Announcements

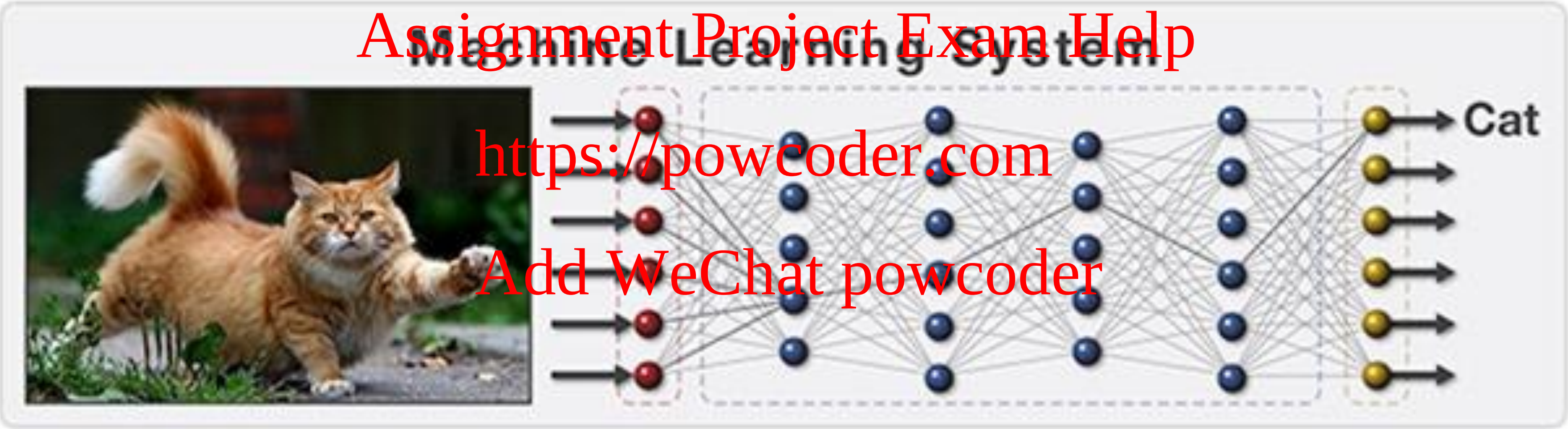
Reminder: Class challenge out! Ends December 10th

- Lab this week – go over pset6 solutions, tips for challenge

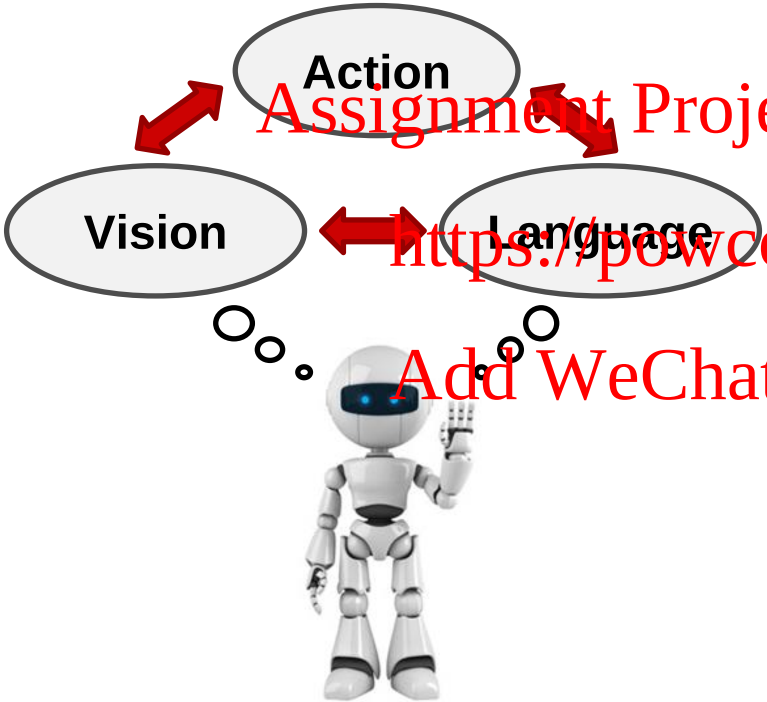
Vision & Language Applications

Slides adapted from Kate Saenko
Machine Learning

so far...

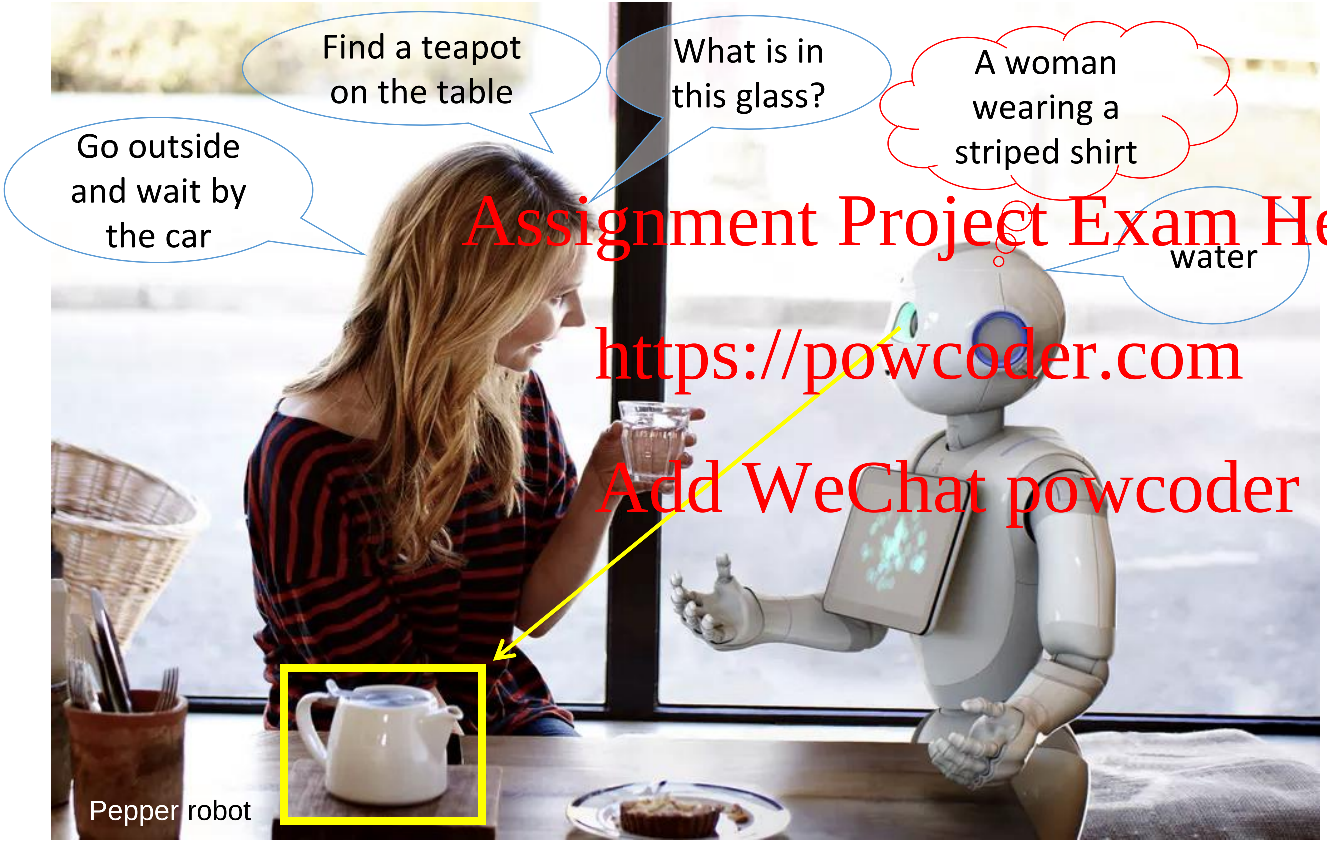


General AI: machines that *see*, *talk*, *act*



- Social media analysis
- Security and smart cameras
- AI assistants
- Helper robots for the elderly
- etc…

More Natural Human-Machine Interaction



- Description
- Visual question answering (VQA)
- Referring expression (REF)
- Instruction following / navigation
- …

Vision & Language problems

A baseball game in progress with the batter up to plate

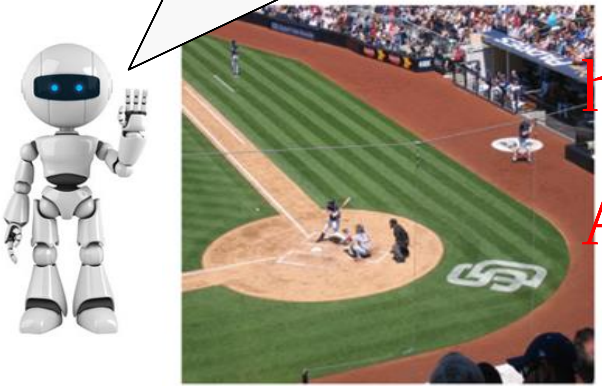


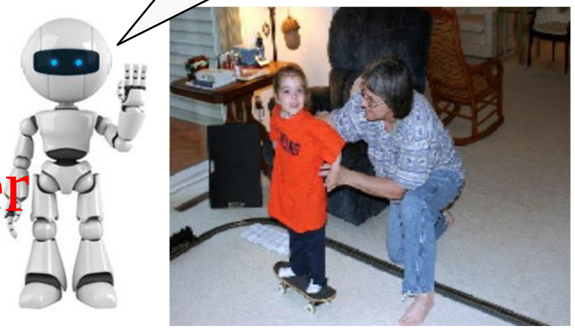
Image captioning

A man is riding a bicycle

Video captioning

Q: What is the child standing on?

