

CS6501: Deep Learning for Visual Recognition

Open Topics

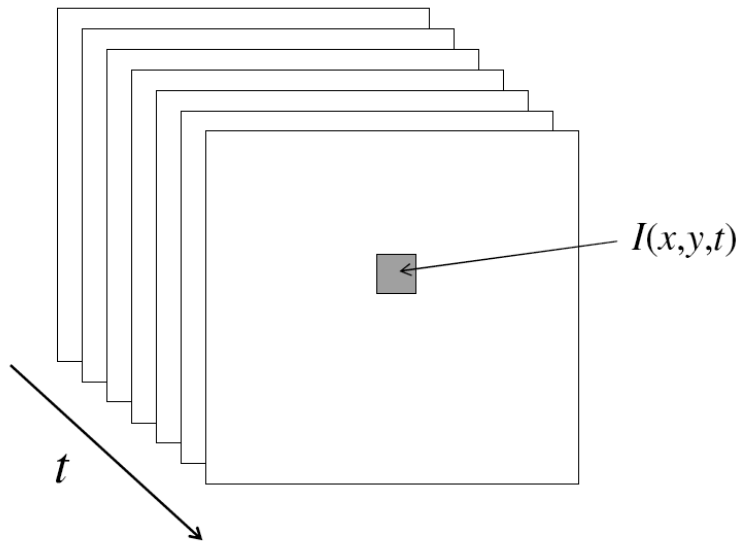


Today's Class

- Video: Optical flow, Two-Stream Networks, Tracking
- VQA: Datasets, Challenges, Models

From images to videos

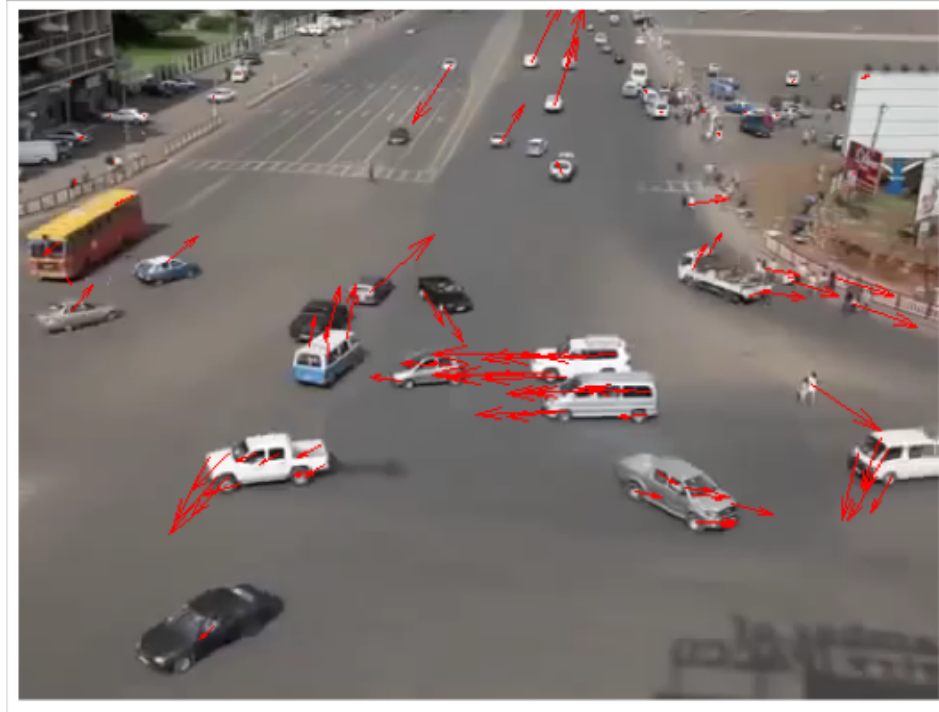
- A video is a sequence of frames captured over time
- Now our image data is a function of space (x, y) and time (t)



Why is motion useful?



Why is motion useful?

