	conv 3, 64	conv 3, 64
	LRN	conv 3, 64	conv 3, 64	conv 3, 64	conv 3, 64
maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2
conv 3, 128	conv 3, 128	conv 3, 128	conv 3, 128	conv 3, 128	conv 3, 128
		conv 3, 128	conv 3, 128	conv 3, 128	conv 3, 128
maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2
conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256
conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256	conv 3, 256
			conv 1, 256	conv 3, 256	conv 3, 256
					conv 3, 256
maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2	maxpool 2

VGG16



프로세스- 이미지 자막 생성

1. 라이브러리 가져오기
2. 데이터 평가/ 픽셀로부터 맥락까지
    a. Feature vector - FC7
3. 자막과 이미지 정렬
    a. 데이터의 서브셋으로 작동
4. 다음 단어 예측
    a. Lab 2와 유사
    b. Parse, tokenize, etc.



이미지 자막 생성

a

[0.2,0.001,5e-2,….,0.2,0.9,0,0,0,0,0,0,0,0,0,0]

image	0	0	0	0	0	0	0	0	0	0
a	1	0	0	0	0	0	0	0	0	0
dog	0	0	0	0	1	0	0	0	0	0
with	0	0	0	1	0	0	0	0	0	0
cake	0	0	1	0	0	0	0	0	0	0

Input

Recurrent layer

h(t)

W

Output

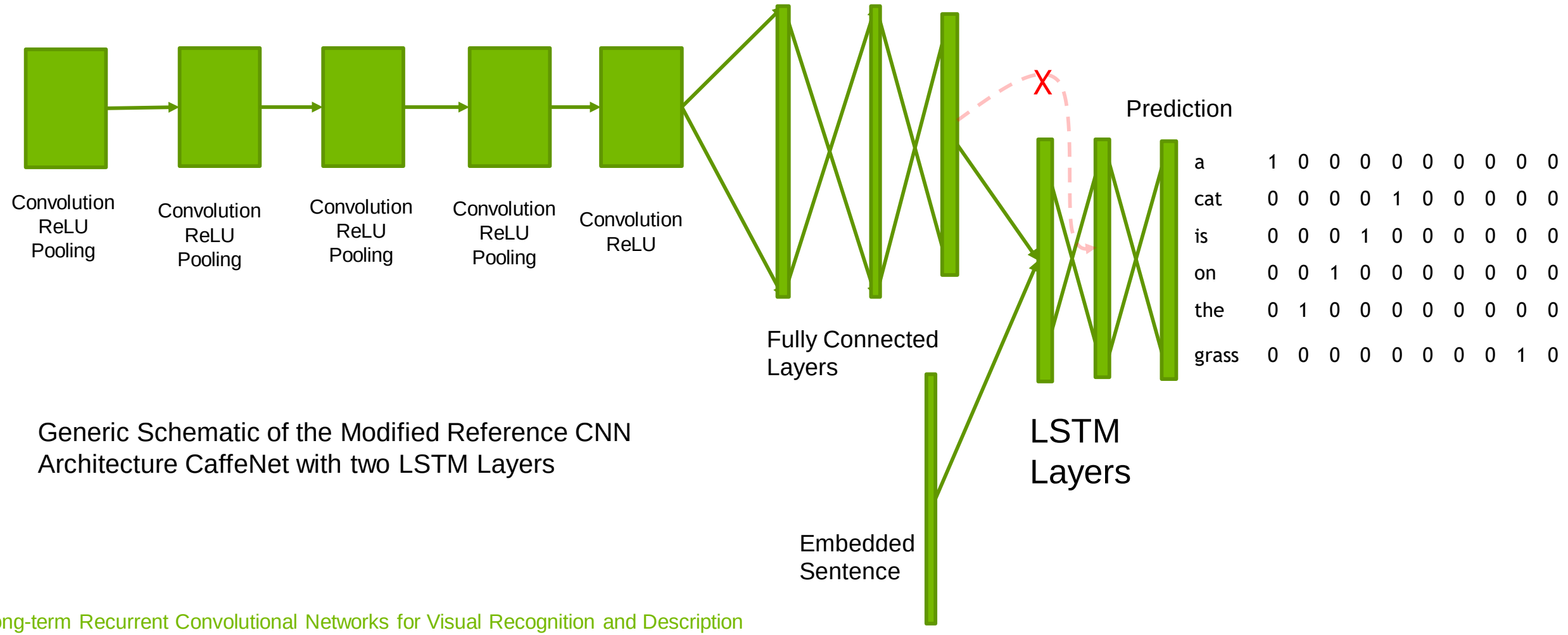
[1,0,0,0,0,0,0,0,0,0,0]

