Machine Intelligence for Large-Scale Image/Video Data Streams

— Advancing Deep Neural Networks for Emerging Applications

Winston H. Hsu (徐宏民)

National Taiwan University, Taipei

November, 2016

Office: R512, CSIE Building
Communication and Multimedia Lab (通訊與多媒體實驗室)
http://www.csie.ntu.edu.tw/~winston

Dr. Winston Hsu (徐宏民) – Short Bio

- Professor in NTU CSIE and GINM, since Feb. 1, 2007
  - Affiliated with Communication and Multimedia Lab (CMLab)
- PhD from Columbia University, New York, 2007
- **4 years in (startup-period) CyberLink Corp.** (訊連科技)
  - Founding Engineer, Project Leader, and RD Manger
- Recognitions & Awards
  - 3500+ Google citations; H-index: 27; i10-index: 51
  - Director for **NVIDIA AI Lab** (NTU), AE for IEEE Trans. on Multimedia; AE for IEEE Multimedia Mag., Organizing Committee for ACM Multimedia 2010/2013/2015/2016, IEEE/ACM Senior Member, MSR Visiting Researcher (2014), Visiting Researcher IBM Watson (2016)
  - Awards: 2011 Ta-You Wu Memorial Award (Young Researcher), FIRST PRIZE in ACM Multimedia Grand Challenge 2011, **FIRST PLACE in MSR-Bing Image Retrieval Challenge** 2013, **Microsoft Research Award** 2009/2012/2014/2015, 2013 National Outstanding IT Elite Award, 2012 NTU EECS Academic Contribution Award (top 3%), etc.

Globally Competitive for Our Research Team

- Recent report by Wealth Magazine (財訊雙週刊)
- Research developments in AI (data learning in large-scale multimodal data streams)
- How we have strived hard to keep our group competitive in the global research communities.
- Our PhD alumni had received offers from the US-based research labs

徐宏民的資工系實驗室 盛產頂尖人才

台大人工智慧幫 全球搶挖角

過去三年，IBM、GE、微軟搶人工智慧人才搶到台灣來，台灣畢業的土博士，挖到美國工作，年薪十二萬美元起跳，讓自己站上國際舞台，讓能力被看見，讓他們身價從此不同。

文／林宏達

台大資工系教授徐宏民，在多媒體和人工智慧方面相關研究，頗受國際大廠青睞。

在美國，人工智慧（AI）相關人才炙手可熱，一名懂人工智慧技術的博士生，一畢業，年薪至少十二萬美元（約三八〇萬元台幣）起跳。這股人工智慧挖角風，竟跨海吹進台大。

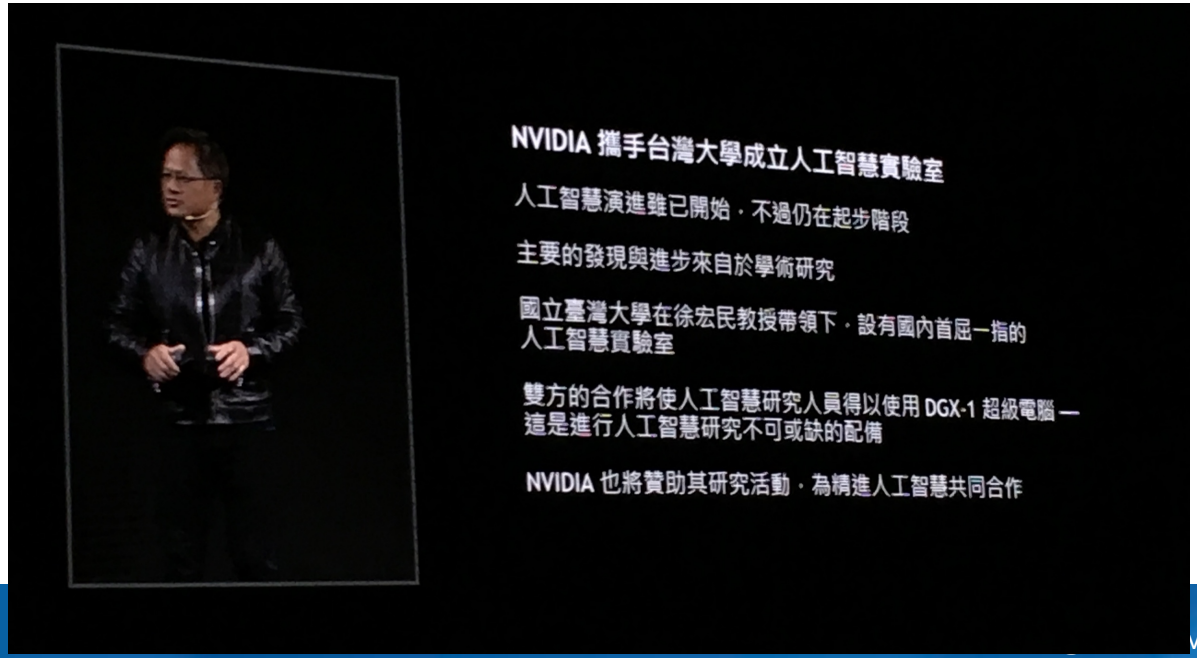
過去三年，台大資工系徐宏民教授的實驗室訓練出來的台灣「土」博士，已有四人分別被美國IBM Spark技術中心、奇異全球研究中心、美國微軟和位於矽谷的Fuji Xerox研究中心挖走，還有一位即將畢業的博士生，已被谷歌和微軟看上，爭相邀請加入團隊。

