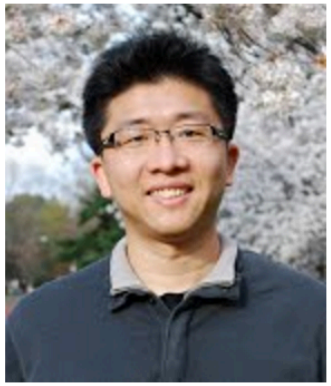
Training a Deep Agent to See and Interact

台灣清華大學
孫民教授

VSLab



Assistant Professor in Electrical Engineering
at National Tsing Hua University (Sept. 2014)
[CV],[Google Scholar]

Postdoctoral Researcher with Steve Seitz and Ali Farhadi
in CSE at University of Washington (Jan. 2013 - Aug. 2014)

2015

專題演講

國際會議廳

海量視覺資料

場次主持: 李維斌 / 臺北市政府資訊局長
孫民 / 國立清華大學電機工程學系助理教授

http://www.slideshare.net/tw_dsconf/ss-51943658

2016

專題演講

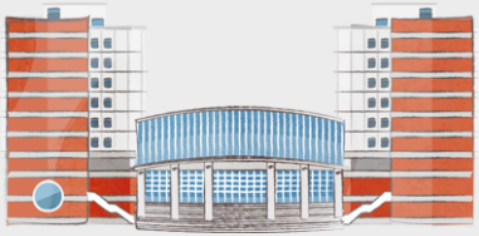
第一會議室

從電腦視覺看人工智慧：下一件大事

主持: 邱泊寰 / CAMEO CTO
孫民 / 清大電機系助理教授

https://www.slideshare.net/tw_dsconf/ss-64091725

2017



人文社會科學館
國際會議廳 R0

Training a Deep Agent to See and Interact

孫民 / 清華大學電機系助理教授

Work@NTHU

- CHI 2016, CHI 2017
- ACCV 2016 (oral)
- 2x ECCV 2016
- 2x CVPR 2017 (oral, spotlight)
- 4x ICCV 2017 (2x spotlight)
- AAAI 2017, IJCAI 2017, 2x AAAI 2018
- ICLR 2017 workshop

http://aliensunmin.github.io/publications.html

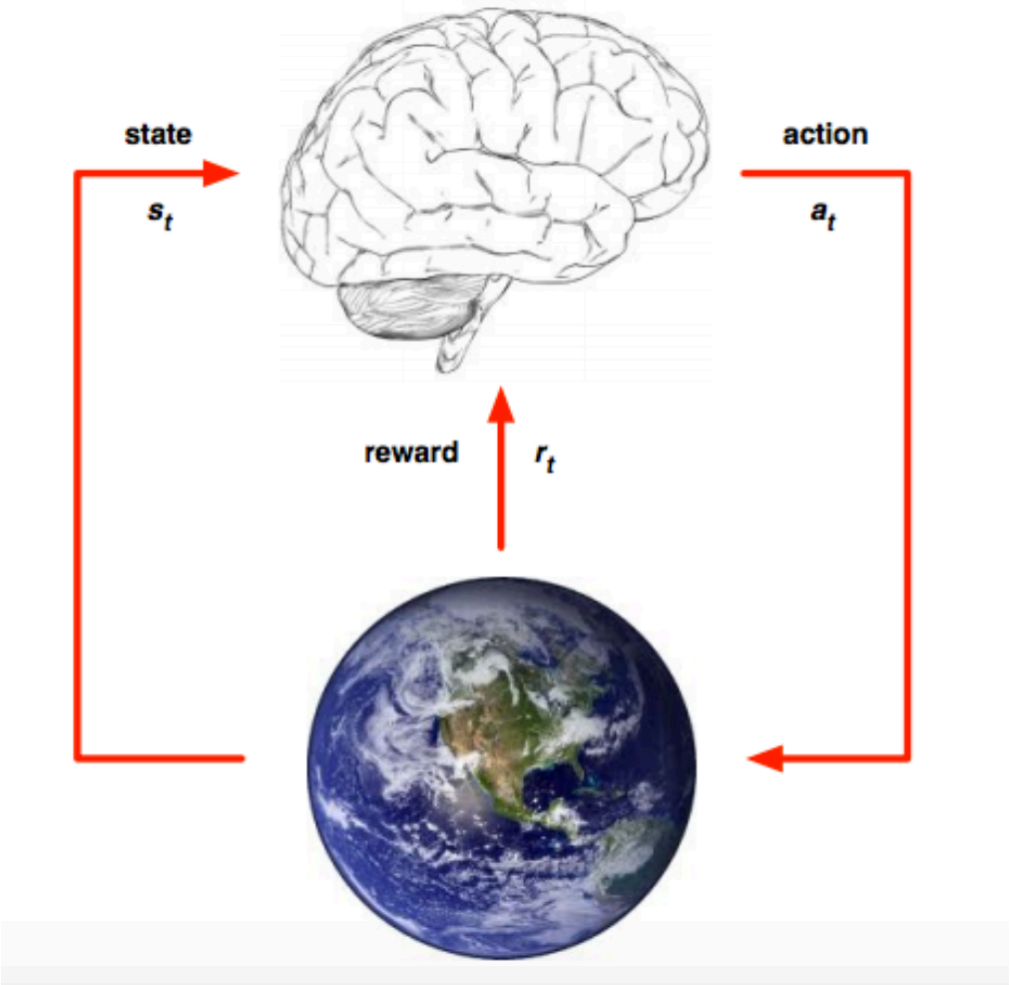
Training a Deep Agent to See and Interact

台灣清華大學
孫民教授

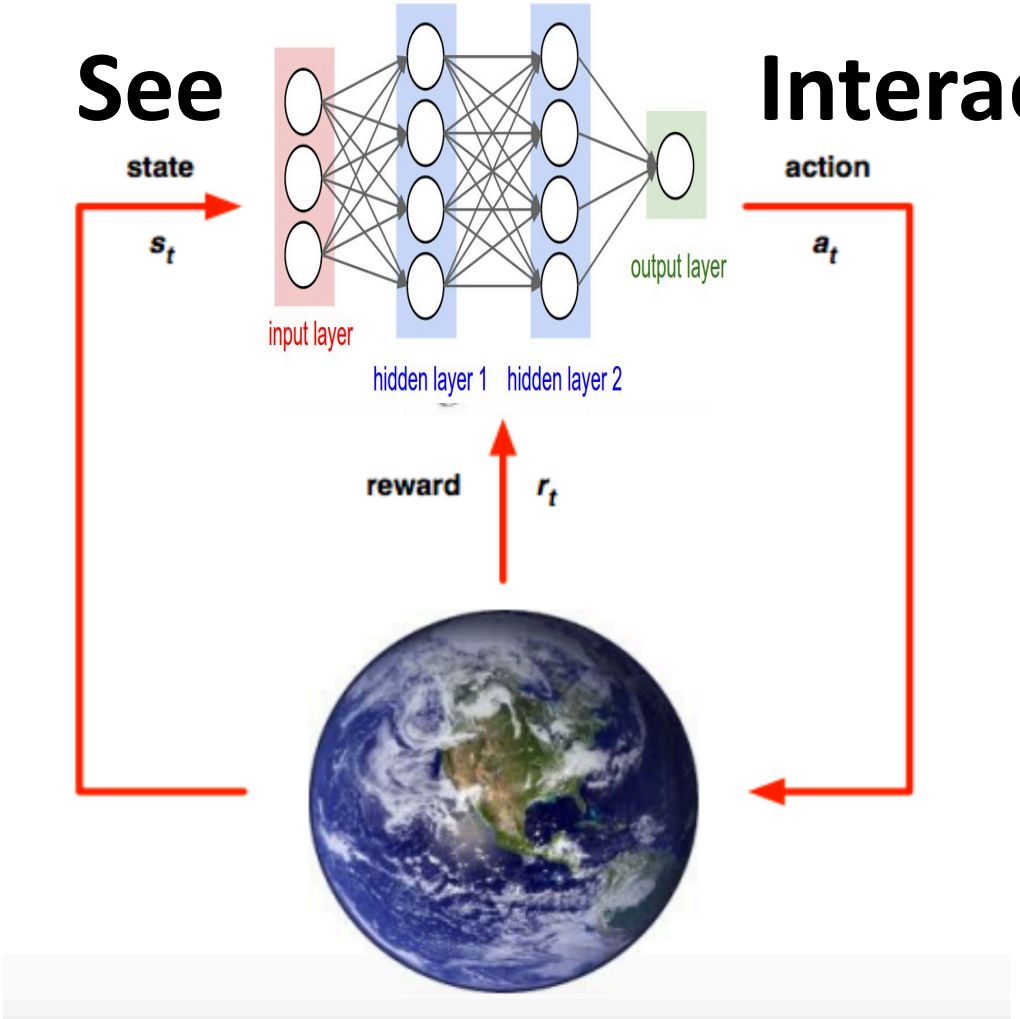
VSLab

What is an Agent?

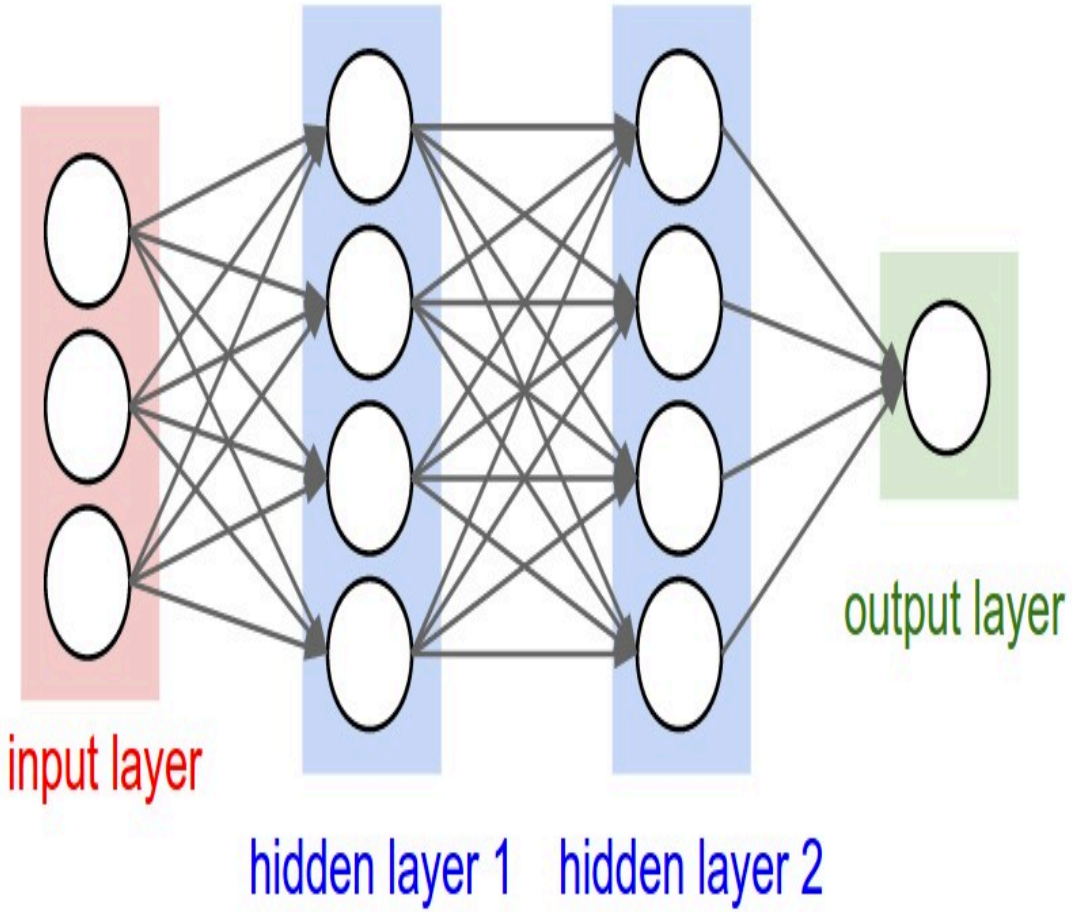
“We” are one of the best Agent



Deep Agent = DNN as the Agent



Why Deep Learning?