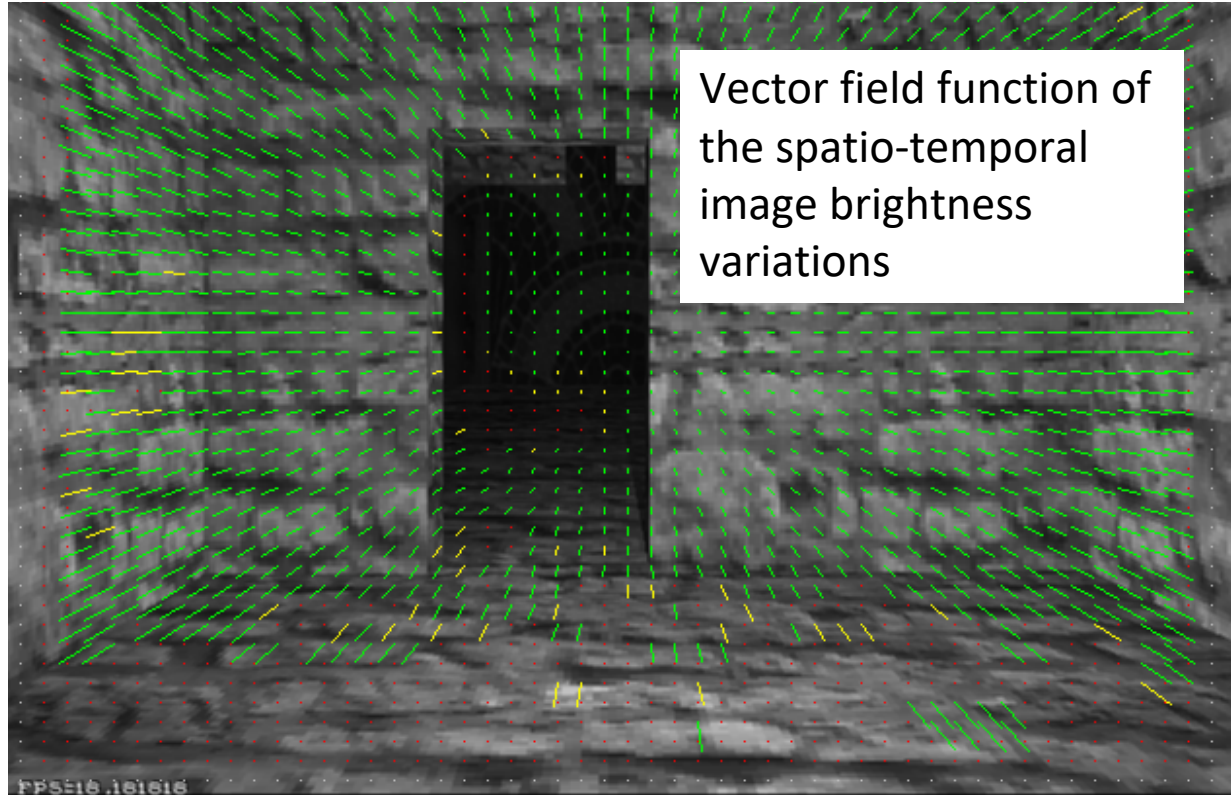


Optical flow

- Definition: optical flow is the *apparent* motion of brightness patterns in the image
- Note: apparent motion can be caused by lighting changes without any actual motion
 - Think of a uniform rotating sphere under fixed lighting vs. a stationary sphere under moving illumination

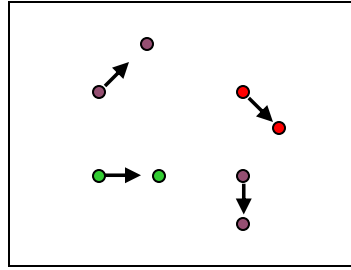
GOAL: Recover image motion at each pixel from optical flow

Optical flow

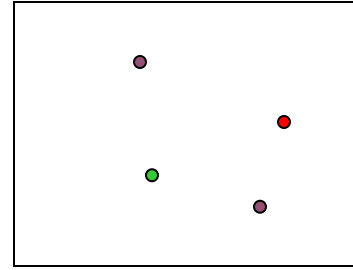


Picture courtesy of Selim Temizer - Learning and Intelligent Systems (LIS) Group, MIT