九月底，NVIDIA執行長黃仁勳宣布和台大合作成立人工智慧實驗室（Nvidia AI

http://www.wealth.com.tw/article_in.aspx?nid=9222

Awarded "NVIDIA AI LAB" – The 1st in Asia, the 4th in the World (GTC Taipei, September 21, 2016)

- Video announcement by NIVIDIA CEO/Co-Founder Jen-Hsun Huang
  - https://www.youtube.com/watch?v=yjhj7bAj9hs#t=57m16s
- For the project, "DeepTutor" – question and answering over large-scale multimodal data streams
- The 4th NVIDIA AI Lab in the world; right after Stanford, Berkeley, and OpenAI

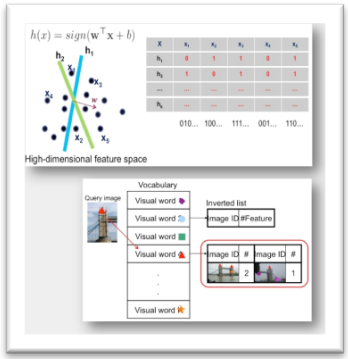


Motivations – Numerous Cameras in Different Forms; and Keep Growing …

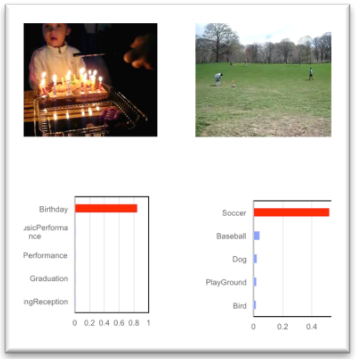
Ongoing Research Projects (Selected) –

More Details and Demos in http://www.csie.ntu.edu.tw/~winston/

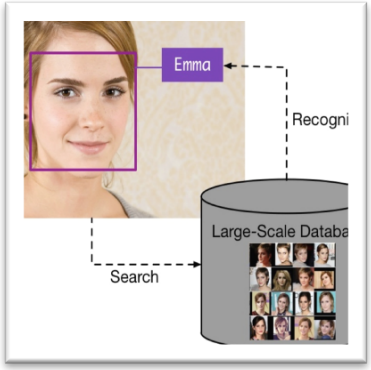
facial/clothing attribute detection/search