Data: IMAGENET

- 開始於 2007 @ Princeton
- 初登場於 2009 @ CVPR
- 照片停止搜集於 2010
  - ➢總共類別：21841
  - ➢總共圖片：1千4百萬

Jia Deng

Fei-Fei Li

POPULAR

QUARTZ

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

1K Image Classification

Label = f(Image)

Deep Learning

深度學習

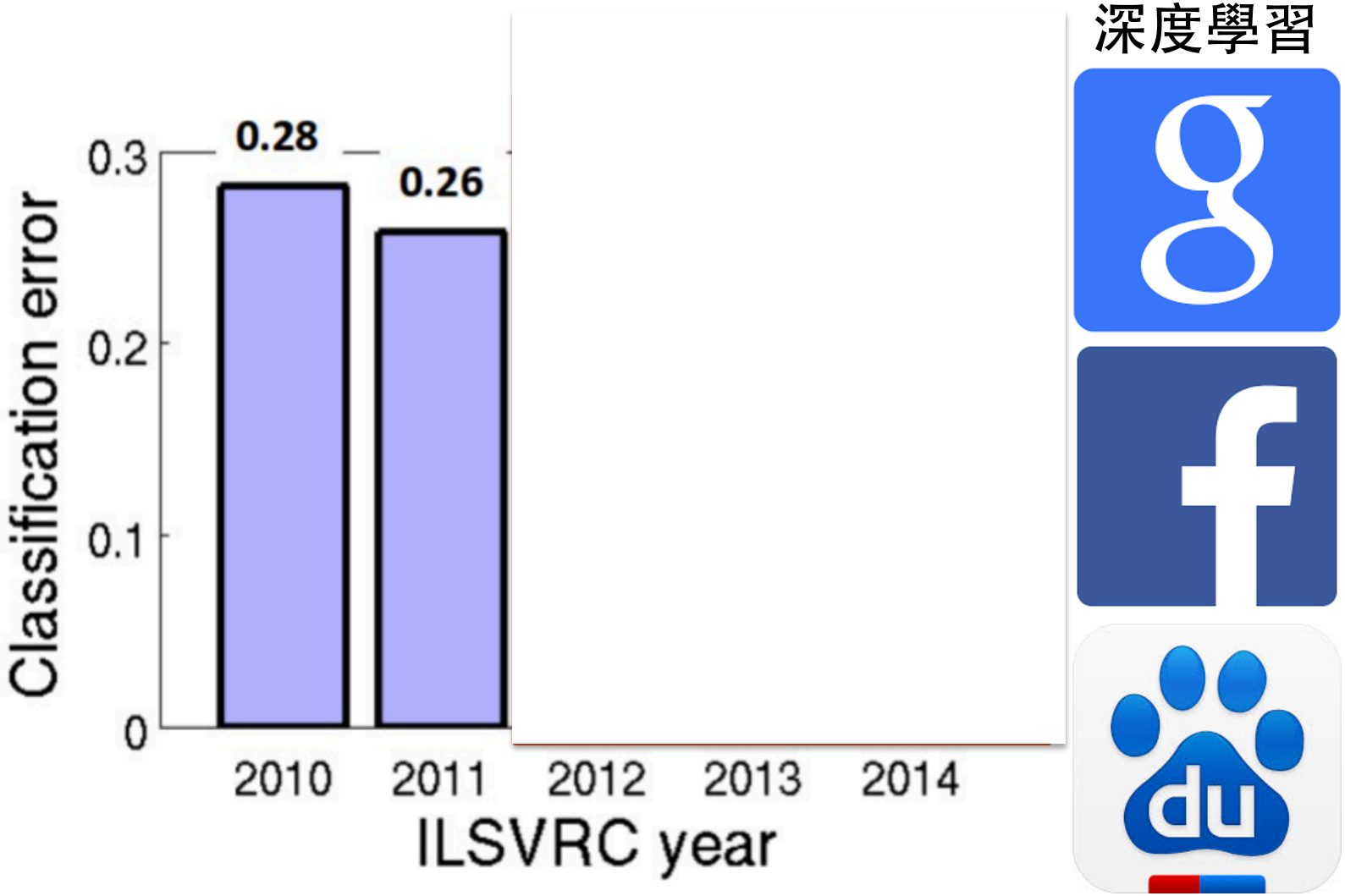
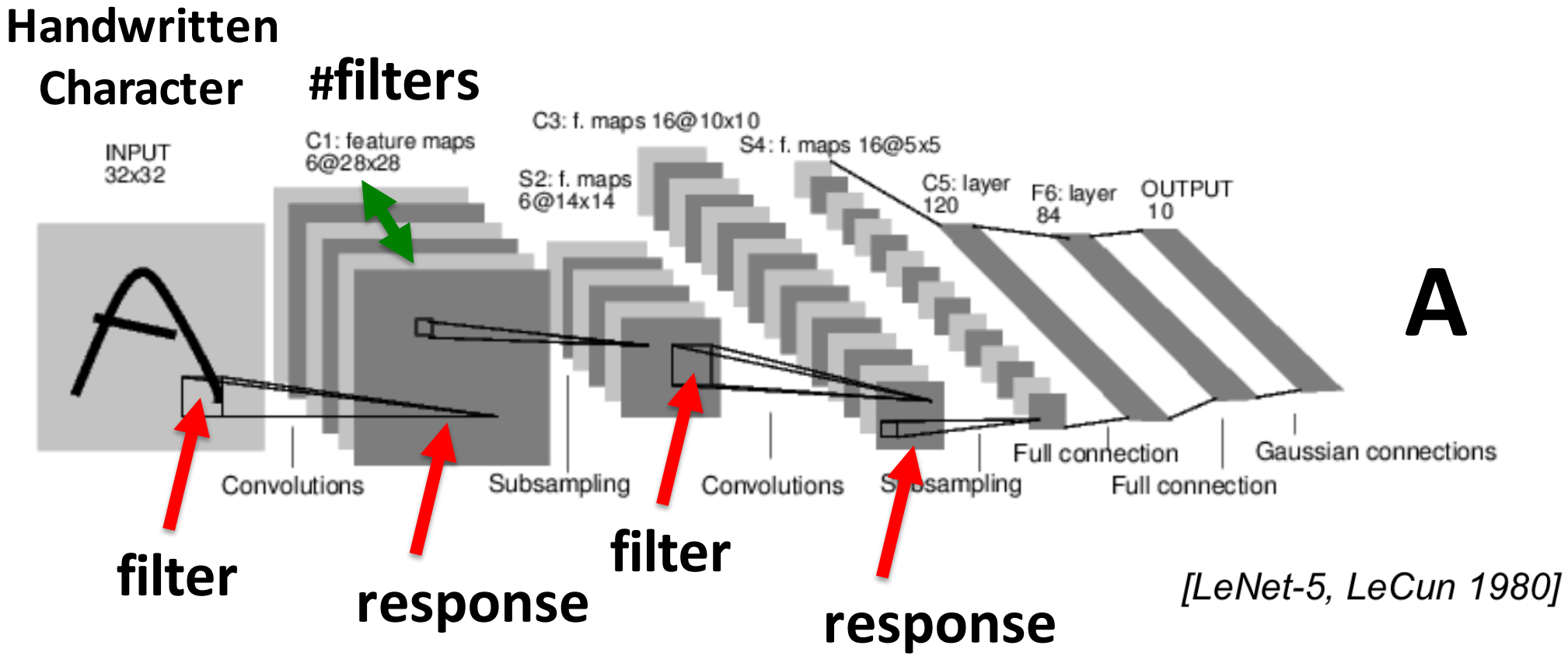


Figure from Olga Russakovsky ECCV'14 workshop

Convolutional Neural Network (CNN)

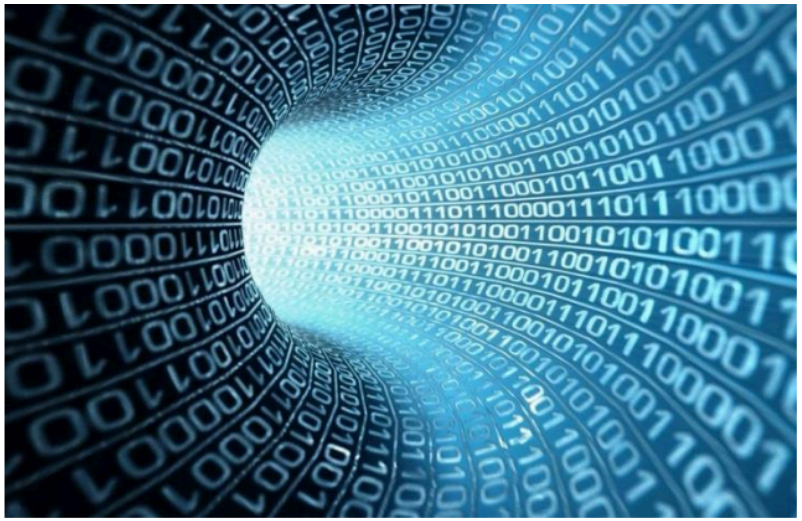


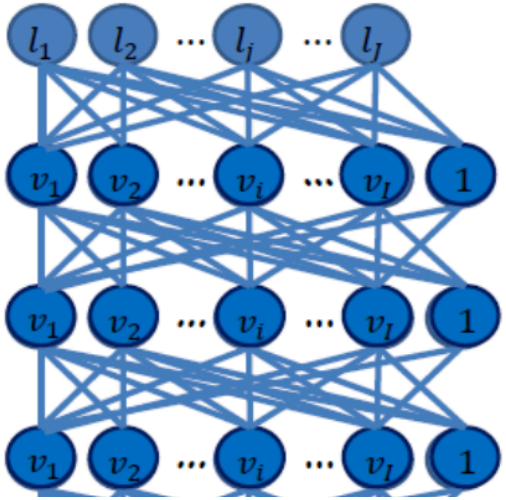
Character = f(Image)

Automatic Speech Recognition (ASR)

- DNN has become the-state-of-the-art for speech recognition since 2010

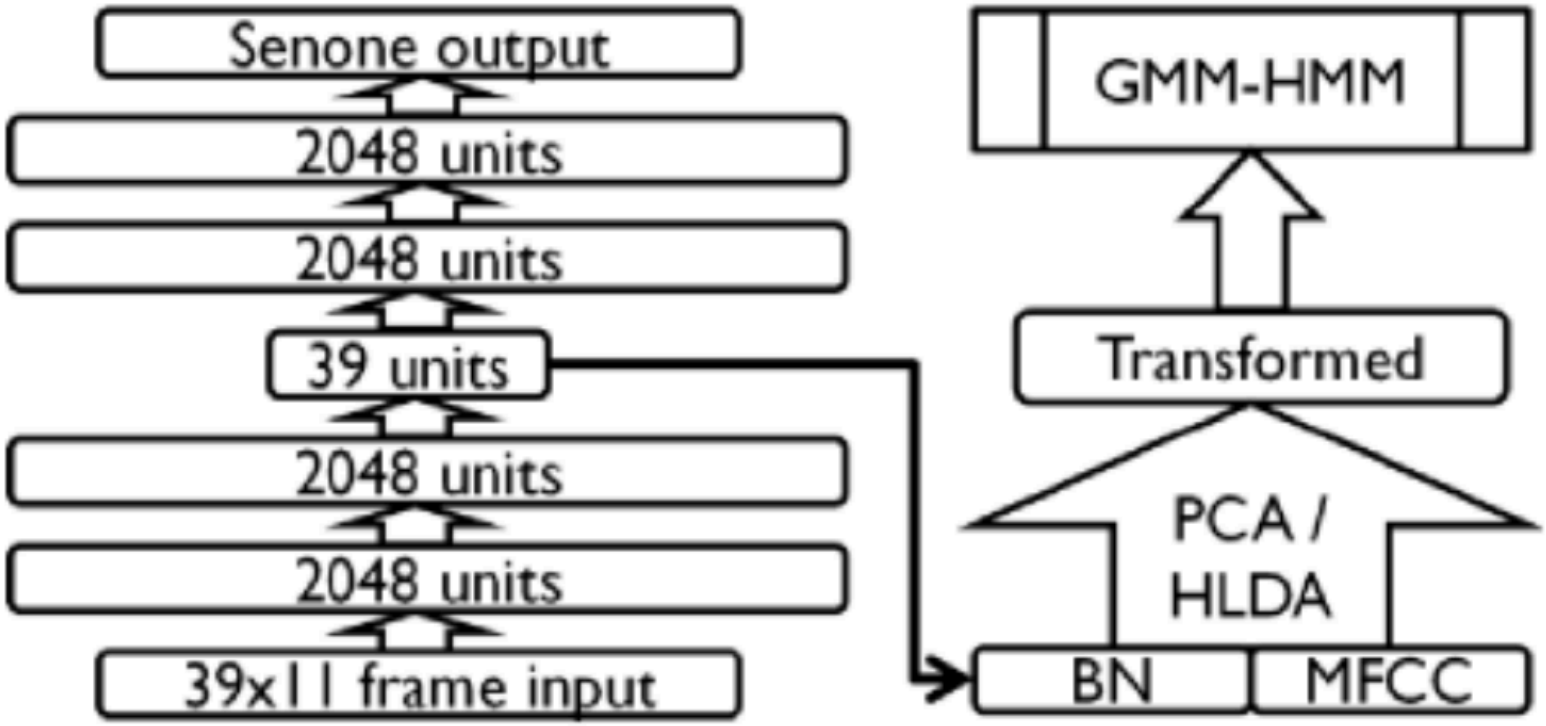
✓ Large volume data available
✓ Deep learning is changing everything. AM, LM…
✓ ASR is going to practice. Siri, google, baidu, Tencent…