Estimating optical flow



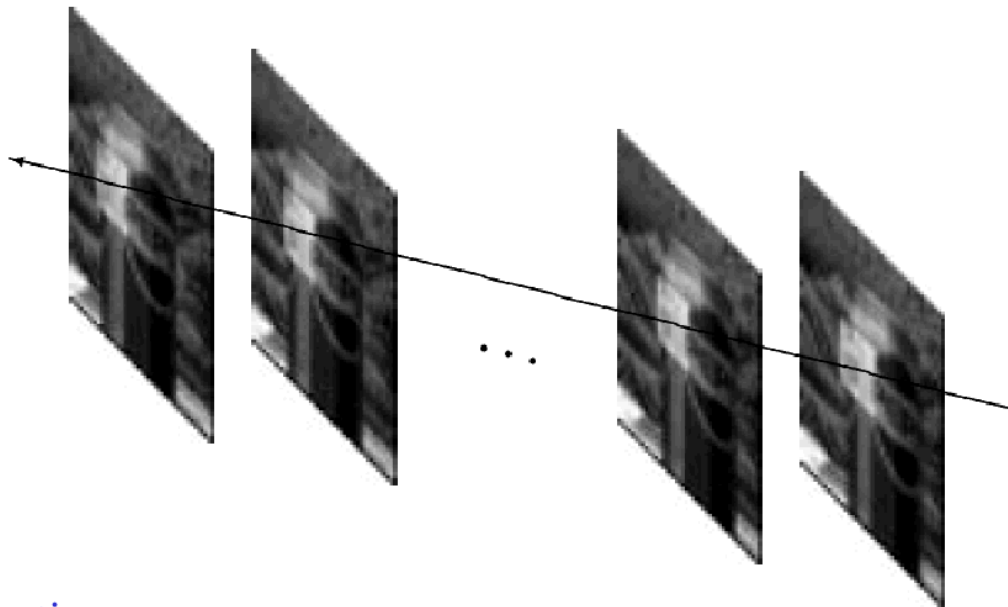
$I(x,y,t-1)$



$I(x,y,t)$

- Given two subsequent frames, estimate the apparent motion field $u(x,y)$, $v(x,y)$ between them
- Key assumptions
 - **Brightness constancy:** projection of the same point looks the same in every frame
 - **Small motion:** points do not move very far
 - **Spatial coherence:** points move like their neighbors

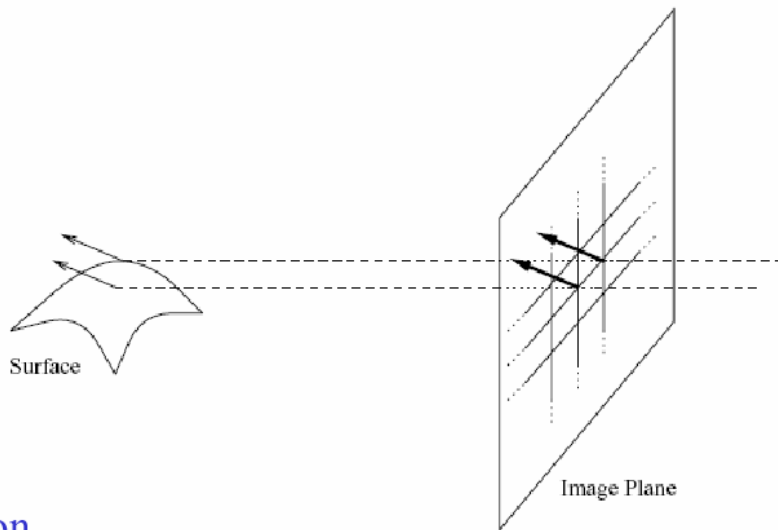
Key Assumptions: small motions



Assumption:

The image motion of a surface patch changes gradually over time.

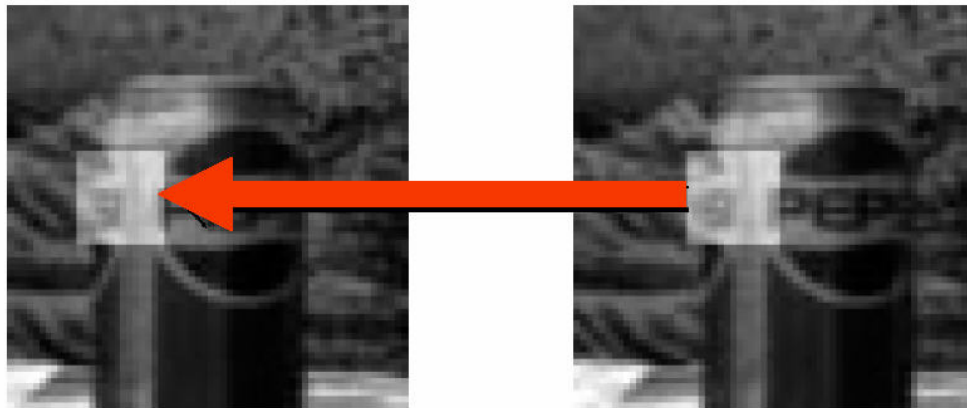
Key Assumptions: spatial coherence



Assumption

- * Neighboring points in the scene typically belong to the same surface and hence typically have similar motions.
- * Since they also project to nearby points in the image, we expect spatial coherence in image flow.