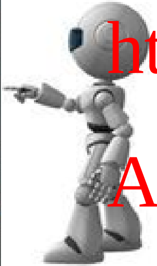
A: skateboard



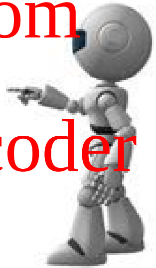
Visual Question Answering

Vision & Language problems

Find “window upper right”



Find the moment when “girl looks up at the camera and smiles”



...and many others...

Referring expressions

Text-to-clip retrieval from video

Demos

https://www.captionbot.ai/

http://vqa.cloudcv.org/

Today: Vision & Language

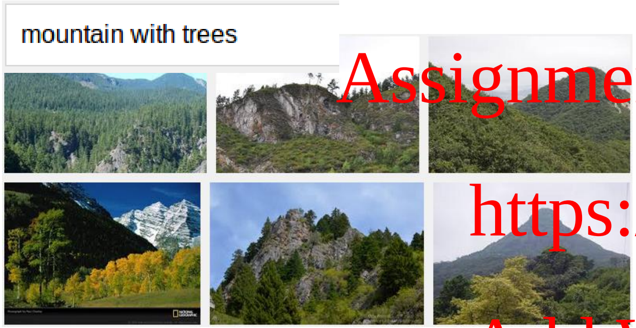
- Video captioning—in detail
- Other tasks
  - Visual question answering (VQA)
  - Video clip search
  - Following instructions to navigate

Video Captioning

Kate Saenko
Machine Learning

Applications of video captioning

Image and video retrieval by content.



Video description service.

Human Robot Interaction

Video surveillance

Image Captioning, B.D. (before deep learning)

Language: Increasingly focused on **grounding** meaning in perception.
Vision: Exploit linguistic ontologies to **"tell a story"** from images.

[Farhadi et. al. ECCV'10]

(animal, stand, ground)

[Kulkarni et. al. CVPR'11]

There are one cow and one sky.
The golden cow is by the blue sky.

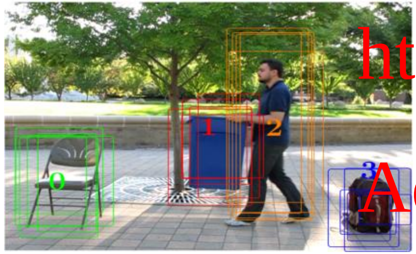
Many early works on Image Description
Farhadi et. al. ECCV'10, Kulkarni et. al. CVPR'11, Mitchell et. al. EACL'12, Kuznetsova et. al. ACL'12 & ACL'13

Identify objects and attributes, and combine with linguistic knowledge to "tell a story".

Dramatic increase in interest since then.
(8 papers in CVPR'15)

Video Description, B.D. (before deep learning)

[Krishnamurthy, et al. AAAI'13]



[Yu and Siskind, ACL'13]

[Rohrbach et. al. ICCV'13]

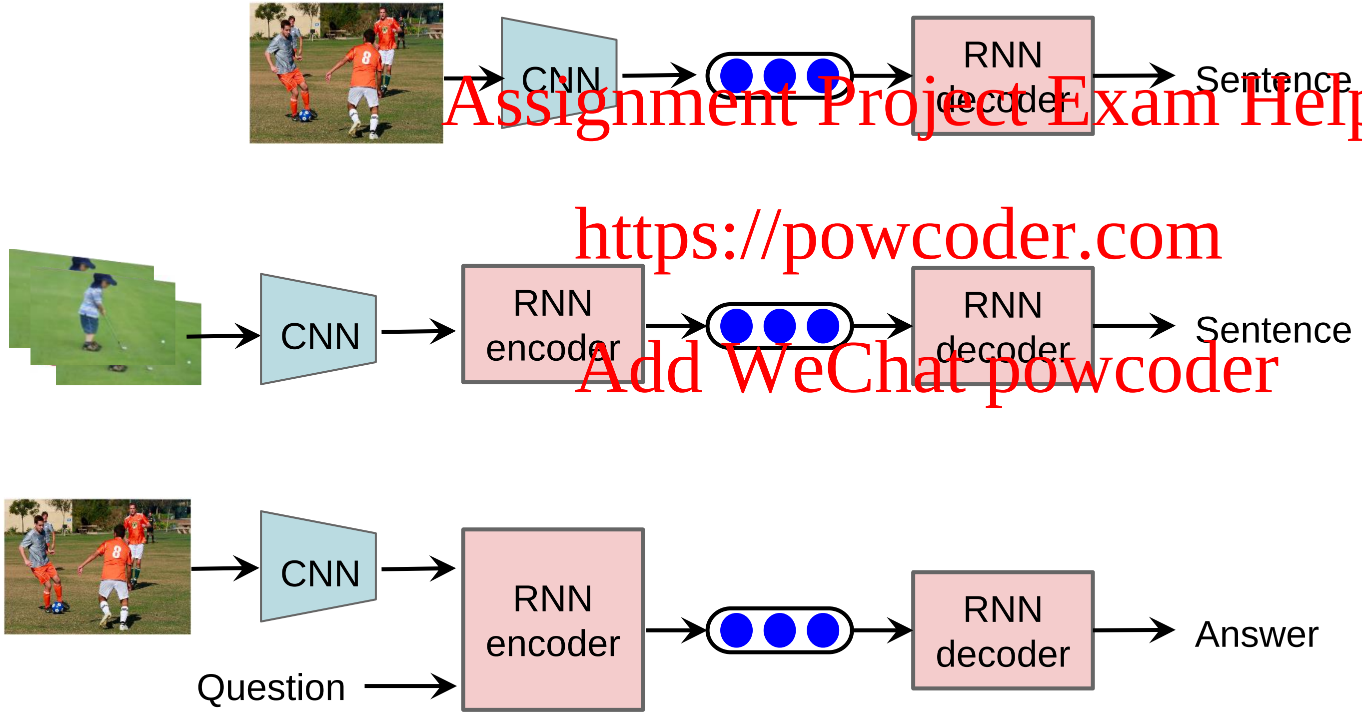
- Extract object and action descriptors.
- Learn object, action, scene classifiers.
- Use language to bias visual interpretation.
- Estimate most likely agents and actions.
- Template to generate sentence.

Others: Guadarrama ICCV'13, Thomason COLING'14

Limitations:

- Narrow Domains
- Small Grammars
- Template based sentences
- Several features and classifiers

After Deep Learning, A.D.:
End-to-End Neural Models based on Recurrent Nets

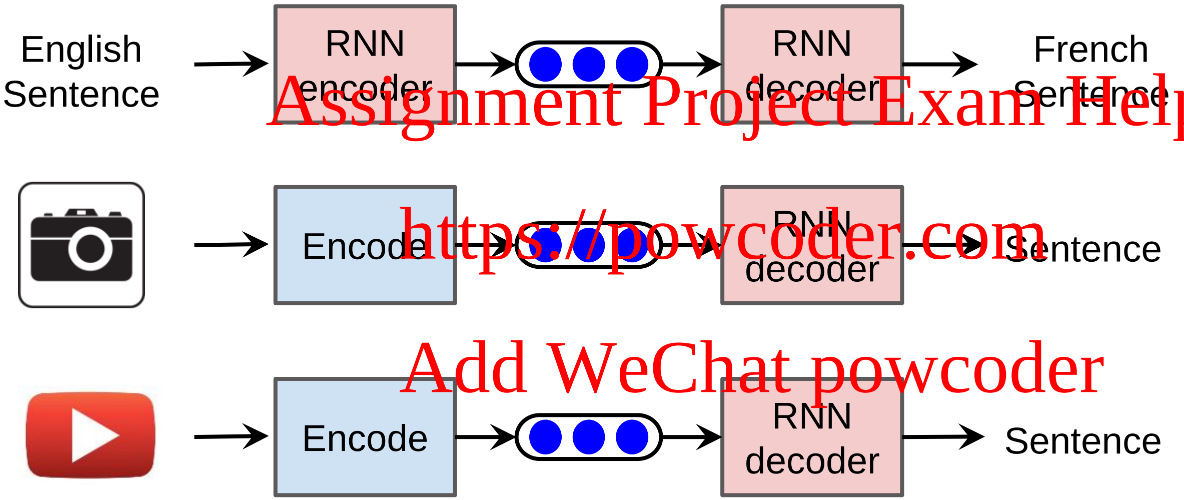


[Donahue et al. CVPR’15]
[Vinyals et al. CVPR’15]

[Venugopalan et. al. ICCV’15]

[Malinowski et. al. ICCV’15]

Recurrent Neural Networks (RNNs) can map a vector to a sequence.