DBN-HMM

Microsoft: 33% Recognition Error Reduction over discriminatively trained GMM-HMMs



F. Seide, Gang Li, Dong Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, Interspeech 2011, Florence, Italy

Natural Language Processing (NLP)

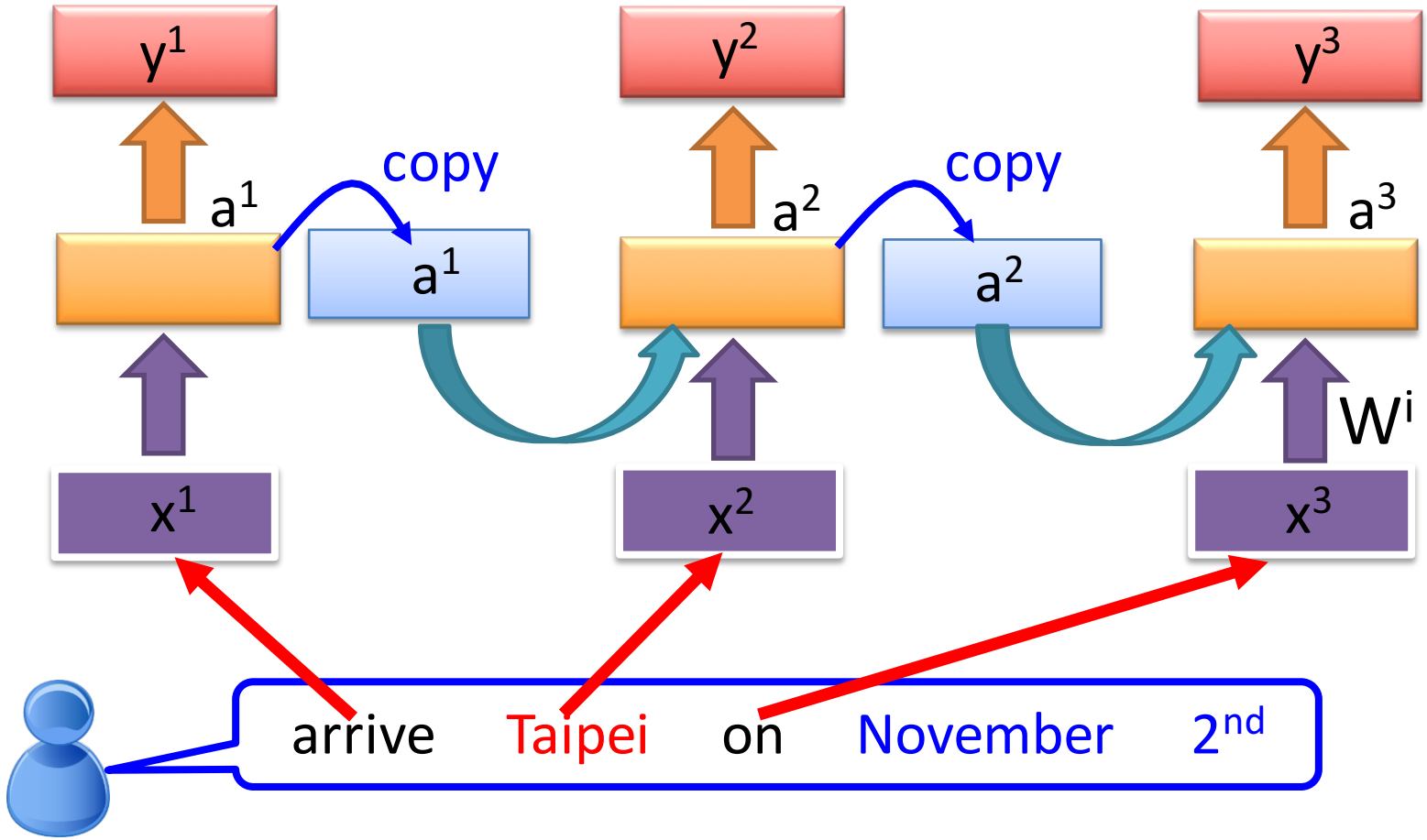
- “Deep Learning waves have lapped at the shores of computational linguistics for several years now, but 2015 seems like the year when the full force of the tsunami hit the major Natural Language Processing (NLP) conferences.” -Dr. Christopher D. Manning, Dec 2015

Recurrent Neural Network

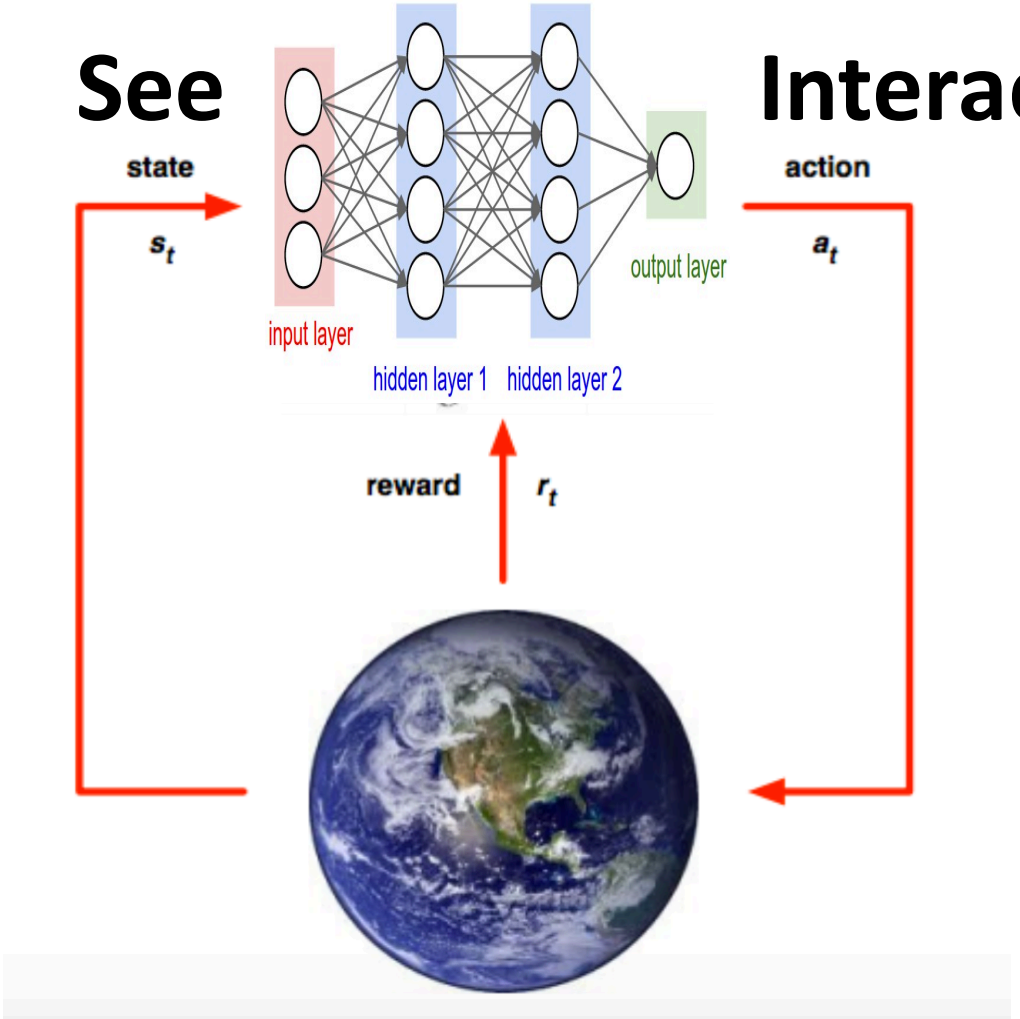
Probability of "arrive" in each slot

Probability of "Taipei" in each slot

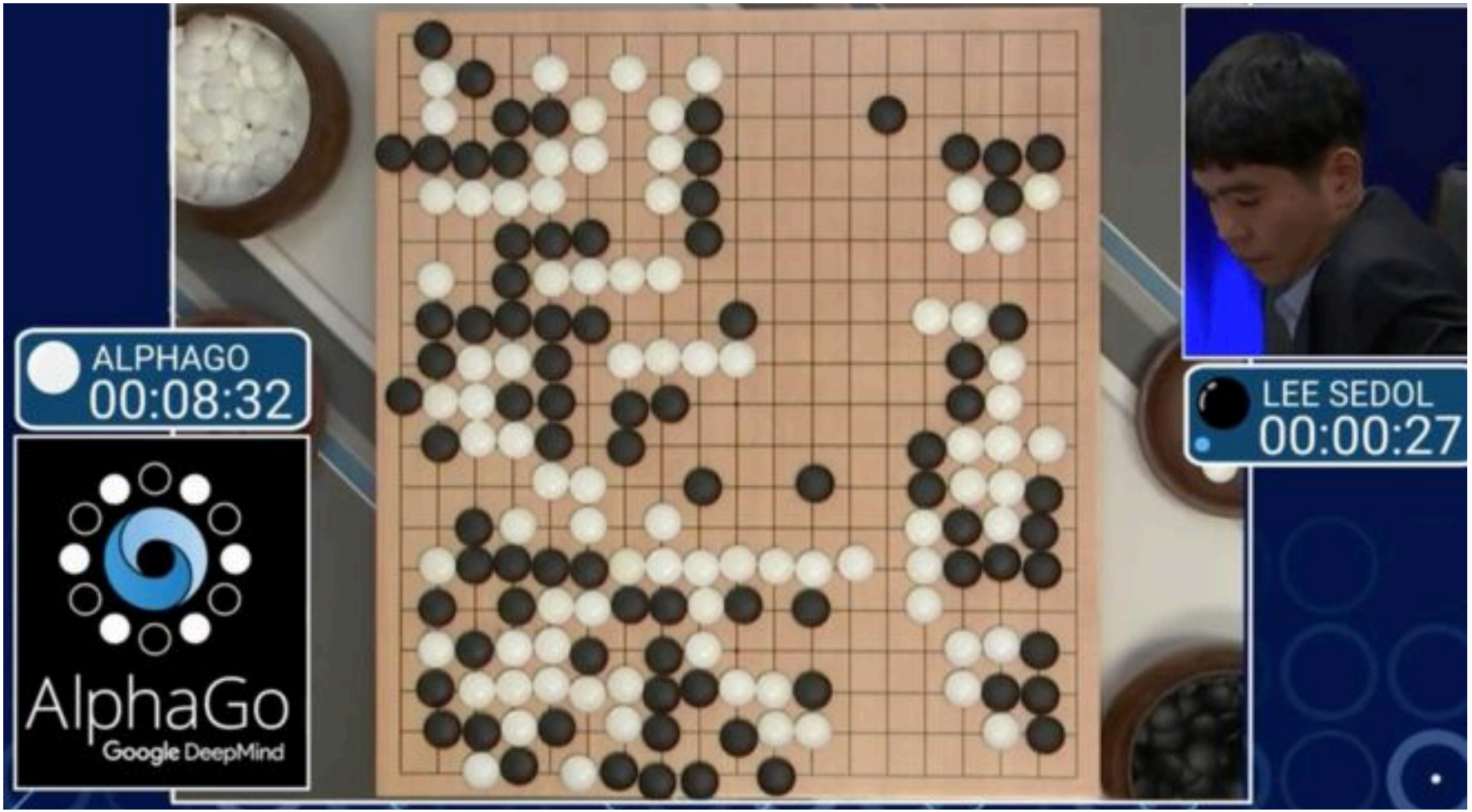
Probability of "on" in each slot



Deep Agent = DNN as the Agent



AlphaGo = See and Place Stone

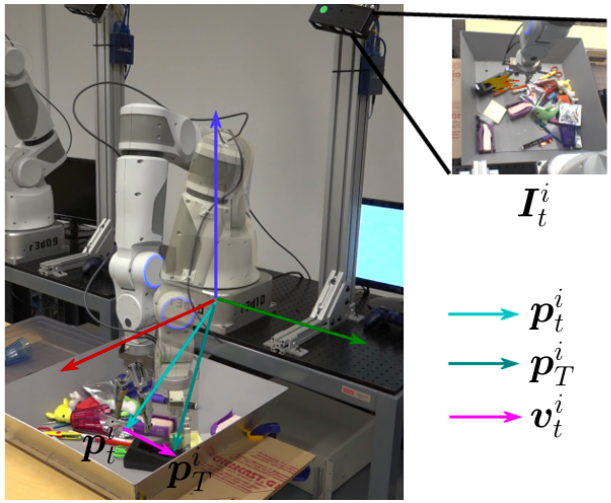


2016 by Google DeepMind

Robotic Applications

- Driving

- Grasping



End-effector
- Rotation
- Translation
- Grasp

Outline

- Interact through **Language**
  - Video Title Generation
  - Transferring Sentence Style
- Interact through **Actions**
  - Deep 360-degree Pilot
  - Target Driven Navigation
- Interact through **Attacks**
  - Adversarial Attacks