Key Assumptions: brightness Constancy



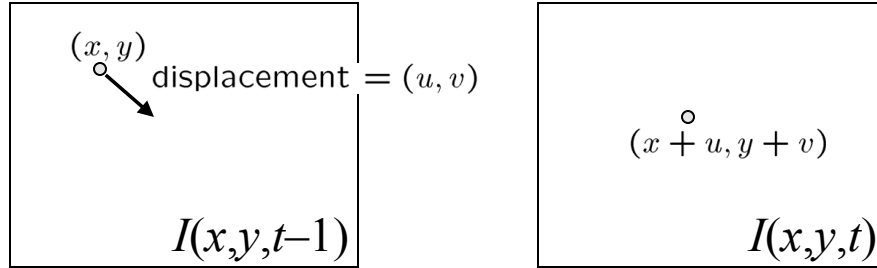
Assumption

Image measurements (e.g. brightness) in a small region remain the same although their location may change.

$$I(x + u, y + v, t + 1) = I(x, y, t)$$

(assumption)

The brightness constancy constraint



- Brightness Constancy Equation:

$$I(x, y, t - 1) = I(x + u(x, y), y + v(x, y), t)$$

Linearizing the right side using Taylor expansion:

$$I(x + u, y + v, t) \approx I(x, y, t - 1) + \overset{\text{Image derivative along x}}{I_x} \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$I(x + u, y + v, t) - I(x, y, t - 1) = I_x \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$\text{Hence, } I_x \cdot u + I_y \cdot v + I_t \approx 0 \quad \rightarrow \quad \nabla I \cdot [u \ v]^T + I_t = 0$$

The brightness constancy constraint

(x, y)
displacement = (u, v)

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

$$I(x+u, y+v, t) - I(x, y, t-1) = I_x \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$\text{Hence, } I_x \cdot u + I_y \cdot v + I_t \approx 0 \quad \rightarrow \quad \nabla I \cdot [u \ v]^T + I_t = 0$$

Action Classification from Video

Recommended Paper to Read:

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

João Carreira[†]
joaoluis@google.com

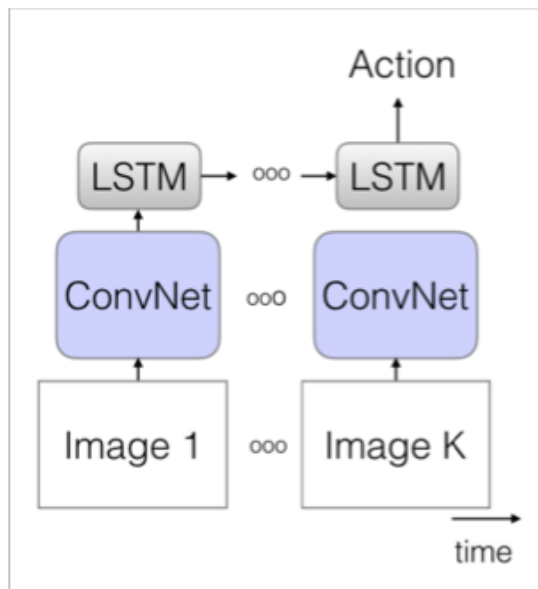
Andrew Zisserman^{†,*}
zisserman@google.com