web-scale indexing & feature learning

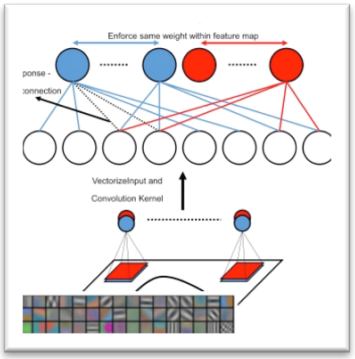


large-scale photo/video recognition

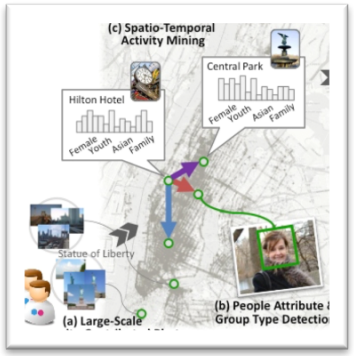


web-scale facial image retrieval

mobile visual recognition

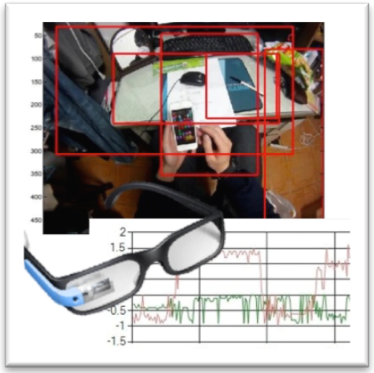


multimodal deep neural network



social media mining

big data analytics and visualization



first-person/wearable cameras

consumer photo retrieval

Image Search by Semantic Understanding –
First Place in MSR-Bing Image Retrieval Challenge 2013

http://web-ngram.research.microsoft.com/GrandChallenge

- Task: Online system (< 12 seconds) to score on each image-query pair that reflects how relevant the query could be used to describe the given image;

drones ?

suri and katie cruise ?

dollar bill ?

- Hosted by **Microsoft Research (Redmond)** and **Bing**
- Dataset: **23M** click logs (query, image, #click) for training set and 77450 image-query pairs for online test

Rank	Team	Name	DCG@25 *	Latency	EvaluationInProgress
1	NTU MIRA	ImgMatch	0.537727656637134	71753.064	False
2	BYRFTRD	boostLearn	0.530992534240603	500928.602	False
3	orange	learn_RF	0.516213951576234	301863.669	False
4	FTRDBJ	Learn_RF	0.507161160461002	231213.845	False
5	NLPR_MMC	Pie	0.503254031251408	446629.263	False
6	NLPR_MMC	TwoFusion	0.50166546329035	121300.56	False

demo

Product Inquiry/Recommendation by Mobiles (2009)

- Product price/information inquiry by mobile phones
- Experienced with indexing high-dimensional & large-scale data

[Lin et al., ICIP'09, Chen et al., JVCI'10]

Amazon Flow

Google Goggle

Pinterest Visual Search

Alibaba Pailitao

demo

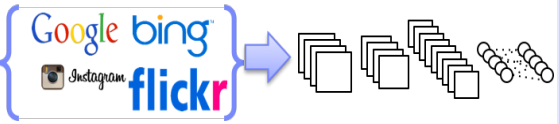
Large-Scale Attribute-based People Search – Search by Impression

[Lei&Hsu, ACM MM 2011]
[Lei&Hsu, SIGIR 2012]

- Search by impression – searching people-related photos by graphically describing the search intentions
- FIRST PRIZE in ACM Multimedia Grand Challenge 2011

Ongoing Projects in Image/Video Analytics with Deep Convolutional Neural Networks

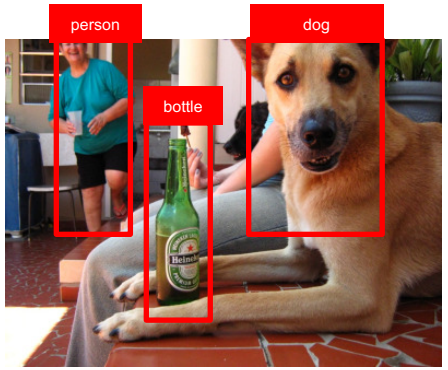
- Goal – Devise **effective** and **efficient learning** methods for **scalable** visual analytic platforms, **applicable** for emerging industry applications



auto. training data for CNN

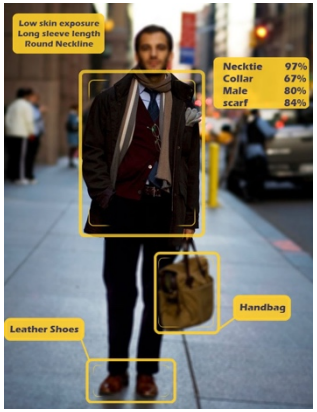
scene cat.

photo annotation



object detection

fine-grained recognition



clothing attributes

Male
African
Glasses
Youth
Smiling
...