Outline

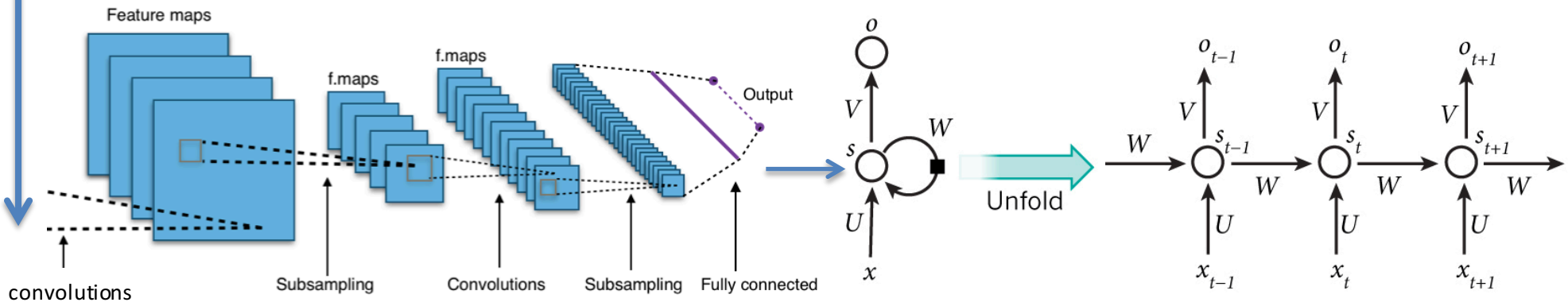
- Interact through **Language**
  - Video Title Generation
  - Transferring Sentence Style
- Interact through **Actions**
  - Deep 360-degree Pilot
  - Target Driven Navigation
- Interact through **Attacks**
  - Adversarial Attacks

See and Interact through Language

Vision

Language

f() = The man at bat is ready to swing at the pitch



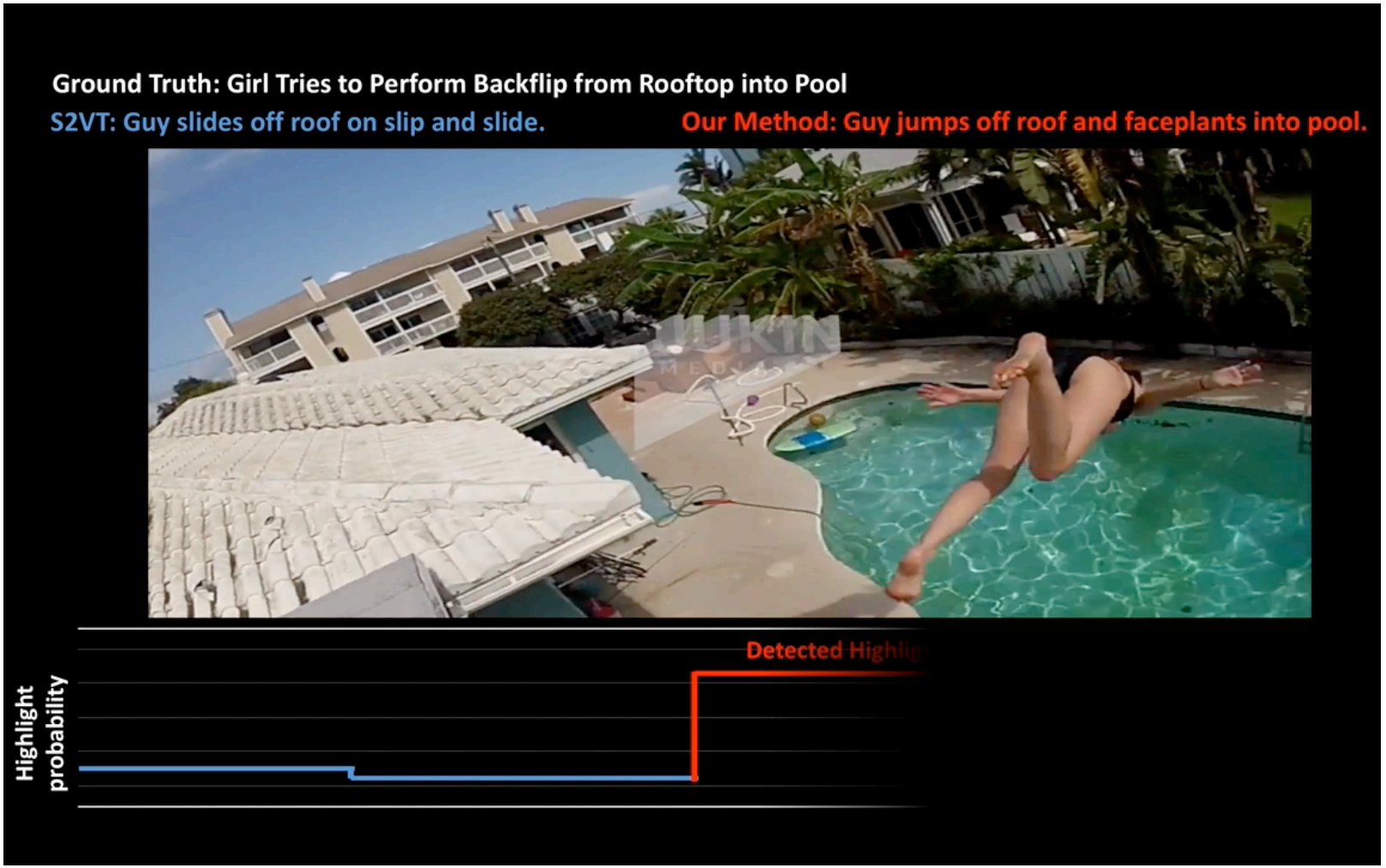
Convolution Neuron Network (CNN)

credit: wiki

Recurrent Neuron Network (RNN)