[†]DeepMind

^{*}Department of Engineering Science, University of Oxford

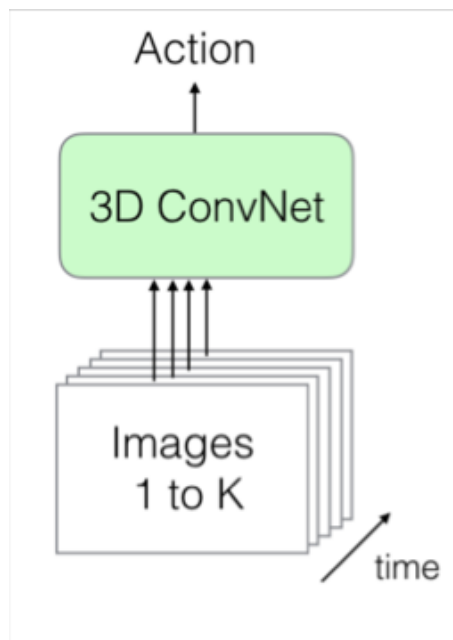
Action Classification from Video

CNN + LSTM over sequence of frames



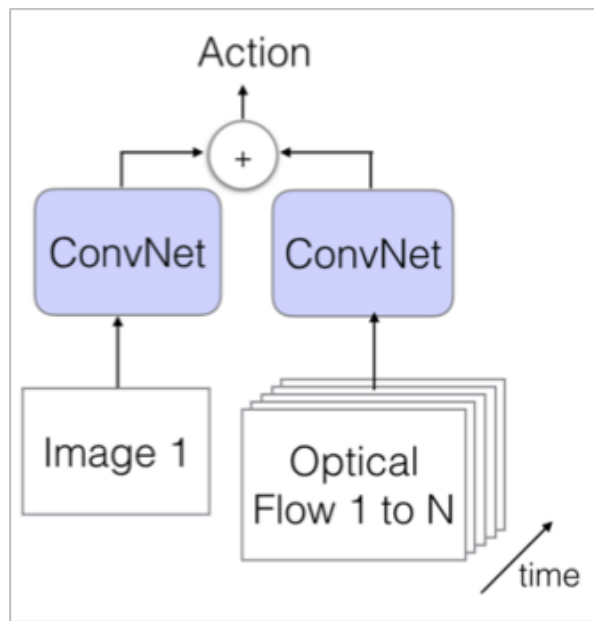
Action Classification from Video

3D CNN of consecutive frames across time



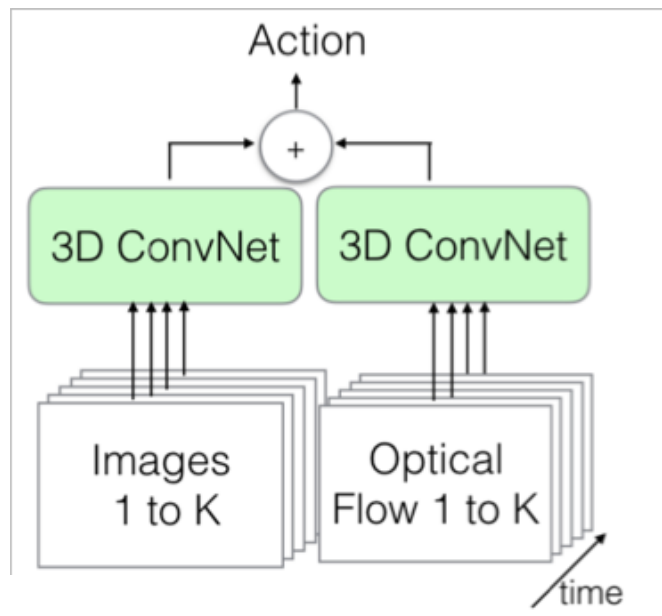
Action Classification from Video

Two Stream CNN: Images + Flow Map



Action Classification from Video

Two Stream 3D CNN: Images + Flow Map

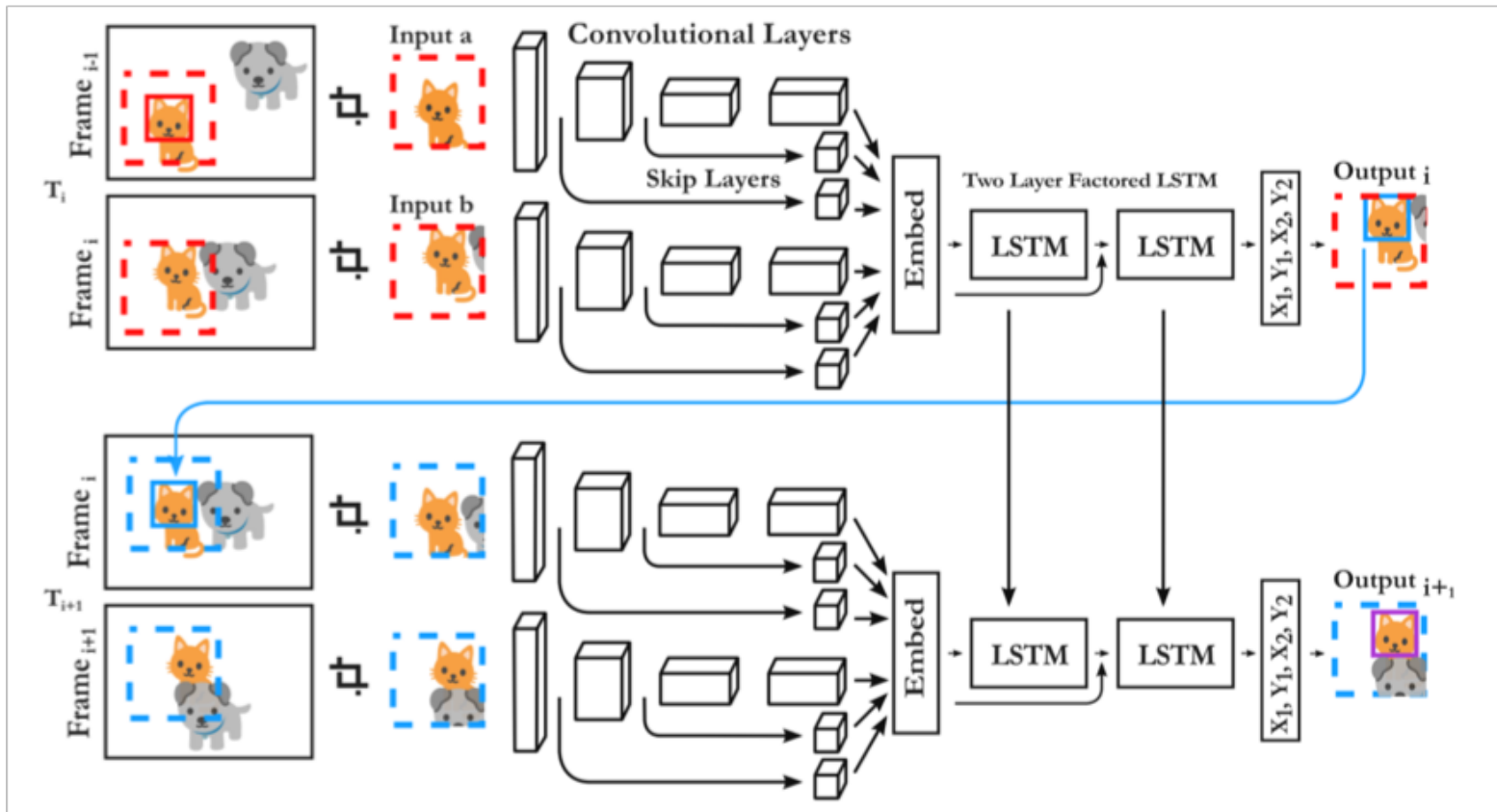