[Sutskever et al. NIPS'14]

[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

[Venugopalan et. al. NAACL'15]

Key Insight:

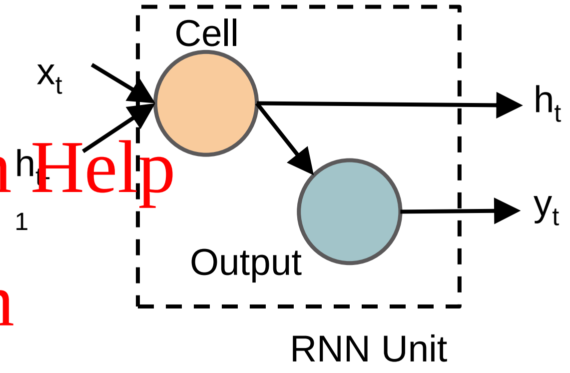
Generate feature representation of the video and "decode" it to a sentence

[review] Recurrent Neural Networks

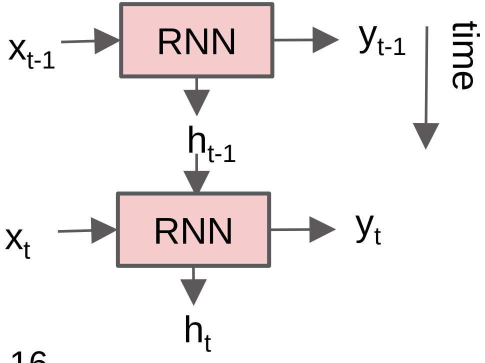
Successful in translation, speech.

RNNs can map an input to an output sequence.

$$\Pr(\text{out } y_t \mid \text{input, out } y_0 \ldots y_{t-1})$$



Insight: Each time step has a layer with the same weights.



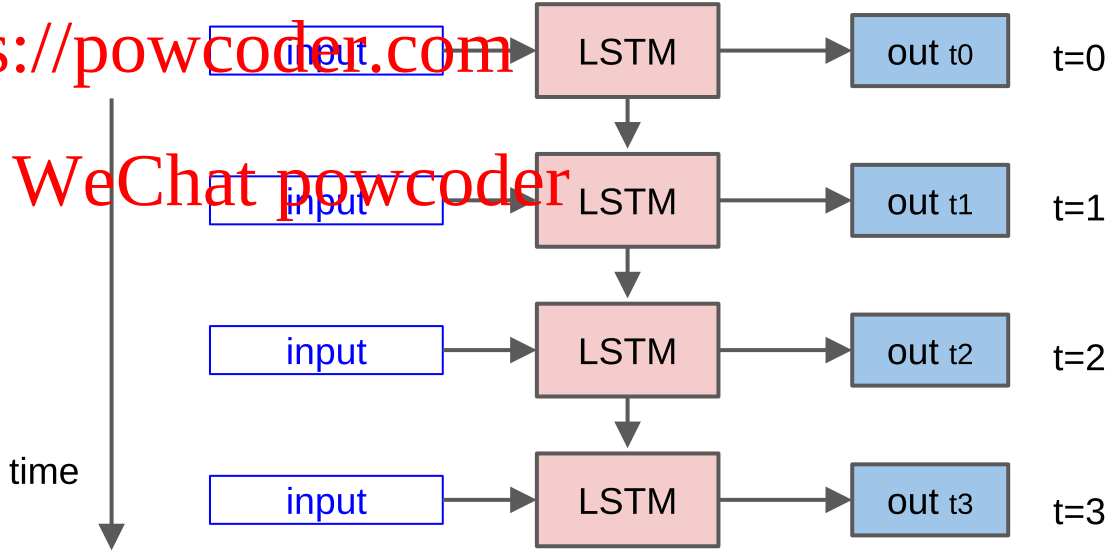
Problems:
1. Hard to capture long term dependencies
2. Vanishing gradients (shrink through many layers)

Solution: Long Short Term Memory (LSTM) unit

LSTM Sequence decoders

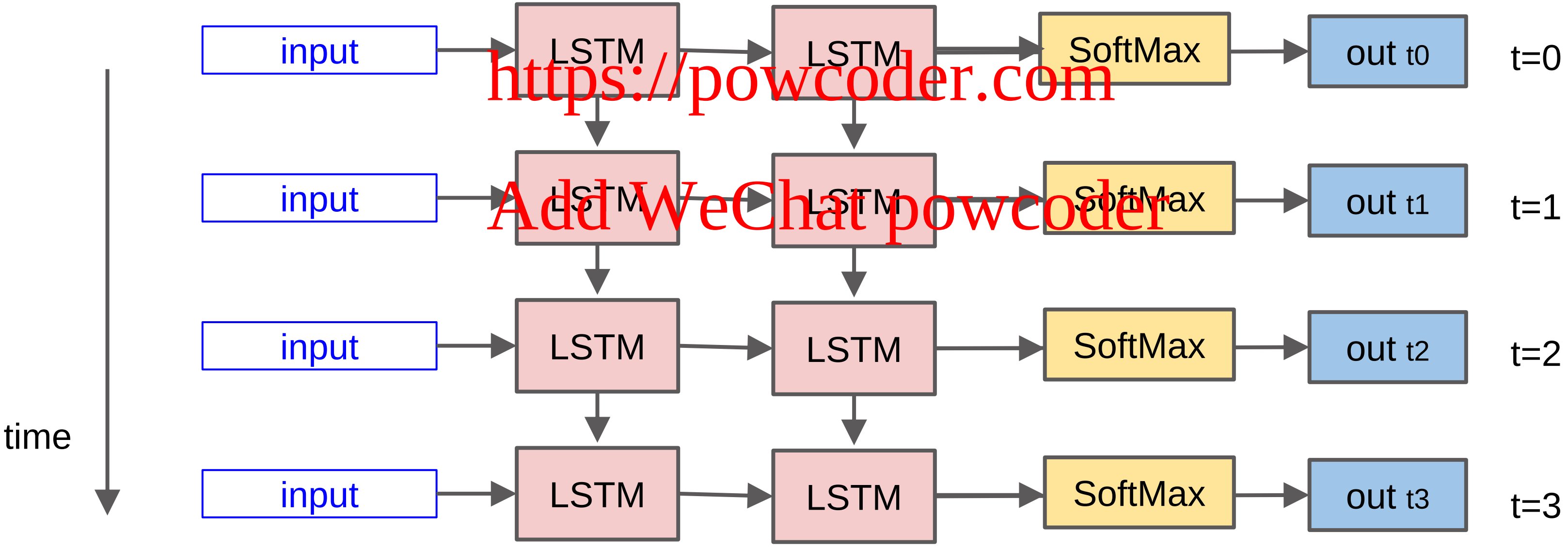
Functions are differentiable.
Full gradient is computed by backpropagating through time.
Weights updated using Stochastic Gradient Descent.

Matches state-of-the-art on:
Speech Recognition
[Graves & Jaitly ICML'14]
Machine Translation (Eng-Fr)
[Sutskever et al. NIPS'14]
Image-Description
[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

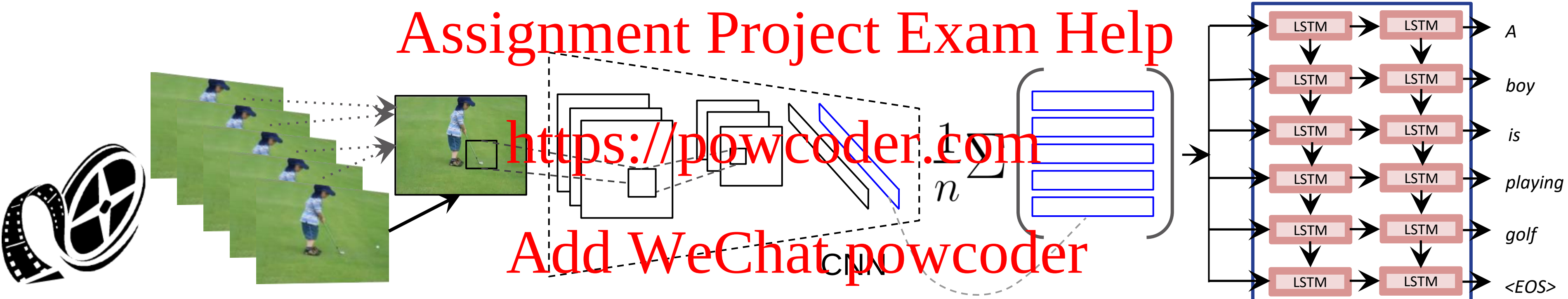


LSTM Sequence decoders

Two LSTM layers - 2nd layer of depth in temporal processing.
Softmax over the vocabulary to predict the output at each time step.