credit: Nature

Video Title Generation

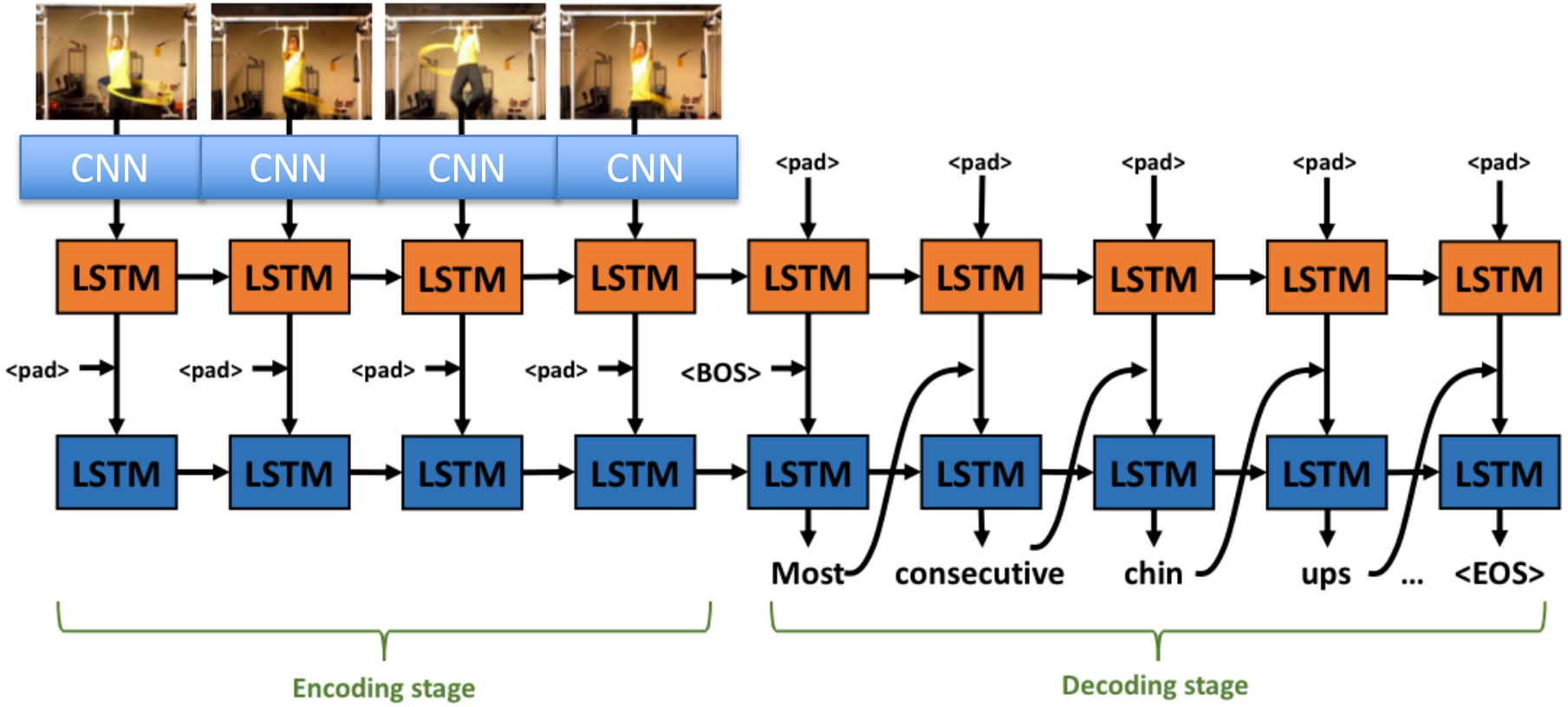


Zhen et al. ECCV 2016 with Stanford AI Lab

Big Video Data with Titles

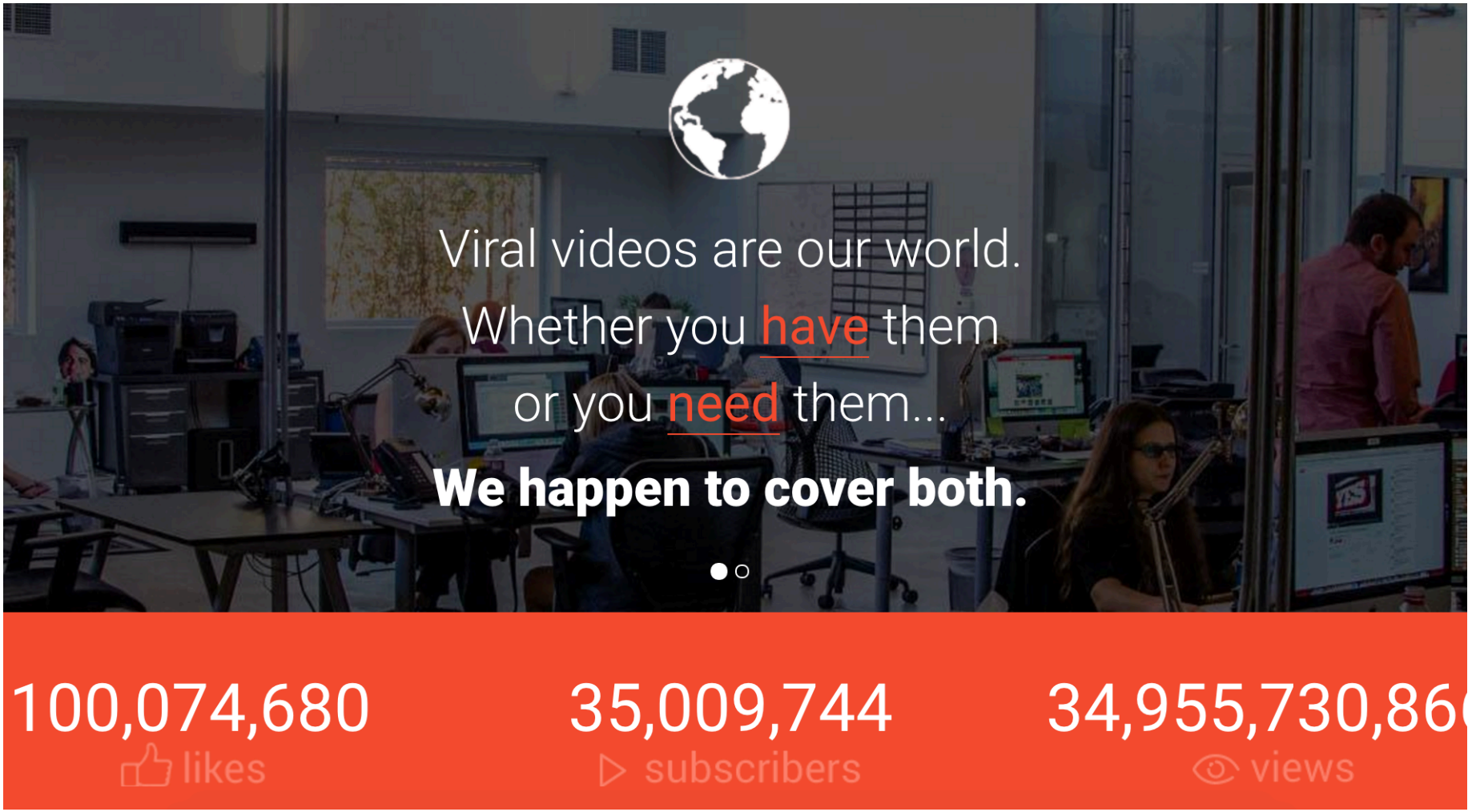
- Pairs of

Raw Video



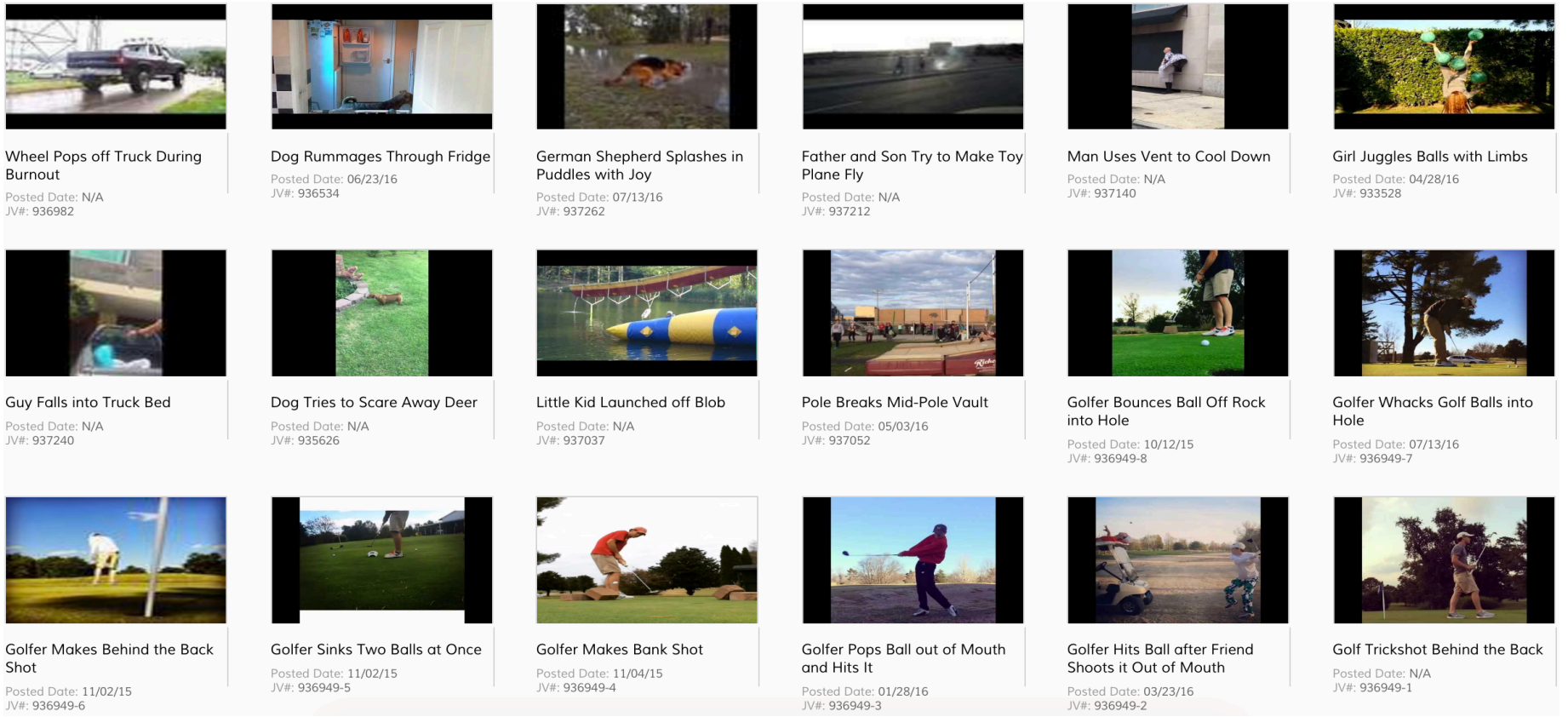
Title

Viral Videos



Google for “viral video company”

Large Video Repository



Currently 28740 videos and keep growing

http://140.128.137.13:5433/UGVideo/index.html

Transferring Sentence Style

Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner

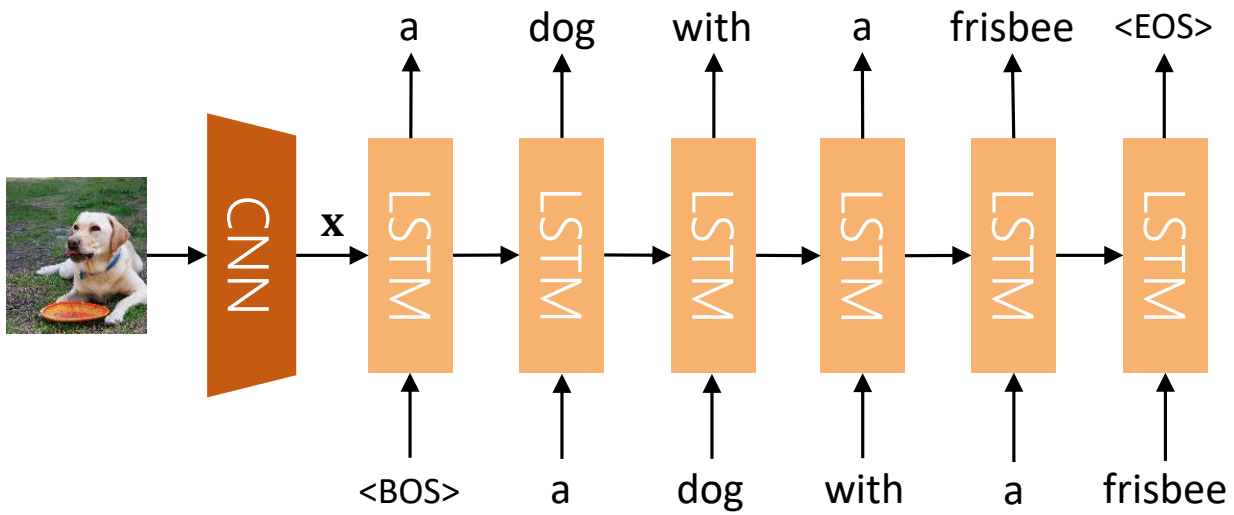
Tseng-Hung Chen[1] Yuan-Hong Liao[1] Ching-Yao Chuang[1]
Wan-Ting Hsu[1] Jianlong Fu[2] Min Sun[1]

[1)]National Tsing Hua University
[2)]Microsoft Research Asia

Accepted by ICCV 2017 with MSRA

Standard Image Captioning

VSLab



VSLab

Cross-domain Setting

We use **paired data** in source domain

A family of ducks swimming in the water.

A hummingbird close to a flower trying to eat.

unpaired data in target domain

images

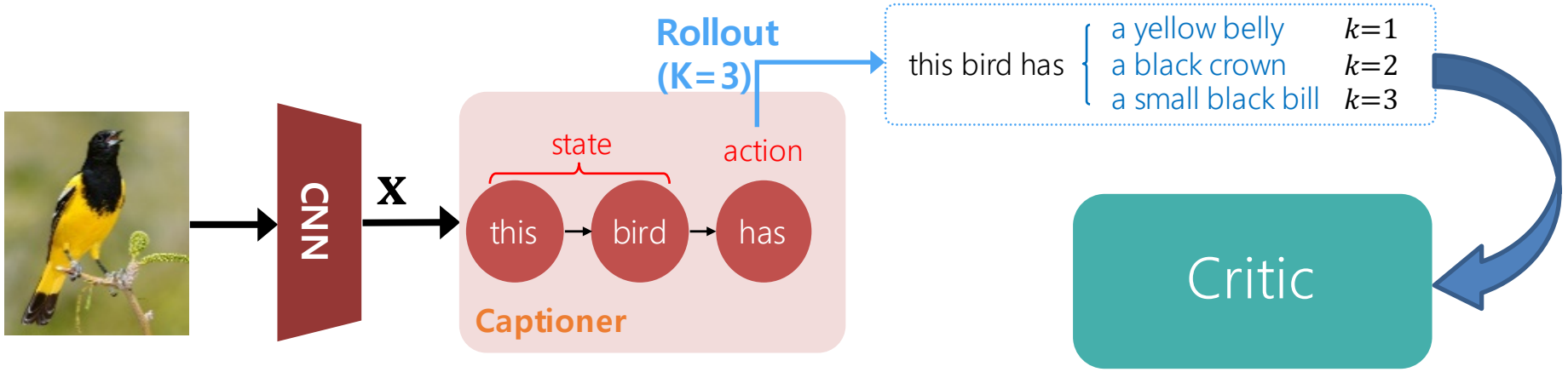
captions

This bird has wings that are brown and has red eyes.

A small bird with orange flank and a long thin black bill.

Captioner as an Agent

VSLab



Estimate $Q((\mathbf{x}, \boldsymbol{y_{t-1}}), y_t)$

$$E_{\mathbf{y}_{(t+1):T}}\left[R(\left[\mathbf{y}_{t-1}, y_t, \mathbf{y}_{(t+1):T}\right]|\mathbf{x}, \mathcal{Y}, \mathcal{P})\right] .$$

$$Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t) \simeq \frac{1}{K}\sum_{k=1}^{K} R(\left[\mathbf{y}_{t-1}, y_t, \mathbf{y}^k_{(t+1):T_k}\right]|\mathbf{x}, \mathcal{Y}, \mathcal{P}) ,$$

For cross-domain image captioning, a good caption needs to satisfy two criteria:

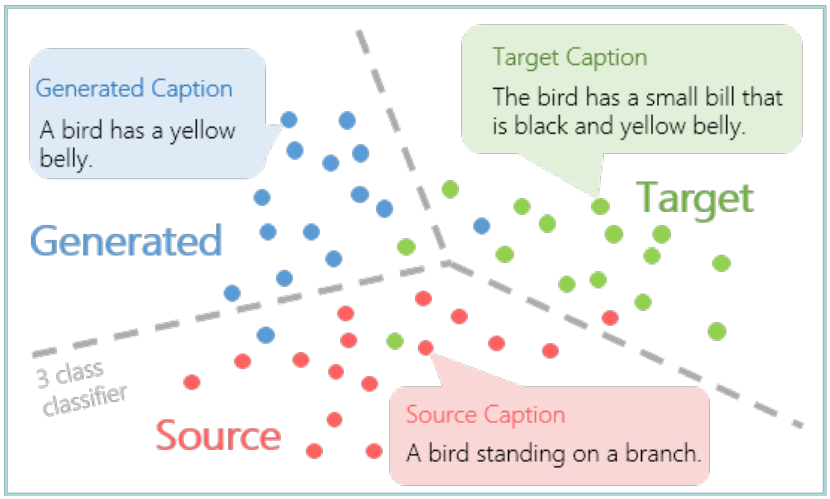

Target Image


Source Image

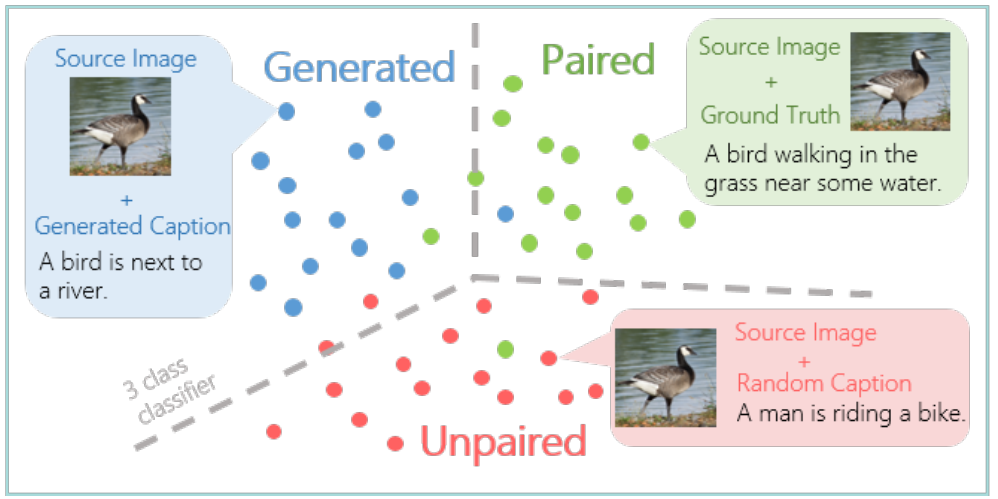
(1) the generated sentence resembles the sentence drawn from the target domain.

(2) the generated sentence is relevant to the input image.