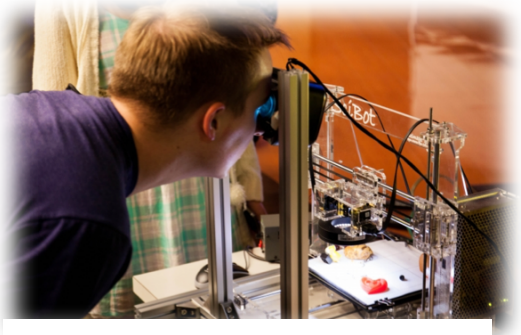
facial attributes

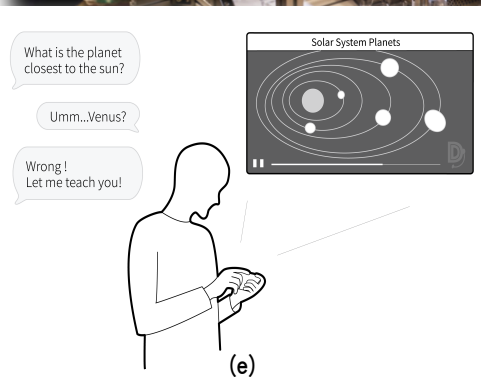
vehicle attributes

video events

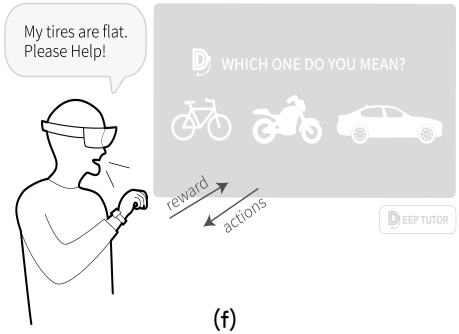
drone AR

DeepTutor for Multimodal Question and Answering





(e)



(f)

Travel	Sentiment	Shopping	Smart City	…	**Education**	Healthcare	Surveillance	Automobile	Robotics

QA Interface (Reinforced + Augmented)

Multimedia QA Engine

Supervised QA → Proactive QA → Self-taught QA → Deep Tutoring

Efficient, Large-Scale, and Multimodal Memory Representations

Scalable Deep Learning Framework

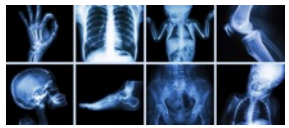
- Multimodal and joint
- Semi-/un- supervised
- Video learning
- Transfer learning
- Scalable platform
- ….

- Memory networks
- Reinforcement
- Attention model for AR
- Deep segmentation
- Deep user modeling
- …

- Hashing for memory networks
- Multimodal memory networks
- Captioning
- Zero-shot query
- Auto. training data acquisition
- …

(diverse media streams)

YouTube OpenData



coursera

UDACITY

Visiting Scientist – Cognitive Computation for IBM Watson AI (New York, USA)

- The first movie trailer generated by AI system (Watson)
  - One of the researchers in the team of three
- Demo video: https://www.youtube.com/watch?v=gJEzuYynaiw
- News
  - "Watson helped make a trailer for a horror movie about AI," Engadget
    - https://www.engadget.com/2016/09/01/ibm-watson-movie-trailer-morgan/
  - "A computer built this trailer for a horror movie about an evil AI," Mashable
    - http://mashable.com/2016/09/01/morgan-watson-ai-trailer
  - ….