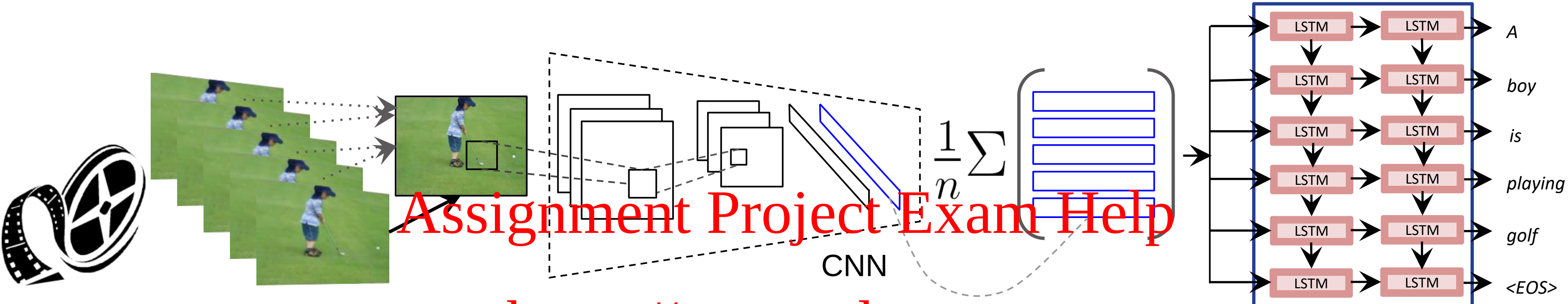


Translating Videos to Natural Language



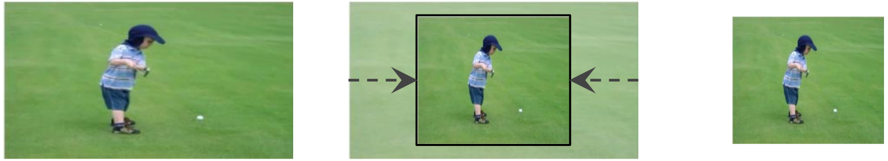
[Venugopalan et. al. NAACL'15]

Test time: Step 1



(a)

Input Video → Sample frames @1/10



227x227

Frame → Scale

(b)

[review] Convolutional Neural Networks (CNNs)

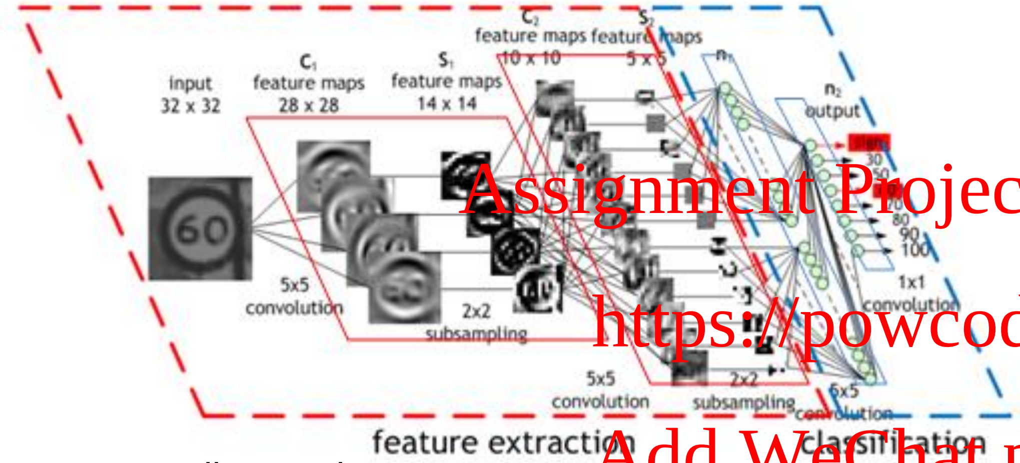


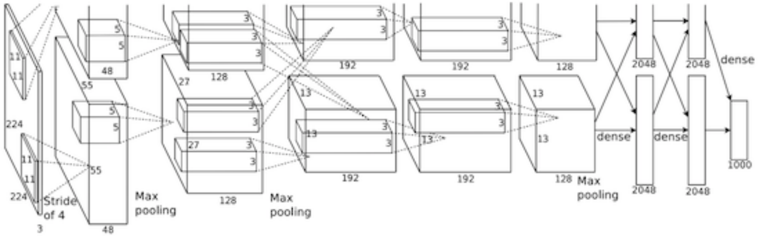
Image Credit: Maurice Peeman

Successful in semantic visual recognition tasks.

Layer - linear filters followed by non linear function. Stack layers.

Learn a hierarchy of features of increasing semantic richness.