Action Classification from Video

Results on UCF101 actions

Architecture	UCF-101		
	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–
(b) 3D-ConvNet	51.6	–	–
(c) Two-Stream	83.6	85.6	91.2
(d) 3D-Fused	83.2	85.8	89.3
(e) Two-Stream I3D	84.5	90.6	93.4

Some Deep Learning (non Flow)-based tracker



Visual Question Answering



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

Visual Question Answering



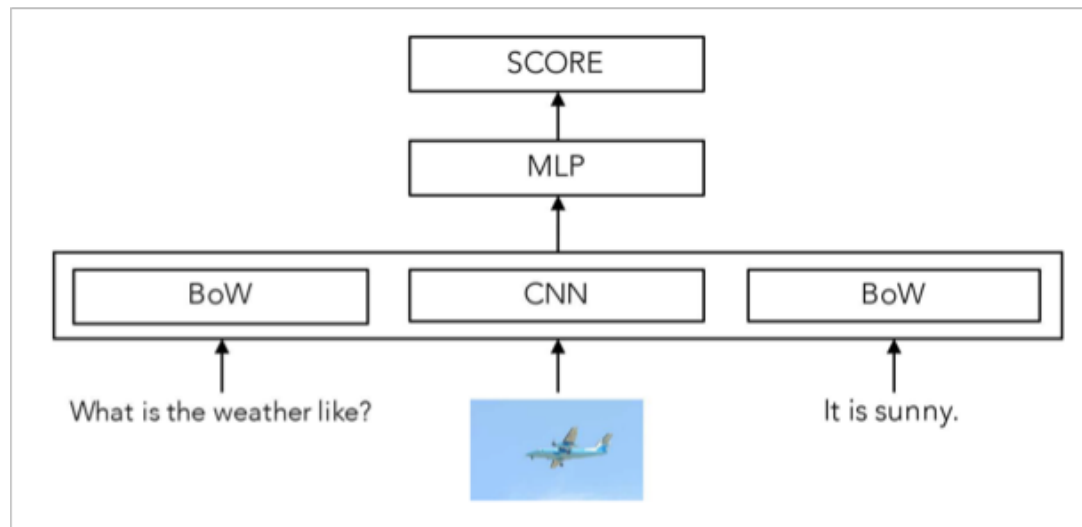
What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

Challenges? Pitfalls?