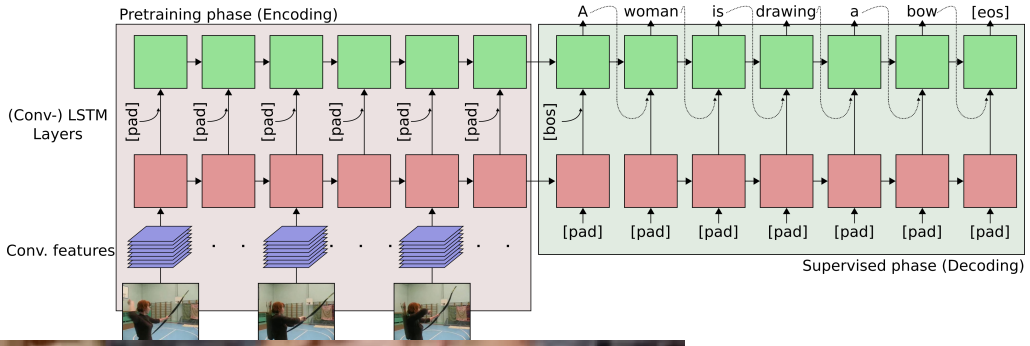
Image/Video Cognition (Machine Perception)

- Problem definition: Given a video (image), describe it in natural language
- Motivations
  - Understanding high-level semantics and intention from video collection
  - Leveraging multiple modalities such as video, time, text, etc., in the unified deep learning framework
  - Enabling technology for video event detection, surveillance, live content filtering, robotics, social media mining, HCI, **question and answering**, etc.