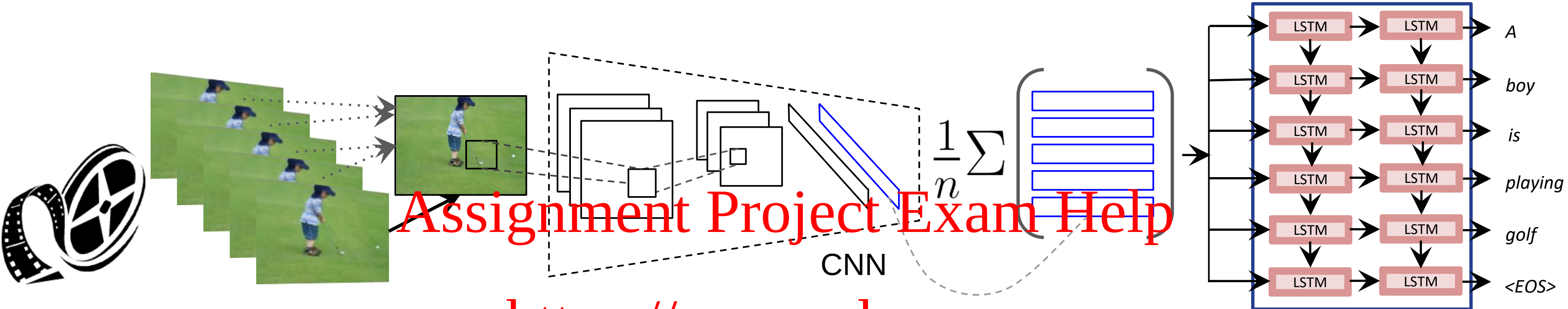
>>

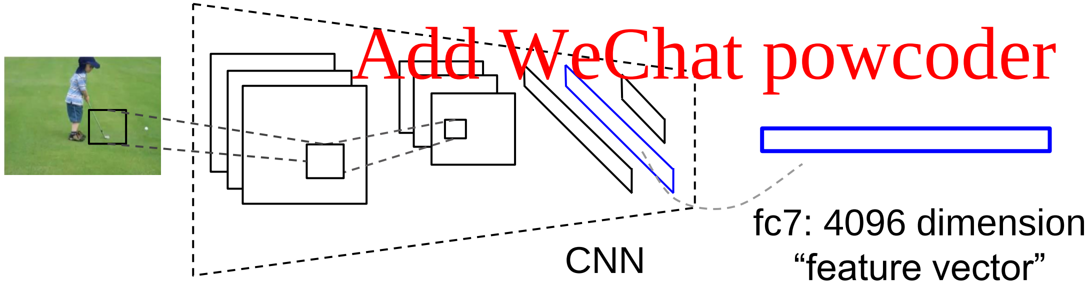


Krizhevsky, Sutskever, Hinton 2012

ImageNet classification breakthrough

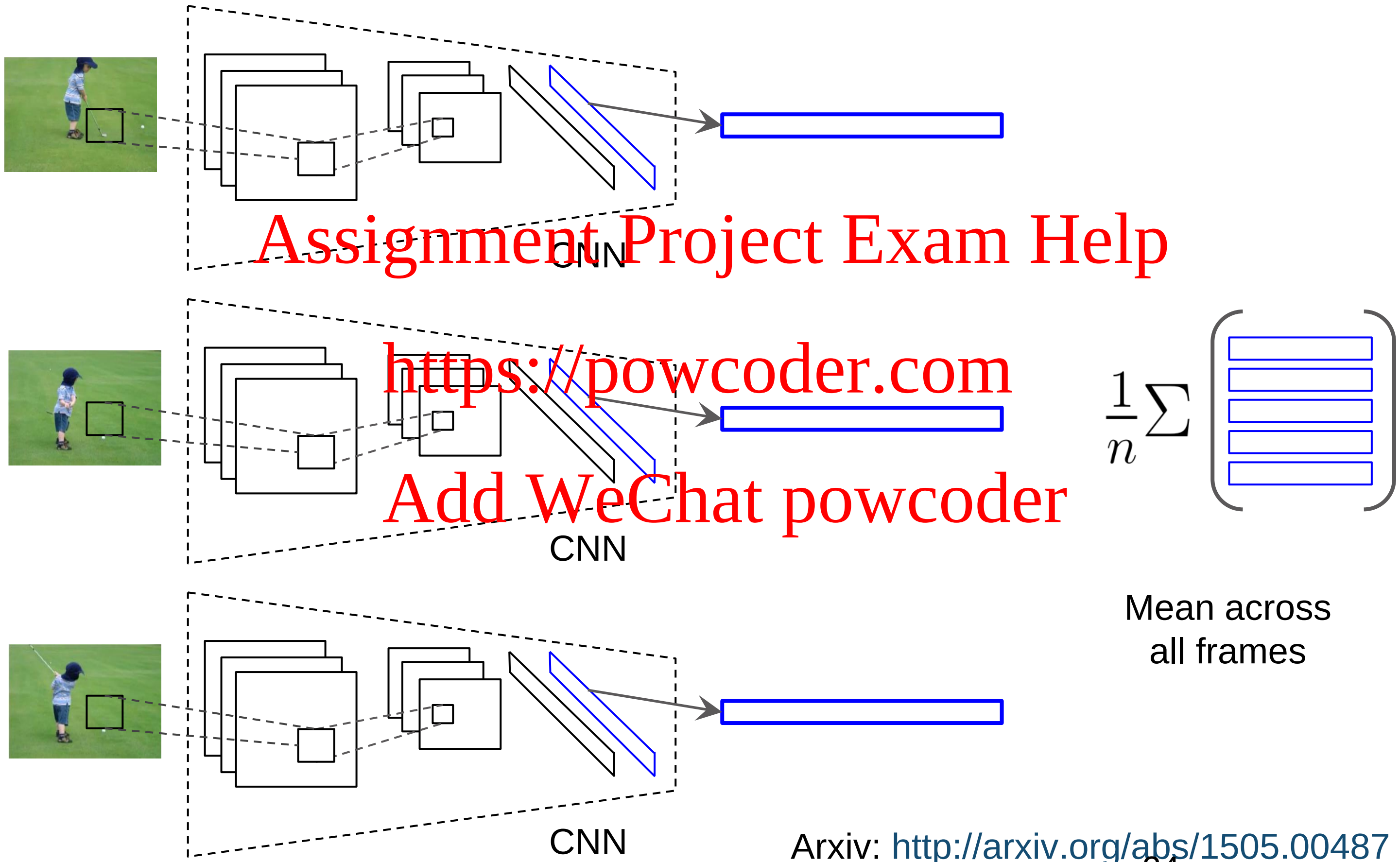
Test time: Step 2 Feature extraction





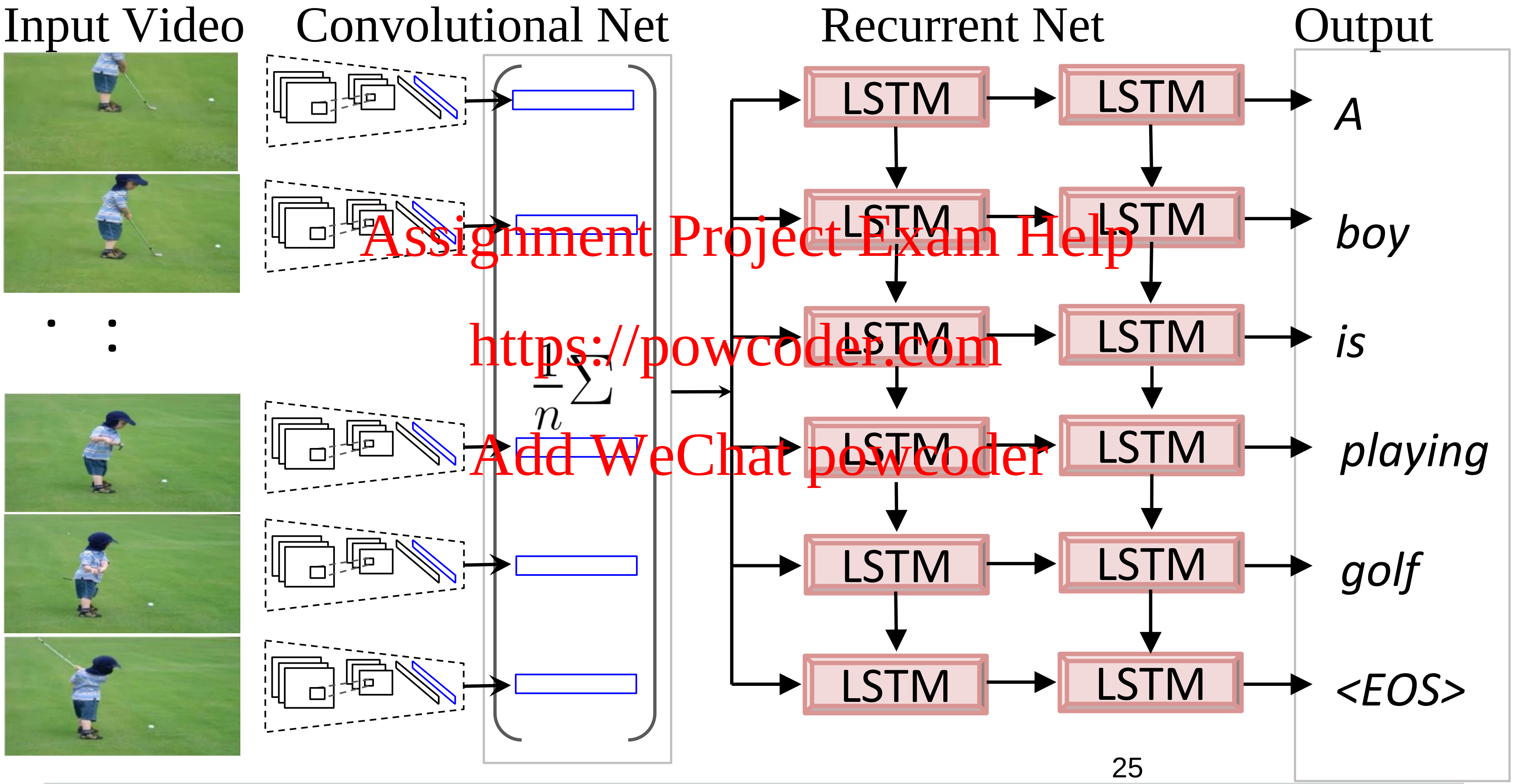
Forward propagate
Output: “fc7” features
(activations before classification layer)