Visual Question Answering: Simplest (but effective) Model

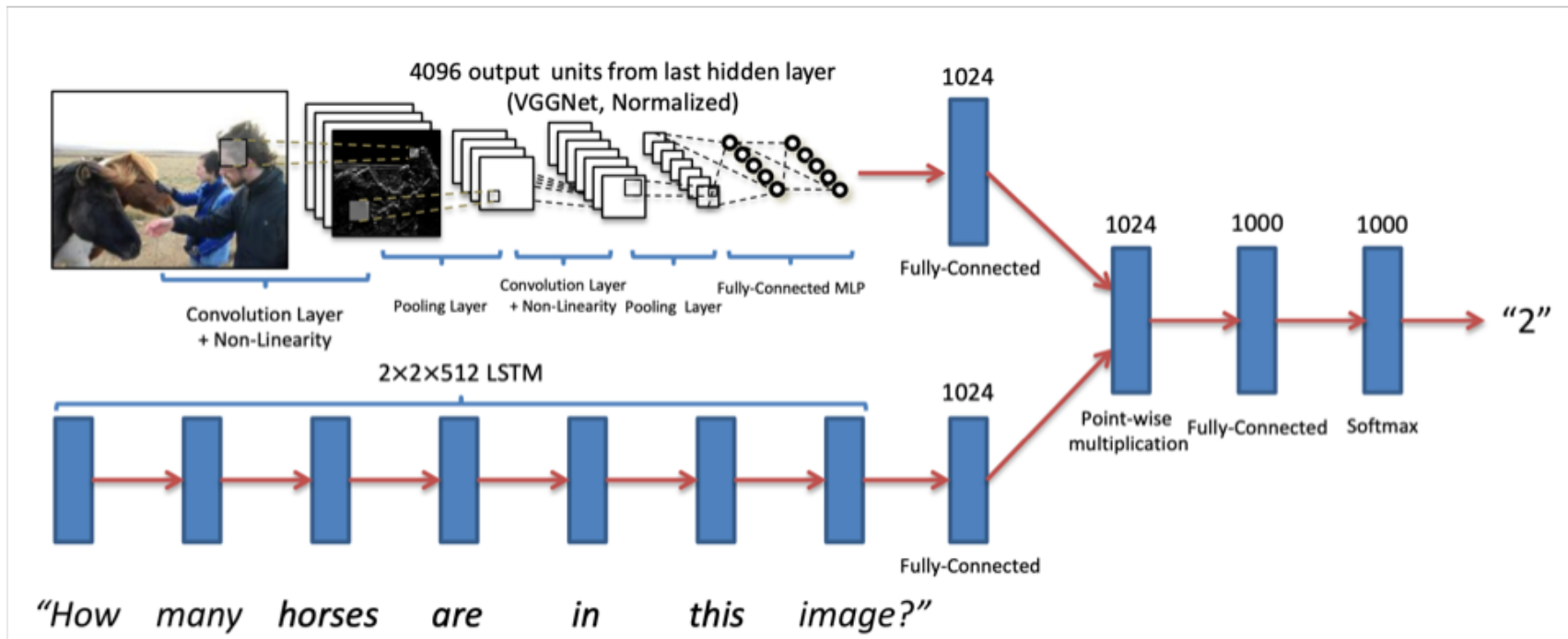


Revisiting Visual Question Answering Baselines

Allan Jabri, Armand Joulin, and Laurens van der Maaten

Facebook AI Research
{ajabri,ajoulin,lvdmaaten}@fb.com

Open Ended Questions Model

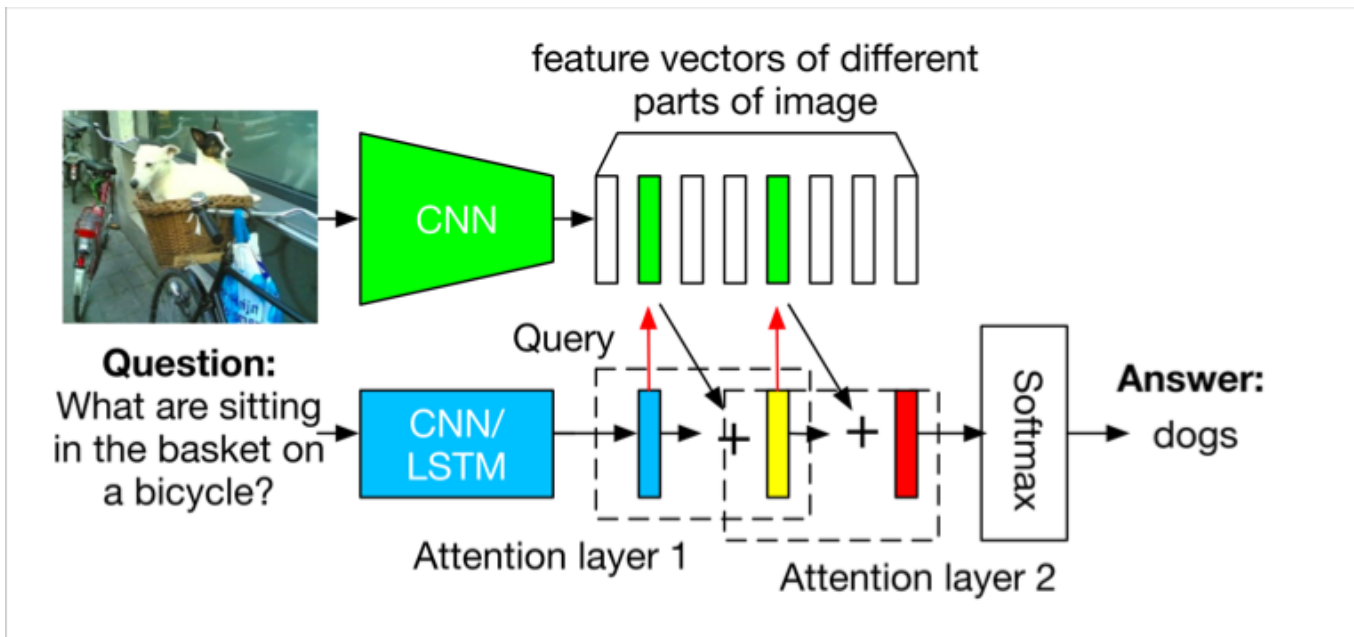


VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Open Ended Questions



Memory Networks (QA but text-only)

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.