Domain Critic

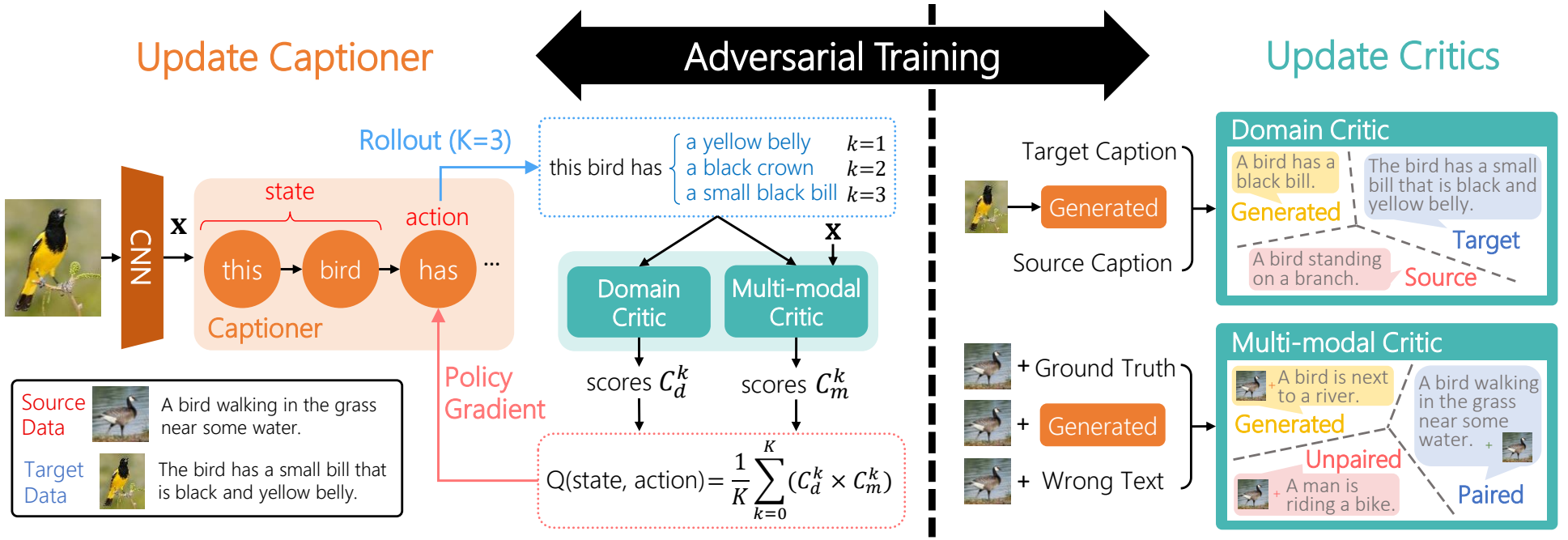


Multi-modal Critic



Training Critics

VSLab



VSLab

Before (trained on MSCOCO):
A yellow and yellow bird is sitting on a branch.

After (adapt on CUB-200):
This is a yellow bird with a black head and a small beak.

VSLab

Before (trained on MSCOCO):
A red flower in a yellow vase on a wooden table.

After (adapt on Oxford-102):
This flower has petals that are pink and has red dots.

Before (trained on MSCOCO):
A cat is standing in a room with a cat.

After (adapt on TGIF):
A cat is playing with a toy in a room.

Before (trained on MSCOCO):
A woman in a black shirt and a white shirt and a blue tie.

After (adapt on TGIF):
A woman is dancing with a crowd of people.

Quantitative Results

VSLab

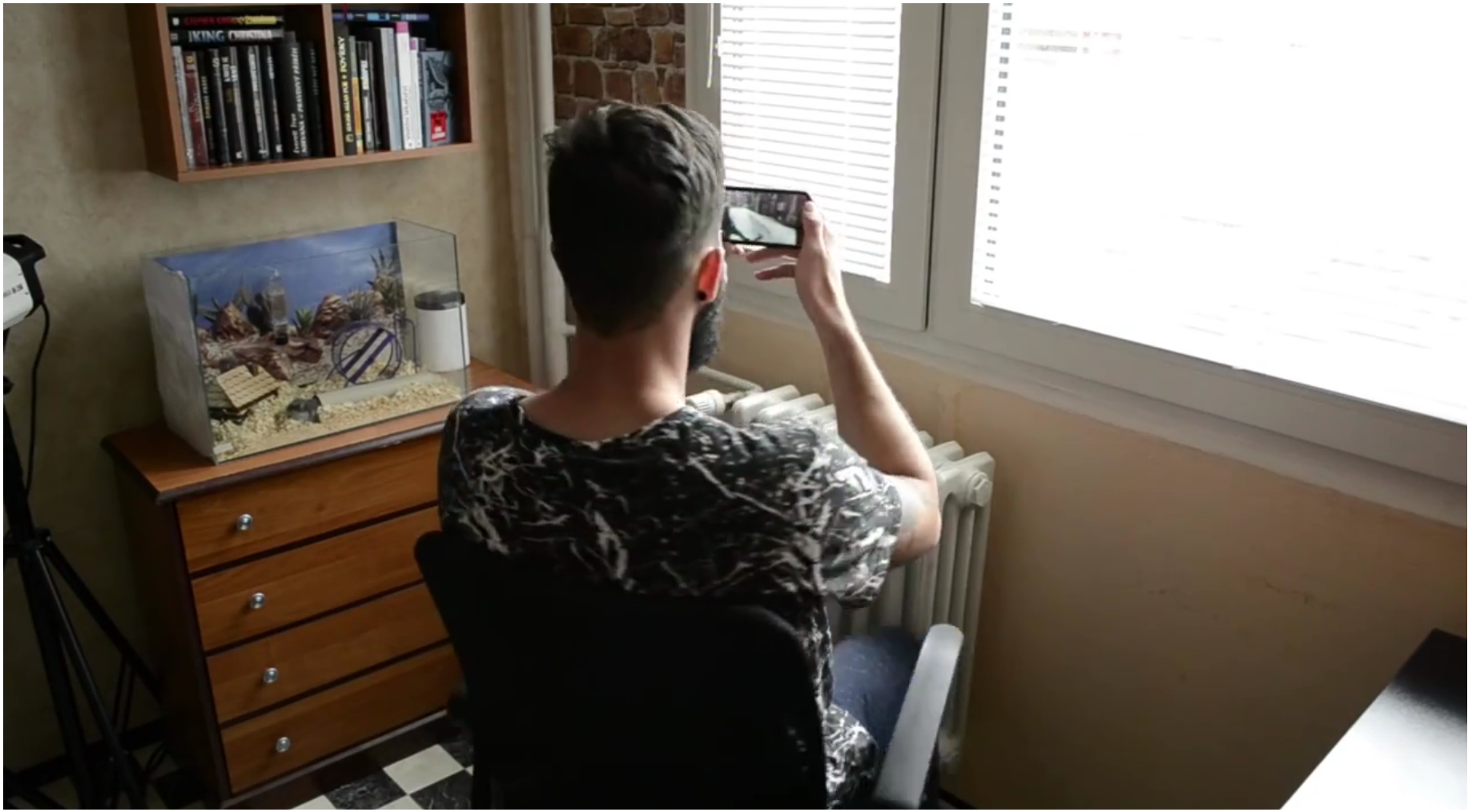
Method	Target (test)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	ROUGE	CIDEr	SPICE
Source	CUB-200	50.8	28.3	13.9	6.1	12.9	33	3	4.6
DCC	CUB-200	68.6	47.3	31.4	21.4	23.8	46.4	11.9	11.1
SA	CUB-200	76.5	61.2	44.6	30.7	22.6	51.1	21.7	12.8
Ours	CUB-200	**91.4**	**73.1**	**51.9**	**32.8**	**27.6**	**58.6**	**24.8**	**13.2**
Fine-tuning	CUB-200	91.3	80.2	69.2	59	36.1	69.7	61.1	17.9
Source	Oxford-102	48.3	21.6	6.2	1.3	10.5	25.8	3.1	4.4
DCC	Oxford-102	51	33.8	24.1	16.7	21.5	38.3	6	9.8
SA	Oxford-102	68.9	48.2	31.6	20.3	19.4	40.7	15.2	10.6
Ours	Oxford-102	**85.6**	**76.9**	**67.4**	**60.5**	**36.4**	**72.1**	**29.3**	**17.9**
Fine-tuning	Oxford-102	87.5	80.1	72.8	66.3	40	75.6	36.3	18.5
Source	TGIF	41.6	23.3	12.6	7	12.7	32.7	14.7	8.5
DCC	TGIF	34.6	17.5	9.3	4.1	11.8	29.5	7.1	7.3
SA	TGIF	45	25.6	14.3	8.2	13.6	34.9	18	9.3
Ours	TGIF	**47.5**	**29.2**	**17.9**	**10.3**	**14.5**	**37**	**22.2**	**10.6**
Fine-tuning	TGIF	51.1	32.2	20.2	11.8	16.2	39.2	29.8	12.1
Source	Flickr30k	57.3	36.2	21.9	13.3	15.1	38.8	25.3	8.6
DCC	Flickr30k	54.3	34.6	21.8	13.8	16.1	38.8	27.7	9.7
SA	Flickr30k	57.9	37.6	23.6	14.7	15.7	39.6	27.2	9.3
Ours	Flickr30k	**62.1**	**41.7**	**27.6**	**17.9**	**16.7**	**42.1**	**32.6**	**9.9**
Fine-tuning	Flickr30k	59.8	41	27.5	18.3	18	42.9	35.9	11.5

VSLab

Outline

- Interact through **Language**
  - Video Title Generation
  - Transferring Sentence Style
- Interact through **Actions**
  - Deep 360-degree Pilot
  - Target Driven Navigation
- Interact through **Attacks**
  - Adversarial Attacks

Watching 360 Videos



Auto Pilot vs. Manually Select

Auto Pilot

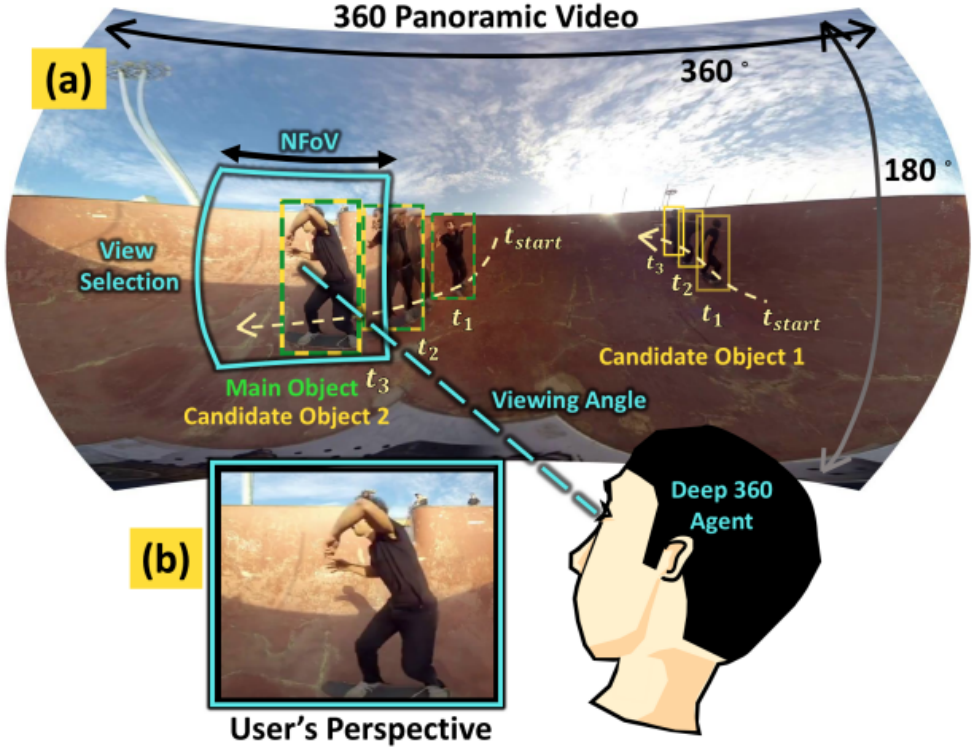
Manually Select

Lin et al., *Tell Me Where to Look: Investigating Ways for Assisting Focus in 360˚Video*, ACM SIGCHI 2017