Test time: Step 3 Mean pooling

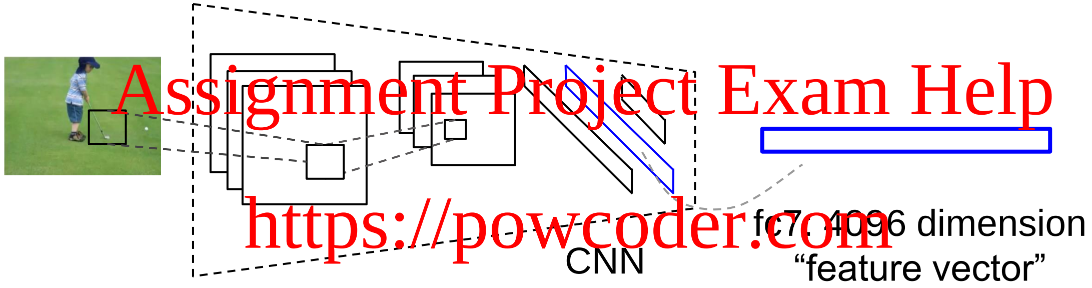


Arxiv: http://arxiv.org/abs/1505.00487

Test time: Step 4 Generation

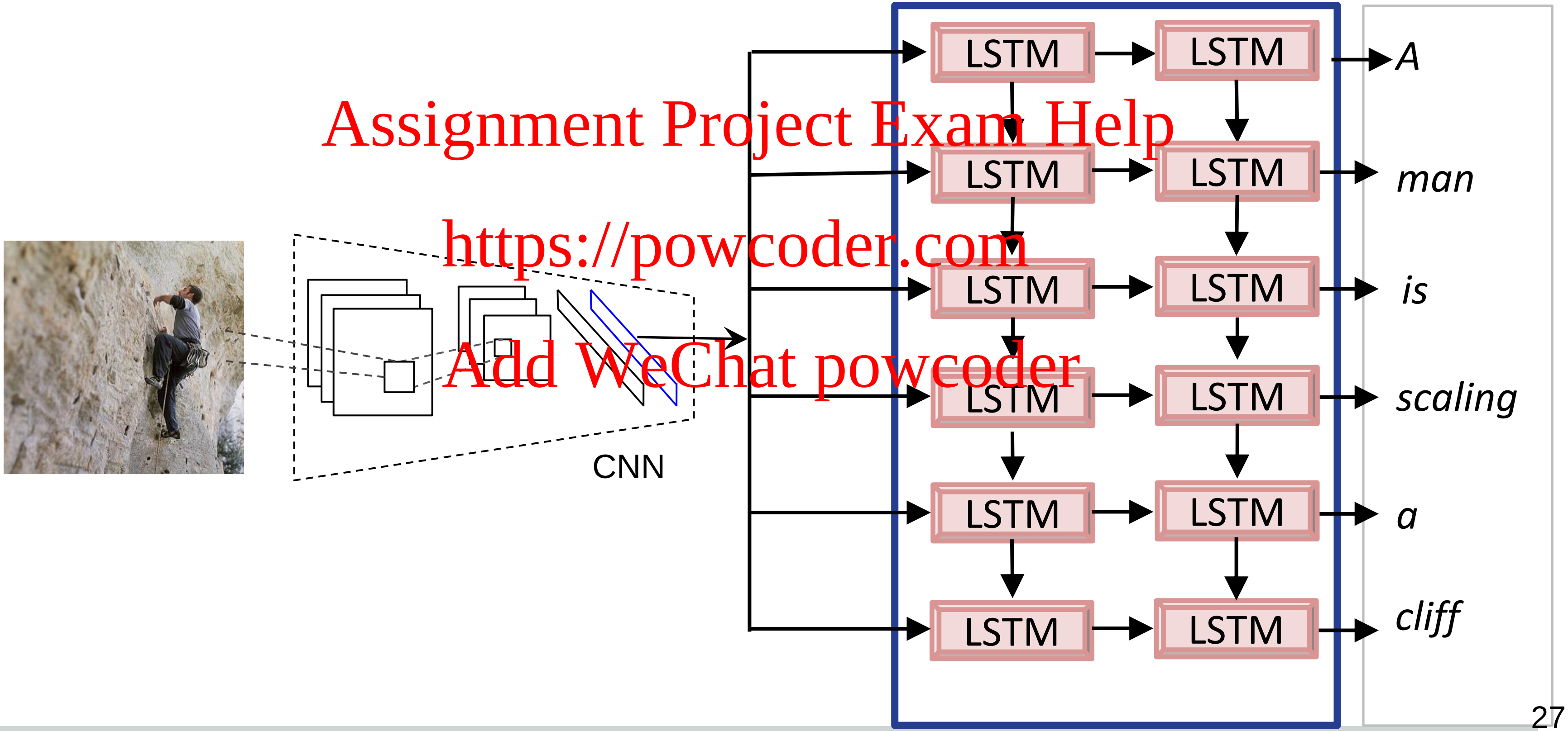


Step1: CNN pre-training

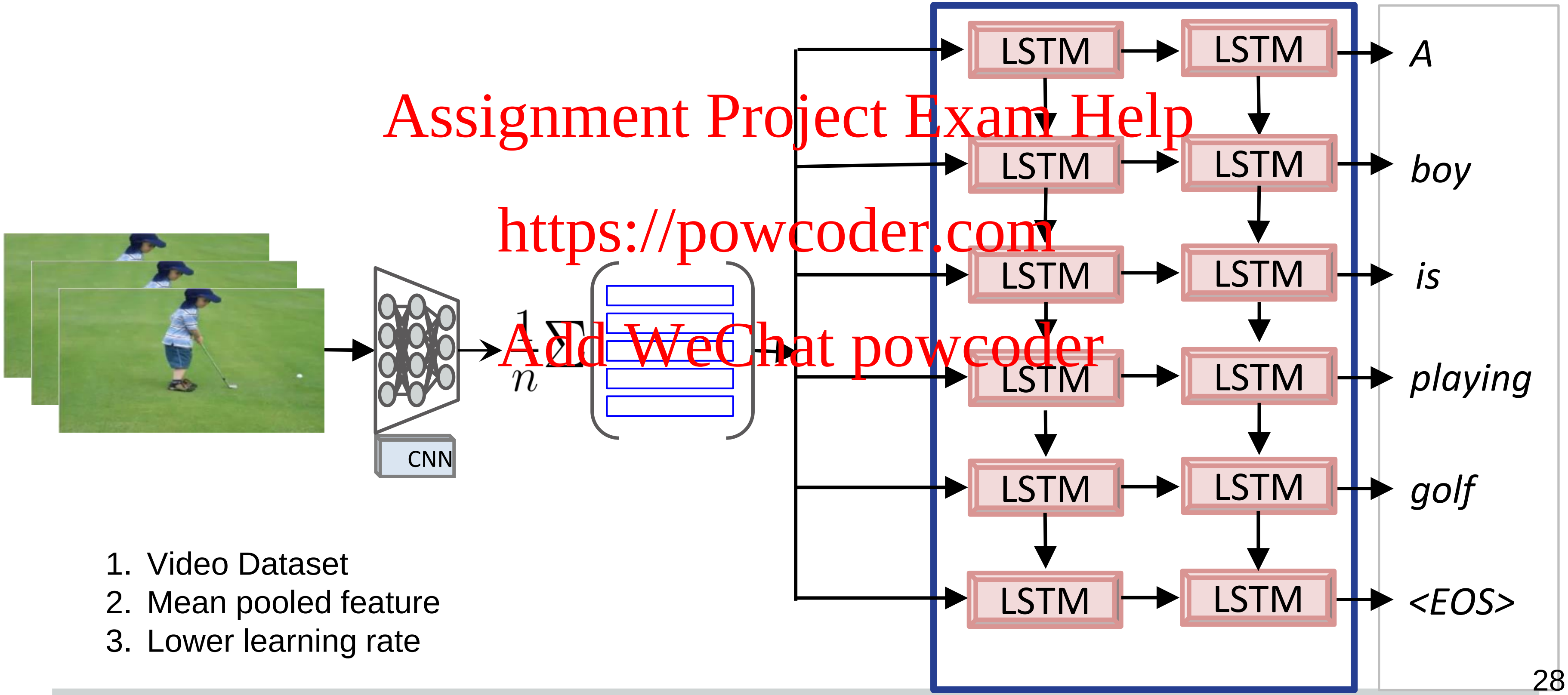


- Based on Alexnet [Krizhevsky et al. NIPS'12]
- 1.2M+ images from ImageNet ILSVRC-12 [Russakovsky et al.]
- Initialize weights of our network.

Step2: Image-Caption training



Step3: Fine-tuning



1. Video Dataset
2. Mean pooled feature
3. Lower learning rate

Experiments: Dataset

Microsoft Research Video Description dataset **[Chen & Dolan, ACL'11]**

Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

1970 YouTube video snippets

- 10-30s each
- typically single activity
- no dialogues
- 1200 training, 100 validation, 670 test

Annotations

- Descriptions in multiple languages
- ~40 English descriptions per video
- descriptions and videos collected on AMT

Sample video and gold descriptions

- A man appears to be **plowing** a rice field with a plow being pulled by two **oxen**.
- A team of **water buffalo pull** a plow through a rice paddy.
- Domesticated **livestock** are helping a man **plow**.
- A man **leads** a team of oxen down a muddy path.
- Two **oxen walk** through some mud.
- A man is **tilling** his land with an **ox pulled** plow.
- **Bulls** are **pulling** an object.
- Two **oxen** are **plowing** a field.
- The farmer is **tilling** the soil.
- A man in **ploughing** the field.