A box of doughnuts on a table <eos>

A man and a woman are kissing <eos>

Image/Video Cognition (Machine Perception) – Tentative Results

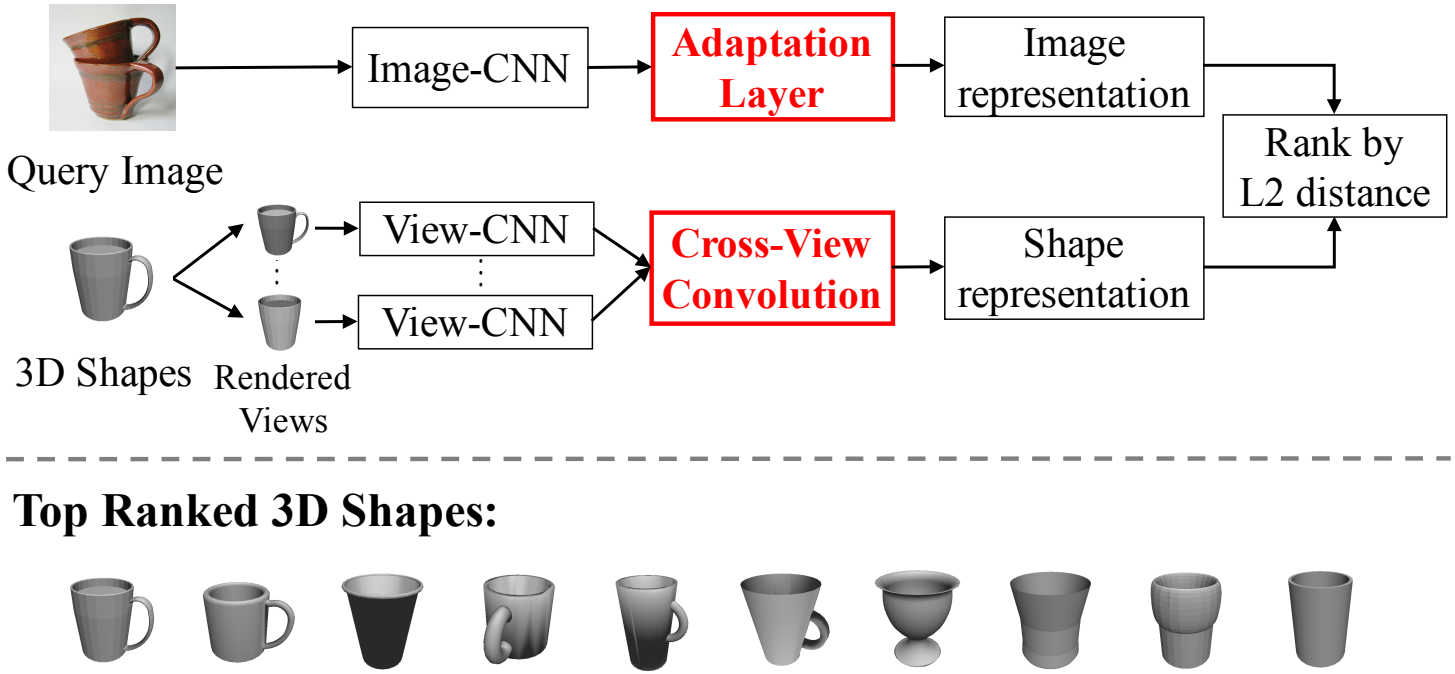


A woman is pouring a bowl of dough and another woman is making something <eos>

Image-based 3D Model Retrieval – Retrieving semantically Related 3D Models by Image

[Lee et al., submitted, 2016]

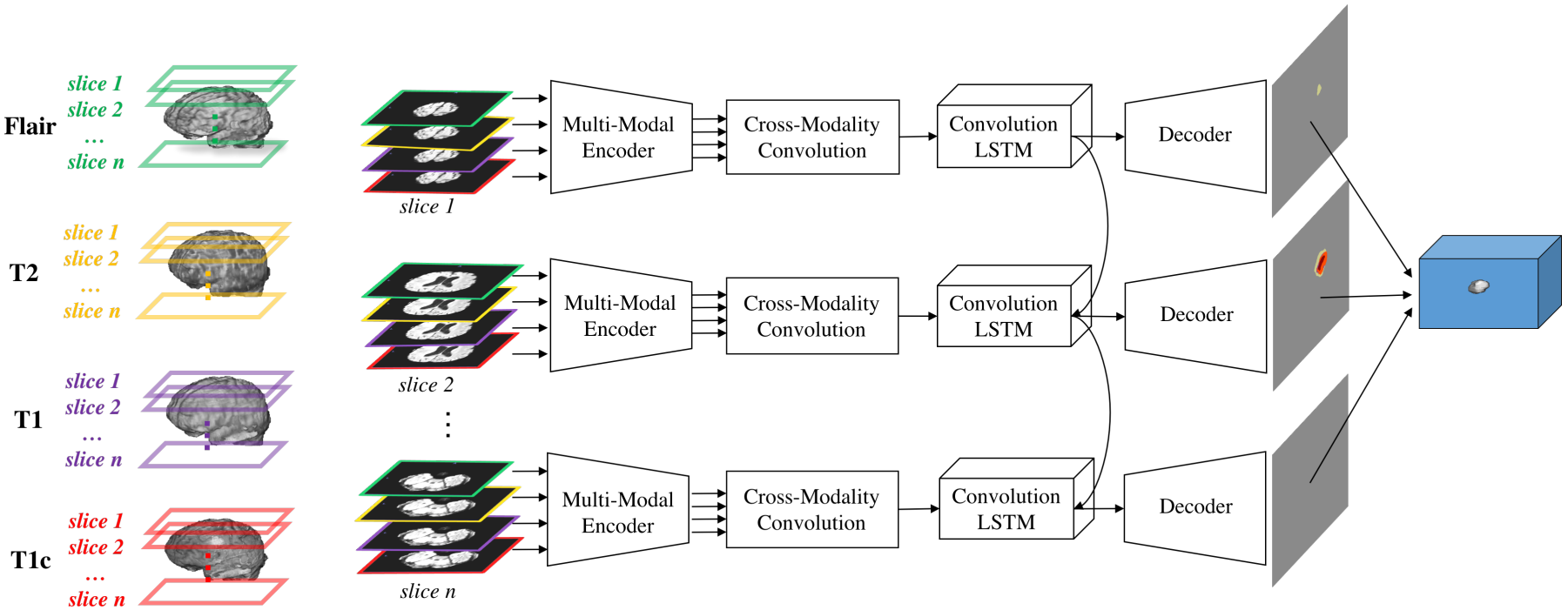
- Novel proposal – End-to-end deep neural networks for cross-domain and cross-view learning and ranking
- Impacts: the brand-new problem and significantly outperforming prior neural networks