Q: Where is the apple?

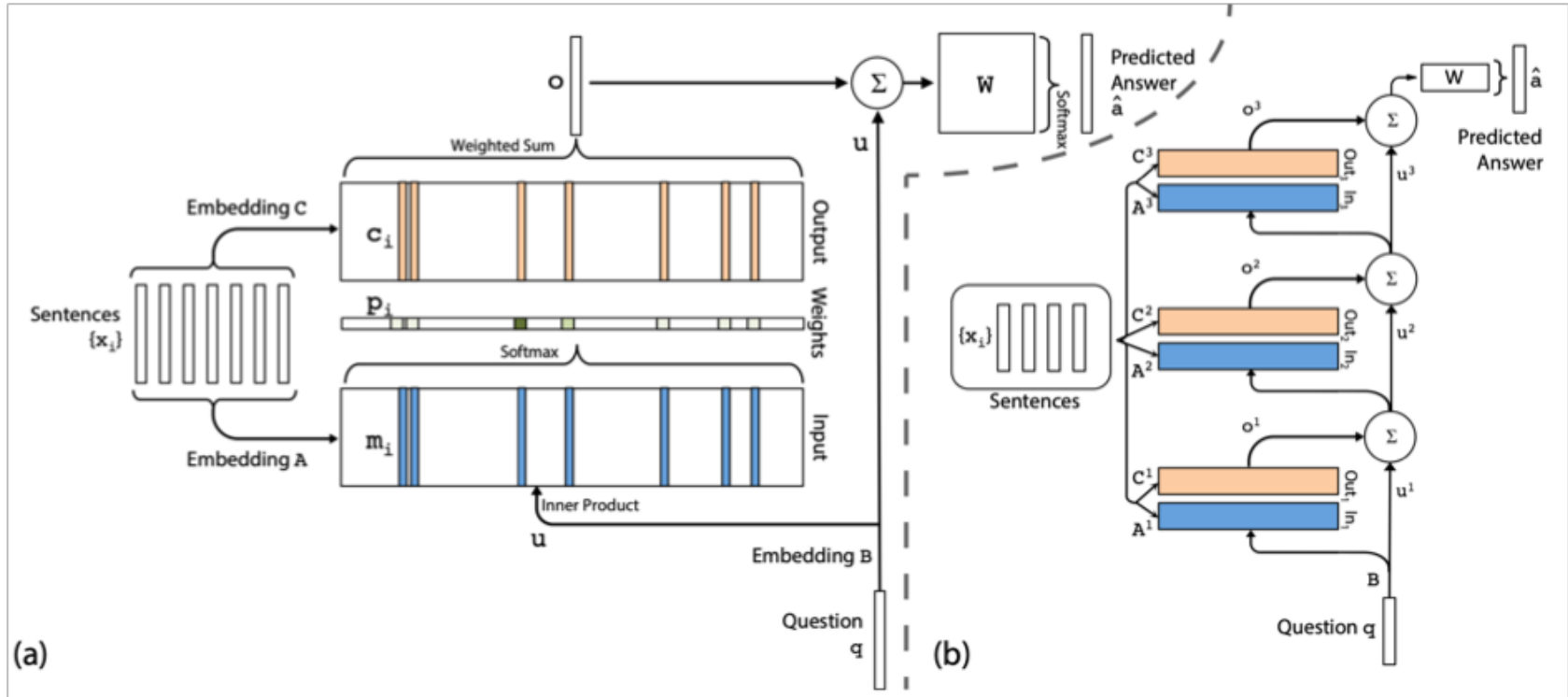
A. Bedroom

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.

Q: What color is Brian?

A. White

Memory Networks



Memory Networks for VQA

Dynamic Memory Networks for Visual and Textual Question Answering

Caiming Xiong*, Stephen Merity*, Richard Socher
MetaMind, Palo Alto, CA USA

{CMXIONG, SMERITY, RICHARD}@METAMIND.IO
*indicates equal contribution.

Learning Visual Knowledge Memory Networks for Visual Question Answering

Zhou Su^{1*}, Chen Zhu², Yinpeng Dong³, Dongqi Cai⁴, Yurong Chen⁴, Jianguo Li⁴
¹Tencent Wechat, ²ShanghaiTech University, ³Tsinghua University, ⁴Intel Labs China
zhousu@tencent.com, zhuchen@shanghaitech.edu.cn, dypl7@mails.tsinghua.edu.cn
{dongqi.cai, yurong.chen, jianguo.li}@intel.com

Questions?