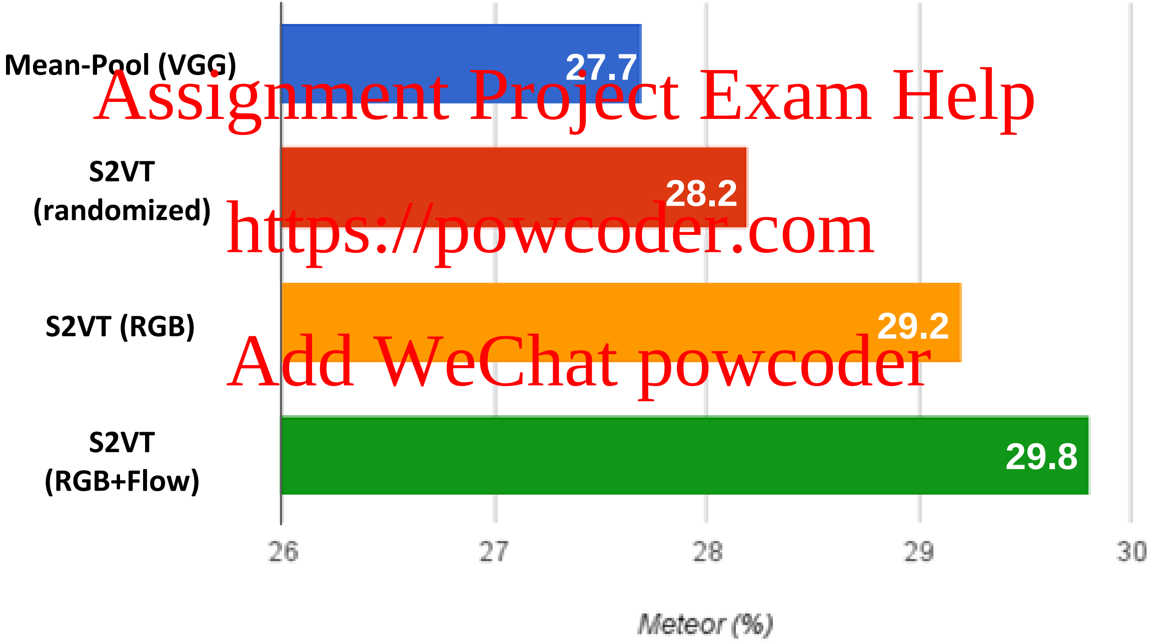
- A man is **walking** on a **rope**.
- A man is **walking** across a **rope**.
- A man is **balancing** on a **rope**.
- A man is **balancing** on a **rope** at the beach.
- A man **walks** on a **tightrope** at the beach.
- A man is **balancing** on a **volleyball net**.
- A man is **walking** on a **rope** held by poles
- A man **balanced** on a **wire**.
- The man is **balancing** on the **wire**.
- A man is **walking** on a **rope**.
- A man is **standing** in the sea shore.

Evaluation

Machine Translation Metrics
- BLEU
- METEOR

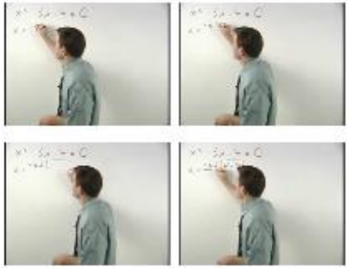
Human evaluation

Results (Youtube)



Example outputs

FGM: A person is playing a guitar in the house.
YT: A group of performing on stage.
YT_C: A man is doing a trick.
YT_CF: **A man is jumping on a pole.**
GT: Two men working on a high building.



FGM: A person is playing a guitar in the house.
YT: A boy is walking.
YT_C: A man is doing a women.
YT_CF: **A man is talking on a wall.**
GT: A man is doing algebraic equations on a white board.

FGM: A person is riding the horse
YT: A group of running.
YT_C: **A group of elephants.**
YT_CF: A group of elephants are walking on a horse.
GT: An elephant leads it's young.

FGM: A person playing the goal of the road.
YT: A player player in a goal.
YT_C: **A man playing a soccer ball.**
YT_CF: **A soccer player is running.**
GT: Two teams are playing soccer.

FGM: A person is running a race on the road.
YT: A group of running.
YT_C: **A group of people are running.**
YT_CF: A man is running.
GT: Eight men are running a race on a track.

Over fitting misses details and hurts.

FGM: A person is riding a motorbike in the kitchen.
YT: A man is jumping on the water.
YT_C: **A man is riding a bike.**
YT_CF: **A man is riding a motorcycle.**
GT: A boy is riding a motorcycle on the seashore .

Movie Corpus - DVS

CC: Queen: "Which estate?"

DVS: Looking troubled, the Queen descends the stairs. | The Queen rushes into the courtyard. She then puts a head scarf on . . . | . . . and gets into the driver's side of a nearby Land Rover. | The Land Rover pulls away. | Three bodyguards quickly jump into a nearby car and follow her.

Processed:
Looking troubled, someone descends the stairs.

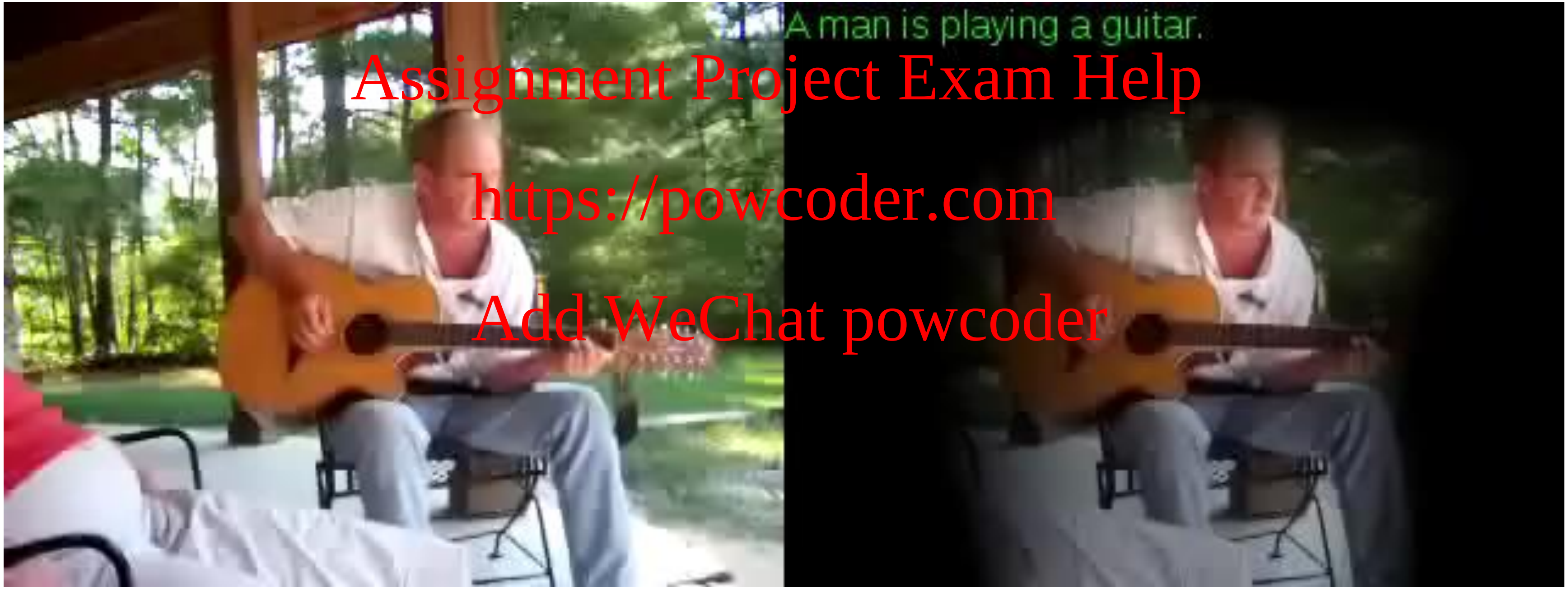
Someone rushes into the courtyard. She then puts a head scarf on ...

Examples (M-VAD Movie Corpus)

MPII-MD: https://youtu.be/XTq0huTXj1M
M-VAD: https://youtu.be/pER0mjzSYaM

Implicit Attention in LSTM

Implicit Attention in LSTM



Other Vision & Language Applications

Kate Saenko
Machine Learning

Visual Question Answering

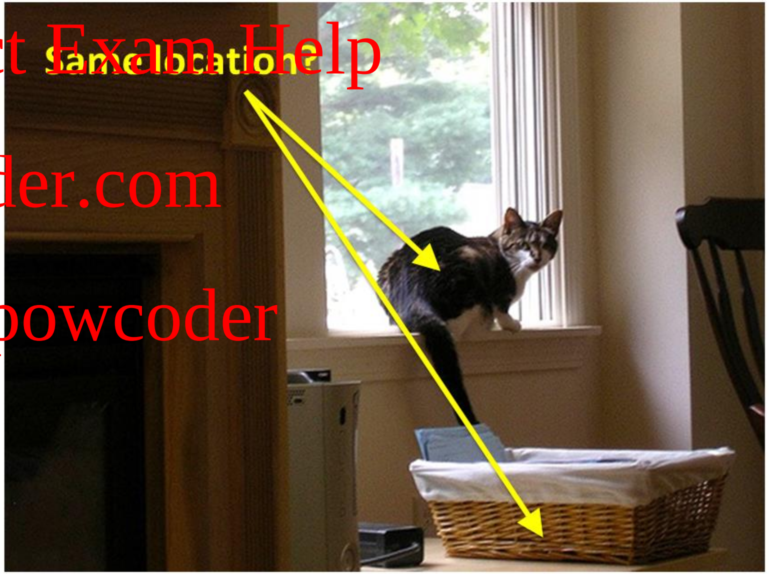
Is there a cat in the basket?