a

프로세스- 이미지 자막 생성

5. 네트워크 설계 (RNN)
6. 모델(model) 학습 및 개발
7. 학습된 이미지 및 캡션 평가
8. 검증 이미지에 캡션 생성
9. GPU 메모리 부하를 막기 위해 마지막 코드 블락 수행

LAB 3 - 이미지 자막 생성



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

Long-term Recurrent Convolutional Networks for Visual Recognition and Description
Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell

캡션 출력의 예

CaffeNet 모래사장에 앉아 있는 하얀 새

VGG 작은 새가 땅 위에 서있다

CaffeNet 의자에 앉아있는 고양이

VGG 하얀 의자에 앉아 있는 하얗고 하얀 고양이

CaffeNet 들판에 서있는 백마

VGG 울타리 옆 들판에 서있는 백마

CaffeNet 테이블 위의 바나나 한 무더기

VGG 하얀 꽃다발의 클로즈업 사진

*Results shown here were generated using work from this paper.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description
Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell

비디오 자막 생성
VIDEO CAPTIONING

학습 데이터/ 네트워크

비디오 자막 생성

- Microsoft Research Video Description Corpus (MSVD)
  - 약 2000개의 비디오 클립
  - 각 비디오 당 10개의 캡션
- VGG
  - Visual Geometry Group