Icon credit: www.flaticon.com

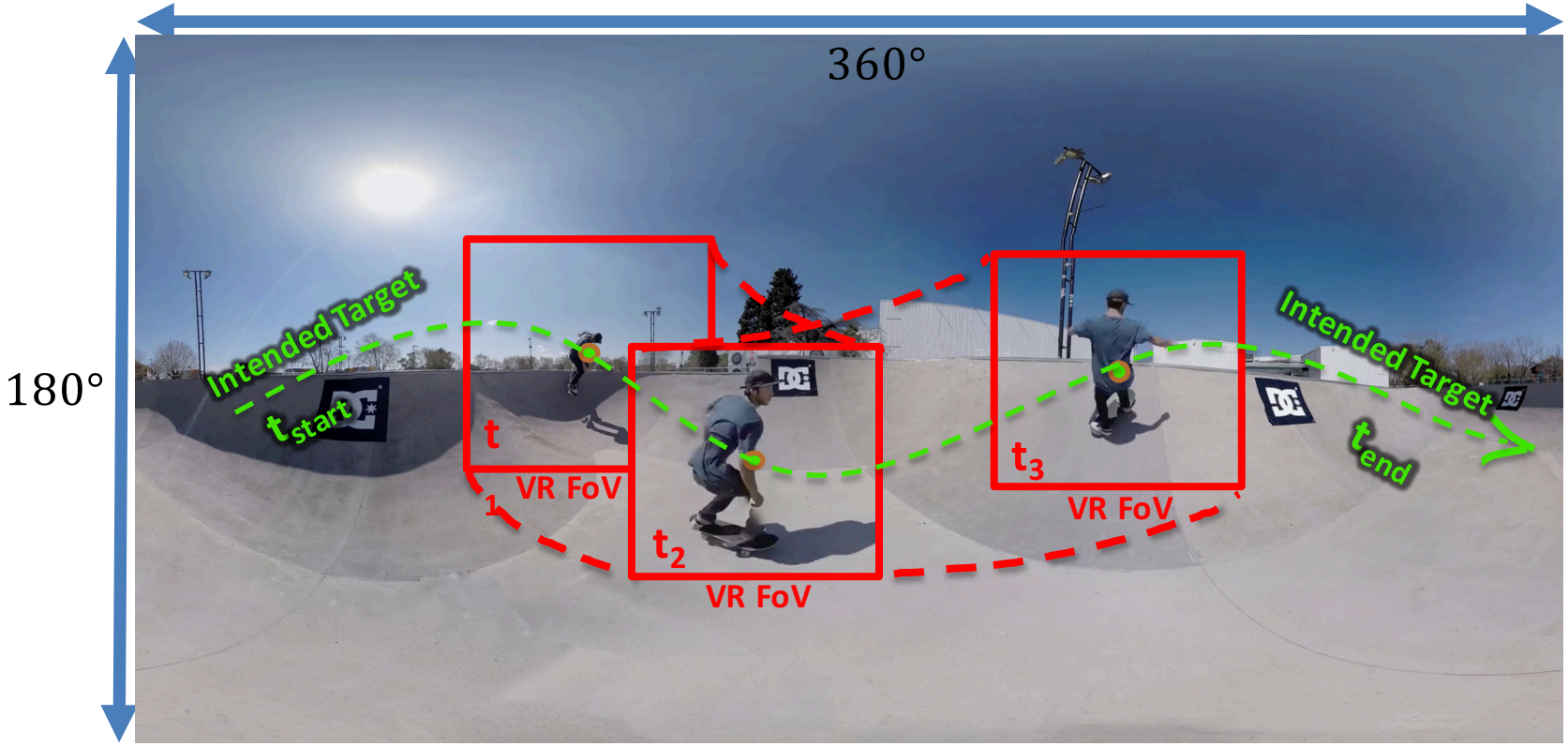
Deep 360 Pilot – A Smart 360° Video Assistant

The first human-like online agent for automatically navigating 360° videos



Hu et al. CVPR 2017 Oral with NVIDIA Research

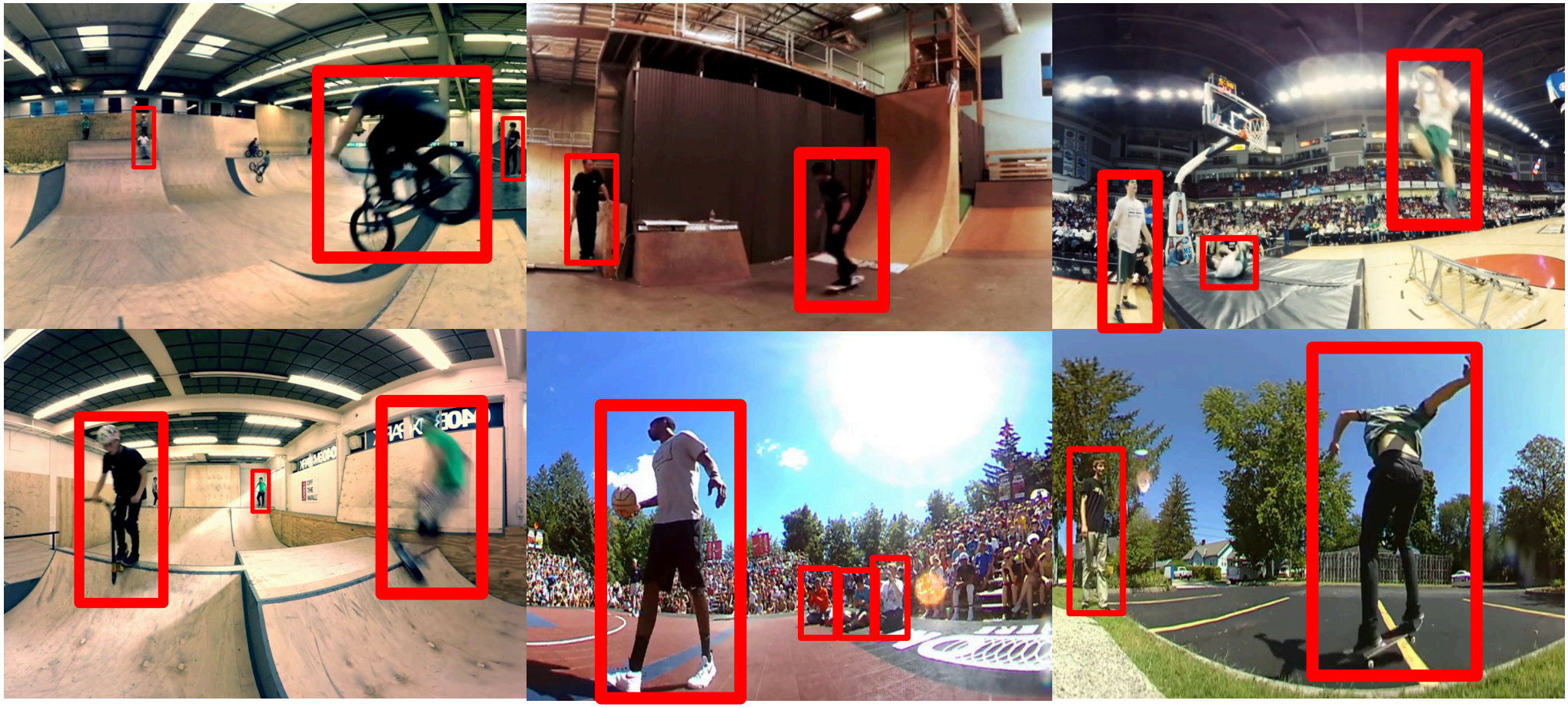
Main Task – Auto Pilot



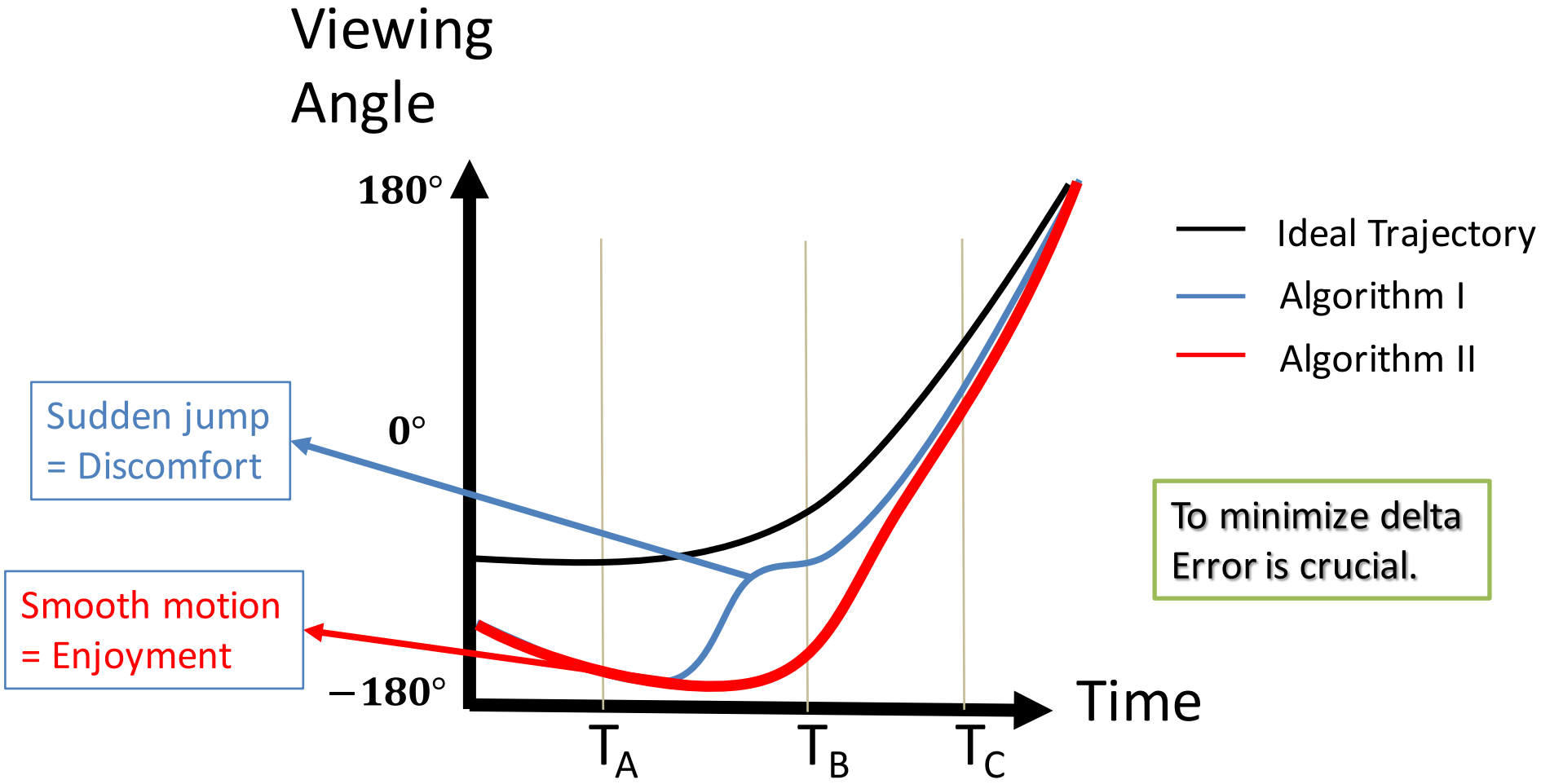
Viewing Angle l_t

Viewing Angle Trajectory $L_t = \{l_1, \ldots, l_t\}$

Assumption – Salient Objects

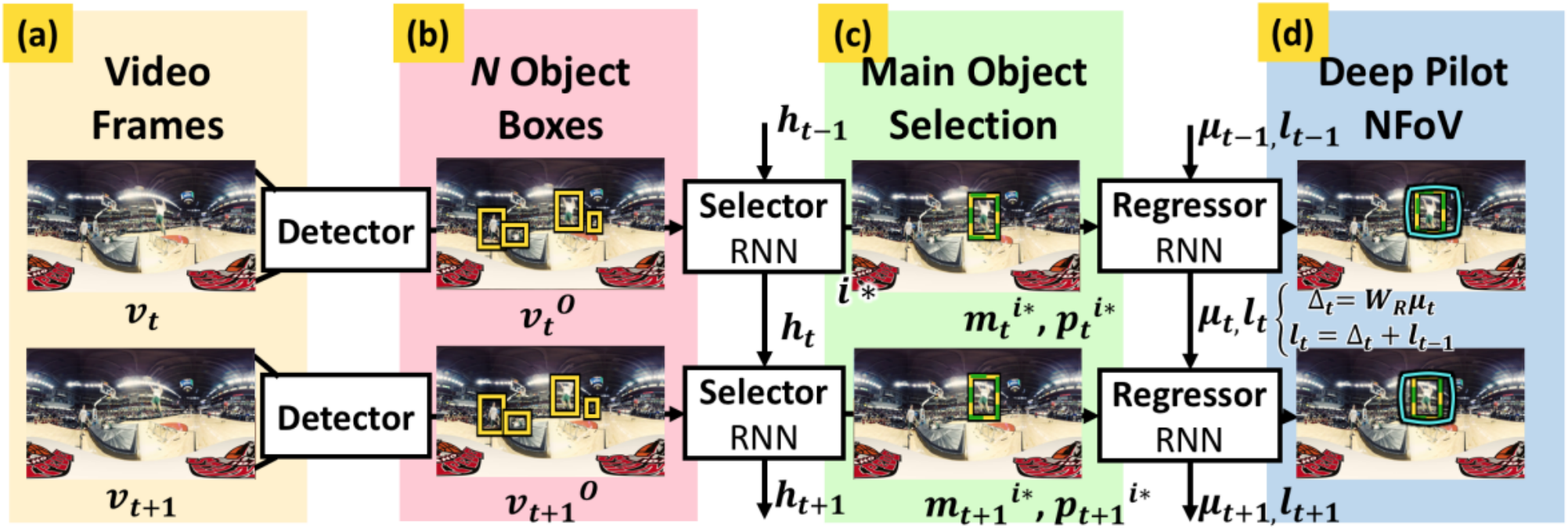


Assumption – Trajectory Smoothness



Chen et al., *Learning online smooth predictors for real-time camera planning using recurrent decision trees*. CVPR, 2016.