What color is the flower?

Where is the giraffe?

What animal is in front of the cow?

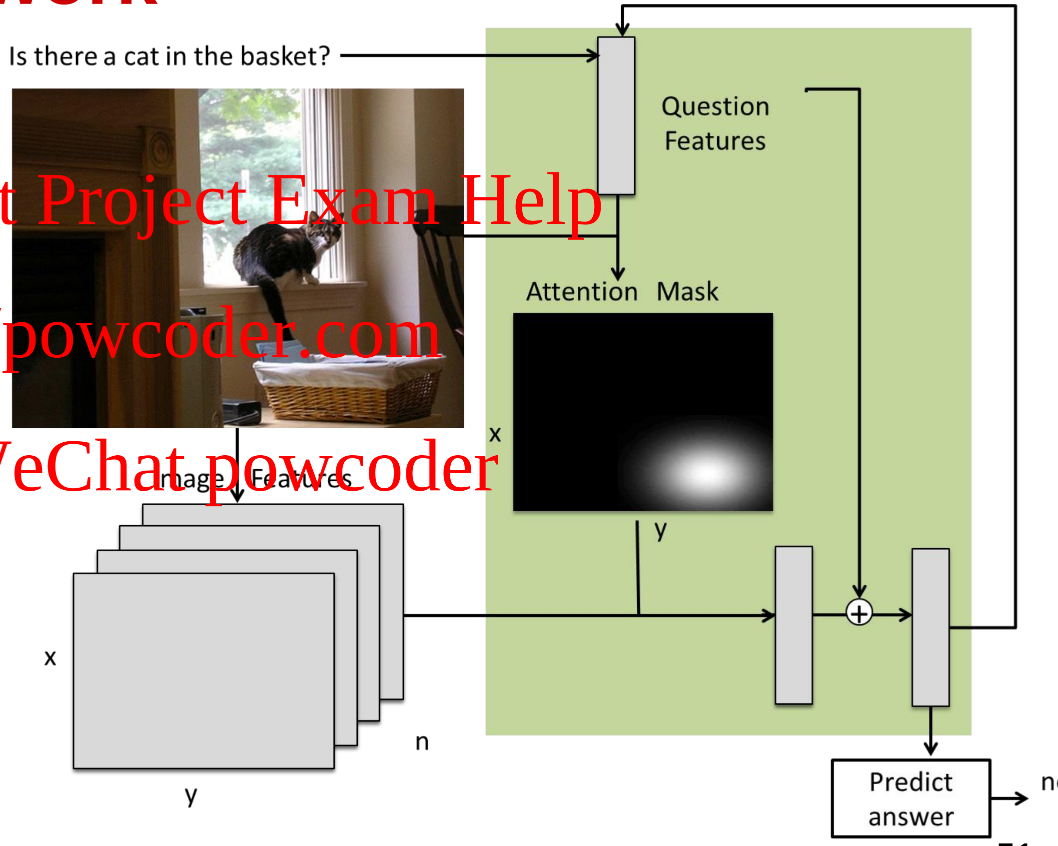
Question: Is there a cat in the basket?



some questions require reasoning

Visual Question Answering: Spatial Memory Network

- Based on Memory Networks [Weston2014], [Sukhbaatar'15]
- Store visual features from image regions in memory



S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks, 2015

J. Weston, S. Chopra, and A. Bordes. Memory networks, 2014.

Huijuan Xu, Kate Saenko, Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering, 2015 https://arxiv.org/abs/1511.05234

VQA Results

What season does this appear to be?

GT: fall Our Model: fall

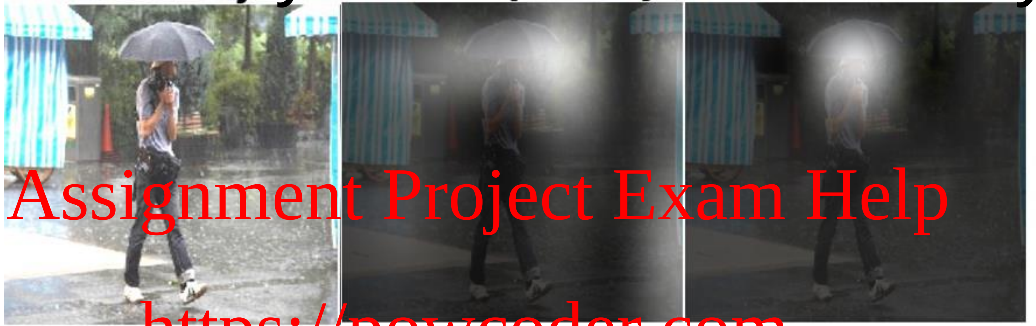
What color is the stitching on the ball?

GT: red Our Model: red

VQA Results

What is the weather?

GT: rainy Our Model: rainy



What color is the fence?

GT: green Our Model: green

Referring Expression Grounding

[Hu et al CVPR16]
[Hu et al CVPR17]
[Hu et al ECCV18]

Text-based object query

query: “lady in black shirt”

prediction

query: “window upper right”

Grounding expressions in video

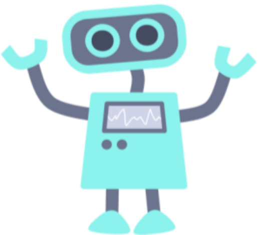
Given a query: Person holding the door to the refrigerator open

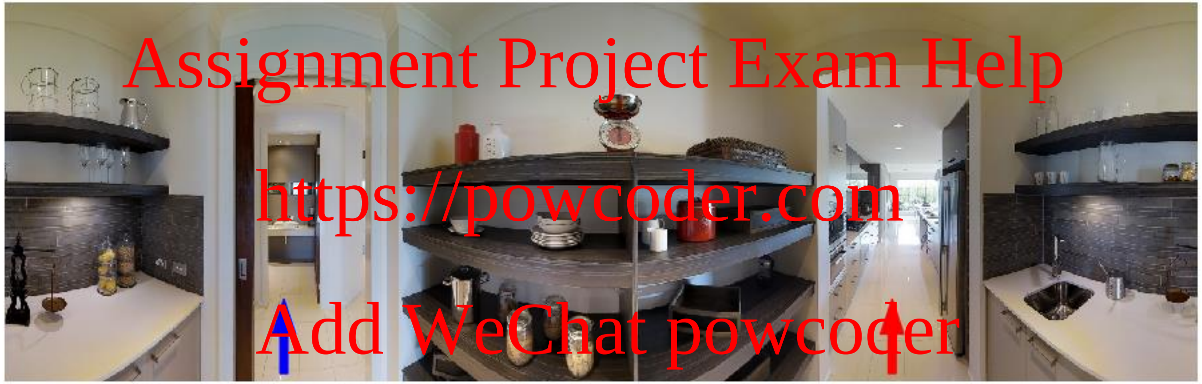
Find it in video

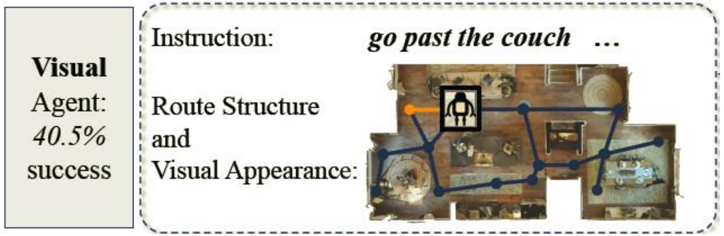
Multilevel Language and Vision Integration for Text-to-Clip Retrieval, Huijuan Xu, Kun He, Bryan A. Plummer, Leonid Sigal, Stan Sclaroff, Kate Saenko, AAAI19

Language based Navigation

Instruction: *Walk into the kitchen and go to the left once you pass the counters. Go straight into the small room with the sink. Stop next to the door.*







Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation, Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, Kate Saenko, ACL 2019

Summary

- variety of language & vision tasks
- active research area

References

[1] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, August 2014.

[2] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. *In IEEE International Conference on Computer Vision (ICCV)*.

[3] *Translating Videos to Natural Language Using Deep Recurrent Neural Networks*. Subhashini Venugopalan, Huijun Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, Kate Saenko. NAACL 2015

[4] *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. Jeff Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell. CVPR 2015.

[5] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko; ICCV 2015

[6] Huijuan Xu, Kate Saenko, Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering, 2015 https://arxiv.org/abs/1511.05234

Next Class

Applications II: Machine Learning Ethics:

Ethics in ML; population bias in machine learning, fairness, transparency, accountability; de-biasing image captioning models