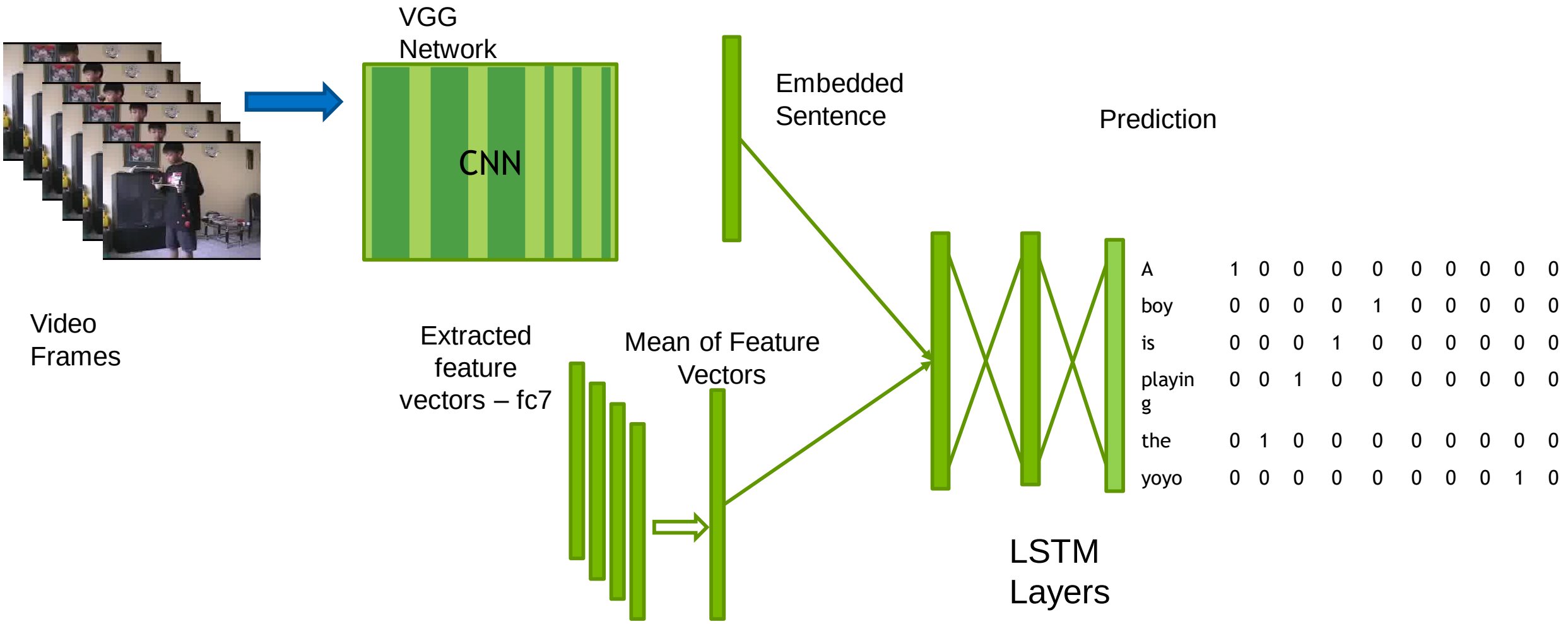
프로세스 - 비디오 자막 생성

1. 라이브러리 가져오기
2. 비디오 및 캡션 평가
    a. 단일 클립의 평균 벡터값 생성
    b. 이를 통해 fc7 레이어에서 각 프레임에 대한 높은 수준의 표현이 생성됨
    c. Lab 2에서의 작업과 유사
3. 캡션과 기능 맵 정렬
    a. 데이터의 서브셋으로 작동
4. 캡션의 다음 단어 예측
    a. Parse, tokenize 등