Deep 360 Agent



Hu et al. CVPR 2017 Oral with NVIDIA Research

Deep 360 Agent

Method	Skateboarding		Parkour		BMX		Dance		Basketball	
	MO	MVD	MO	MVD	MO	MVD	MO	MVD	MO	MVD
Ours **w/o Regressor.**	**0.71**	6.03	**0.74**	4.72	**0.71**	10.73	**0.79**	4.32	**0.67**	8.62
Ours	0.68	**3.06**	**0.74**	**4.41**	0.69	**8.36**	0.76	**2.45**	0.66	**6.50**
AUTOCAM [51]	*0.56*	*0.25*	*0.56*	*0.71*	*0.47*	*0.55*	*0.73*	*0.15*	0.51	*0.66*
RCNN+BMS.	0.25	37.5	0.2	30.8	0.22	32.4	0.24	40.5	0.2	25.27
RCNN+Motion.	0.56	34.8	0.47	26.2	0.42	25.2	0.72	31.4	*0.54*	25.2

	Skateboarding	Parkour	BMX	Dance	Basketball
Comparison	win / loss	win / loss	win / loss	win / loss	win / loss
vs AUTOCAM	34 / 2	35 / 1	31 / 5	34 / 2	36 / 0
vs Ours **w/o Regressor**	28 / 8	29 / 7	26 / 10	31 / 5	34 / 2
vs human	15 / 21	10 / 26	7 / 29	14 / 22	7 / 29

[51] Su et al. *Pano2vid: Automatic cinematography for watching 360 videos*. In ACCV, 2016.

Deep 360 Agent

Hu et al. CVPR 2017 Oral with NVIDIA Research

Collaboration with AILabs.tw

https://drive.google.com/file/d/0B1ZLJz8x8mnqN2w0TEdkV1o4MGc/view

財團法人台灣人工智慧發展基金會
感謝以下贊助廠商

經緯航太科技股份有限公司　　國家高速網路與計算中心

南部科學工業園區管理局　　孫民老師 Vision Science Lab

台南市政府資訊中心　　國立成功大學航太所賴維祥主任團隊

Target Driven Navigation: Knowledge Based Exploration with Online Curiosity Guidance

Ching-Yao Chuang

National Tsing Hua University

Sanja Fidler

University of Toronto

Min Sun

National Tsing Hua University

CVPR 2017 Workshop on Visual Understanding Across Modalities

Winner of THOR Challenge (discrete)

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

CVPR July 21-26 HONOLULU 2017


THOR
Challenge

THOR Challenge

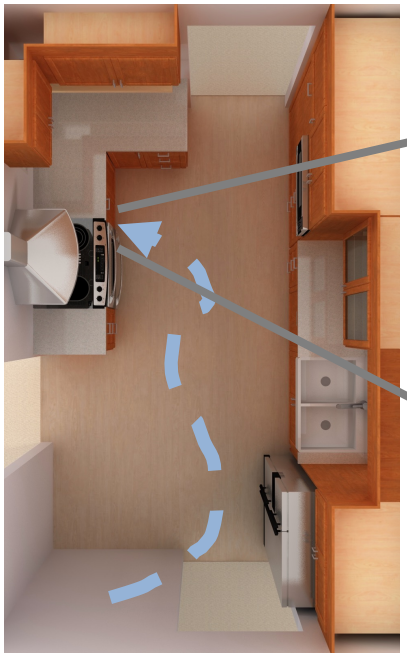
- Navigate and find objects in a virtual environment.

Target image

Observation

Agent

Take actions
Discrete/Continuous



Target positions

Zhu et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In ICRA, 2017.

Qualitative Result

Outline

- Interact through **Language**
  - Video Title Generation
  - Transferring Sentence Style
- Interact through **Actions**
  - Deep 360-degree Pilot
  - Target Driven Navigation
- Interact through **Attacks**
  - Adversarial Attacks

Adversarial Attacks

Normal

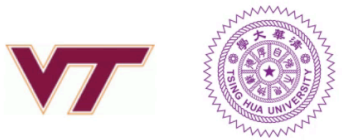
Attacked

http://yclin.me/adversarial_attack_RL/

Adversarial Attacks

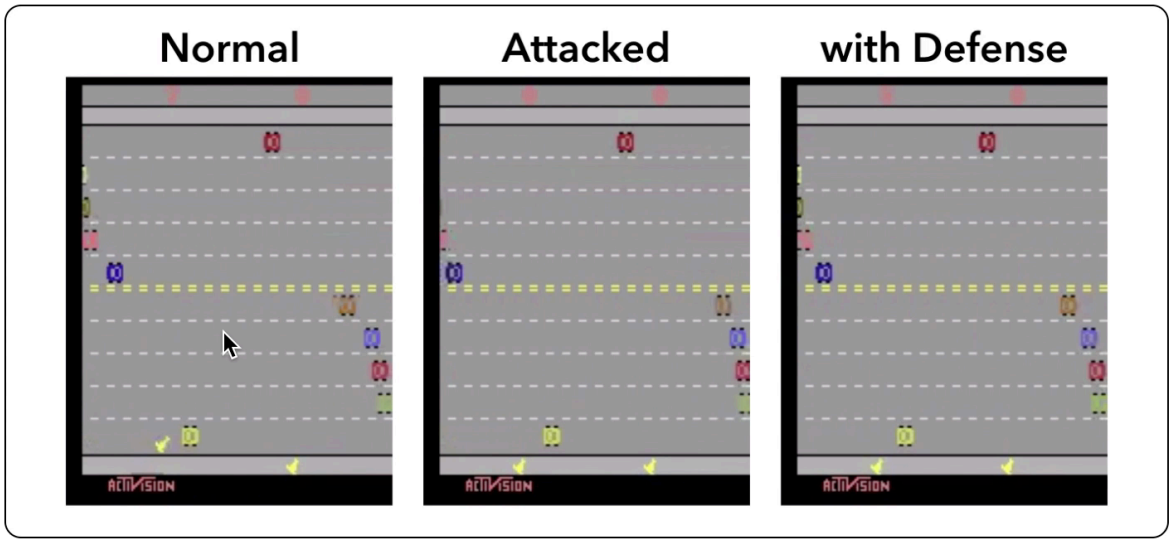
Lin et al. ICLR workshop 2017 and IJCAI 2017

Defense



Detecting Adversarial Attacks on Neural Network Policies with Visual Foresight

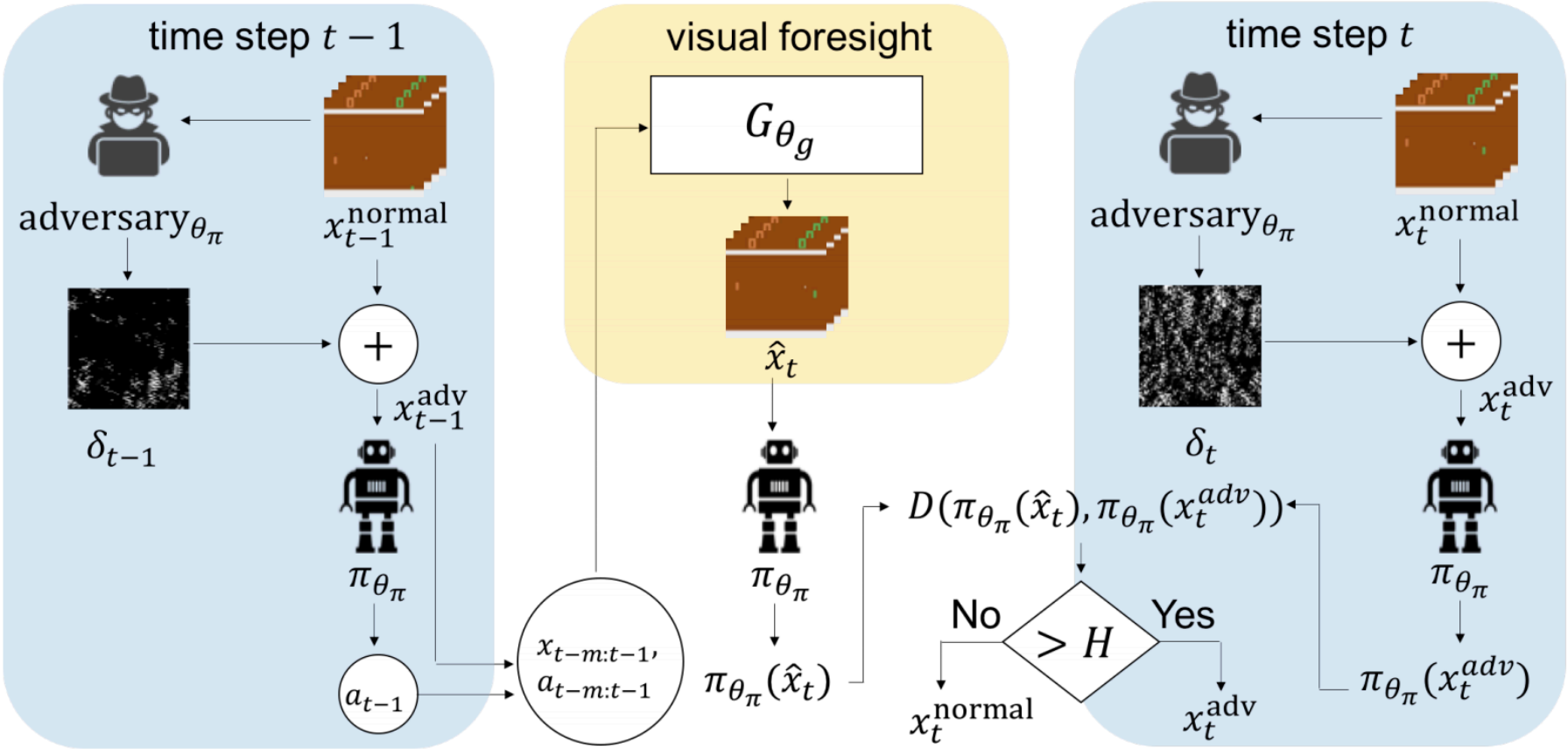
Yen-Chen Lin Ming-Yu Liu Min Sun Jia-Bin Huang



(Click on image to see how our defense work.)

http://yclin.me/RL_attack_detection/

Defense


time step $t-1$
visual foresight
time step t
adversary$_{\theta_\pi}$
x_{t-1}^{normal}
G_{θ_g}
$\hat{x}_t$
adversary$_{\theta_\pi}$
x_t^{normal}
δ_{t-1}
x_{t-1}^{adv}
π_{θ_π}
a_{t-1}
$x_{t-m:t-1}$,
$a_{t-m:t-1}$
π_{θ_π}
$\pi_{\theta_\pi}(\hat{x}_t)$
$D(\pi_{\theta_\pi}(\hat{x}_t), \pi_{\theta_\pi}(x_t^{adv}))$
No
$> H$
Yes
x_t^{normal}
x_t^{adv}
δ_t
x_t^{adv}
π_{θ_π}
$\pi_{\theta_\pi}(x_t^{adv})$

Summary

- Interact through **Language**
  - Video Title Generation
  - Transferring Sentence Style
- Interact through **Actions**
  - Deep 360-degree Pilot
  - Target Driven Navigation
- Interact through **Attacks**
  - Adversarial Attacks

增強學習初探以及最新發展趨勢

活動簡介

增強學習 (Reinforcement Learning) 近年來搭配深度網路，讓電腦在學習打電動、下圍棋、以及操控機器人的應用上獲得超乎預期的成果。

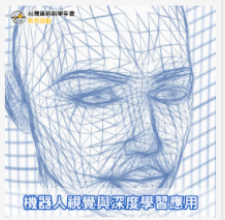
本課程從經典的 Markov Decision Process (MDP) 開始，先介紹傳統增強學習的算法，之後逐步講解著名的 Deep Q-Network (DQN) 以及 Policy Gradient (PG) 演算法。最後分享相關領域最新研究發展以及使用增強學習的小訣竅。

- 日期 2017-12-16(六)
- 地點 中央研究院人文社會科學館
- 票價 $1980 元

2017/11/7 中午12:00 開放報名
2017/11/28 中午12:00 截止報名

更多資料科學相關課程

一天搞懂對話機器人



機器人視覺與深度學習應用

What's happening in

人工智慧與機器學習在推薦

無所不在的自然語言處理—

http://foundation.datasci.tw/reinforcement-learning-171216/

Thanks!

VSLab