프로세스 - 비디오 자막 생성

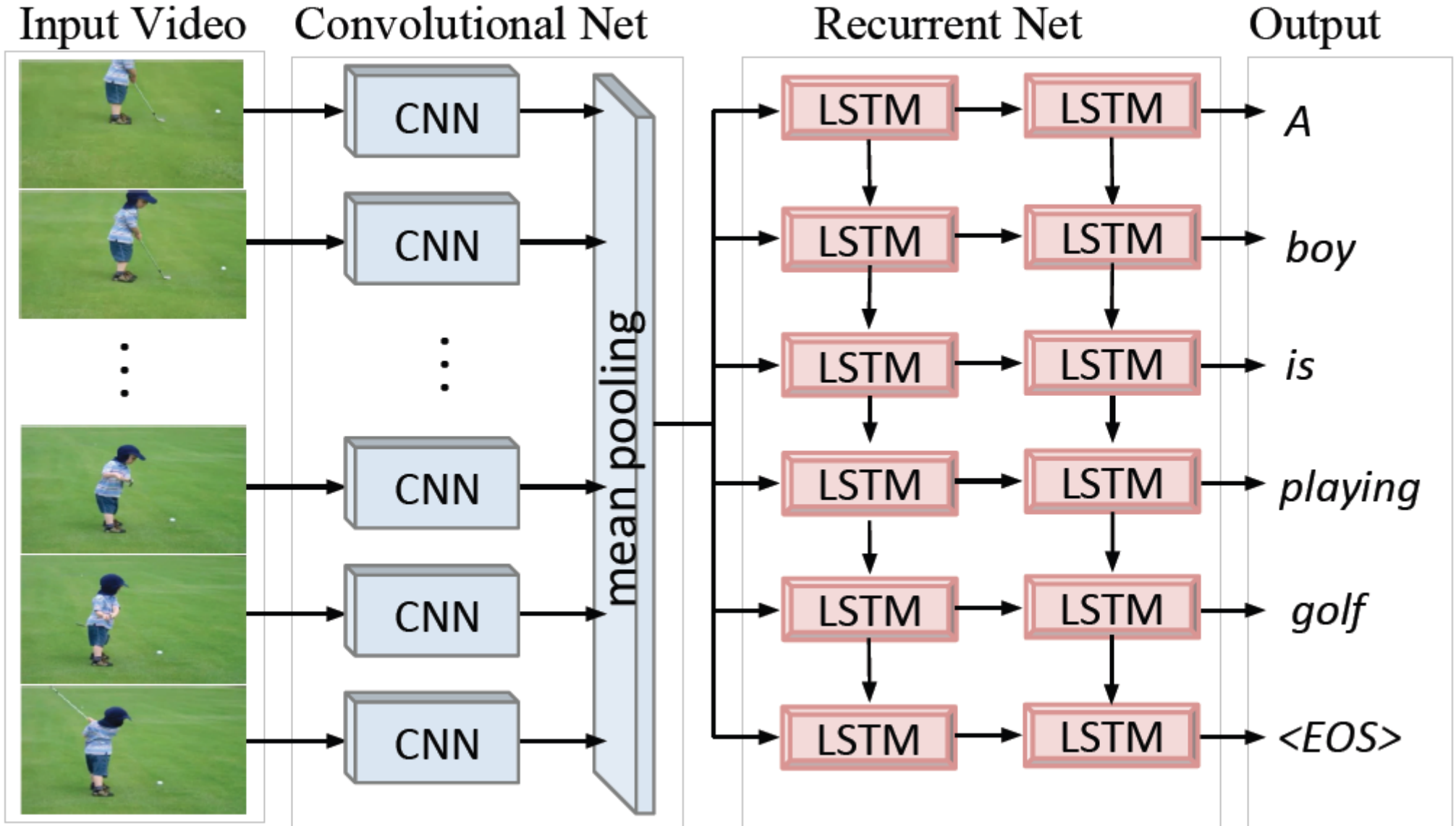
5. 네트워크 설계 (RNN)
    a. 참조: 코드 블락 수행 전 워딩 소스코드 수정
6. 모델 학습 및 개발
    a. 참조: 코드 블락 수행 전 워딩 소스코드 수정
7. 학습된 이미지 및 캡션 평가
8. 검증된 이미지의 캡션 생성

비디오 자막 생성



Translating Videos to Natural Language Using Deep Recurrent Neural Networks
Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko

LAB 4 - 비디오 자막 생성



프로세스 - 비디오 자막 생성

말 타는 한 남자

먹고 있는 동물

서 있는 개

MSVD - Collecting Highly Parallel Data for Paraphrase Evaluation
David L. Chen and William B. Dolan

Translating Videos to Natural Language Using Deep Recurrent Neural Networks
Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko

결 론

- 두 개의 논문을 기반으로 한 이미지 및 비디오 캡션 생성
  - 순환신경망을 이용한 자연 언어로 비디오 변환 Translating Videos to Natural Language Using Deep Recurrent Neural Networks(Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko)
  - 시각인식 및 설명을 위한 Long-term 순환 콘볼루션 네트워크 (Convolutional Networks)
- 이미지 및 비디오 캡션을 위한 여러 가지 접근 방식 존재 - 여기서는 하나만 사용됨
  - 이전과 같이 고정된 문장 템플릿을 사용하지 않고 시퀀스 생성 작업 모델링
    작업(Guadarama et al., 2013). 이미지 및 텍스트 데이터에 대한 사전 교육: 관련 데이터를
    자연스럽게 활용 (이전 슬라이드의 CNN과 LSTM연결 구성도)
    제한된 분량의 설명 자막을 통해 비디오를 보충하다.

www.nvidia.com/dli