3D Medical Segmentation by Deep Neural Networks

[Tseng et al., submitted, 2016]

- Novel proposal – Utilizing cross-modal learning in the sequential and convolutional neural networks
- Impacts: Significantly outperforming prior works (e.g., U-Net) in open benchmarks



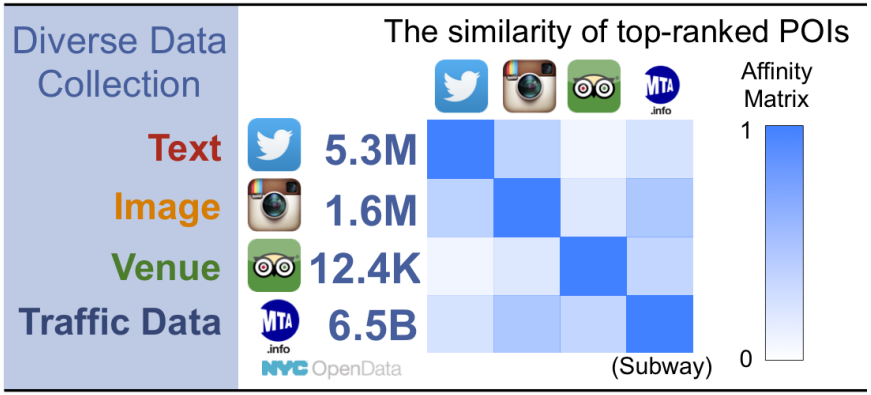
Social Media Mining – Huge Photos/Videos Shared for Human Activities

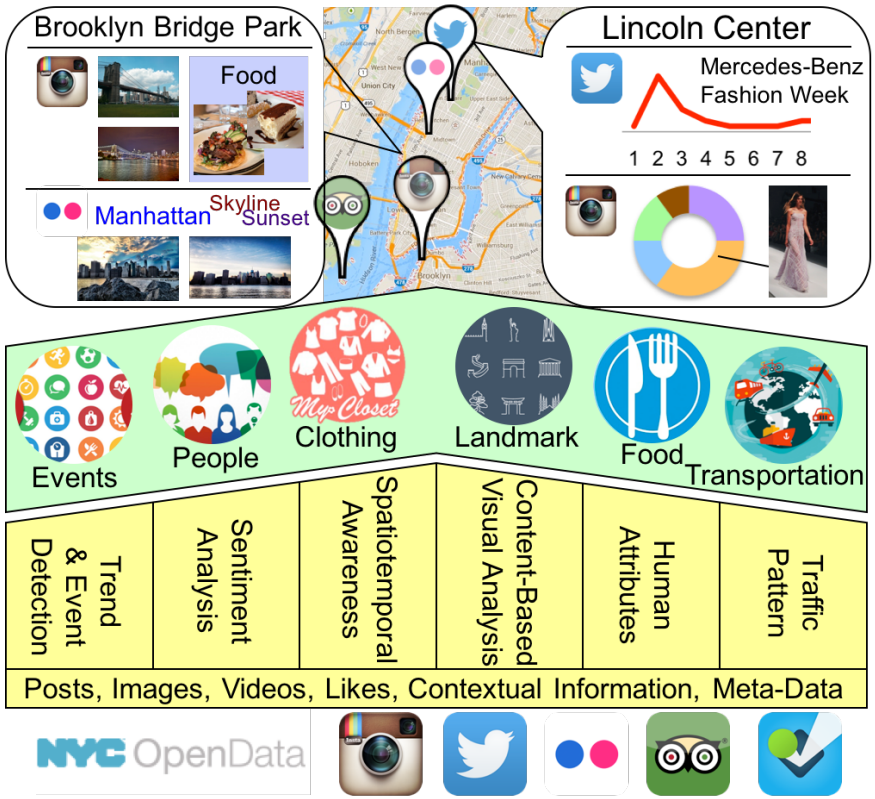
Sampled from 100M Photos from Flickr (24hr)

Discovering the City by Mining Diverse and Multimodal Data Streams – IBM Grand Challenge: New York City 360

[Kuo, ACM MM'14]

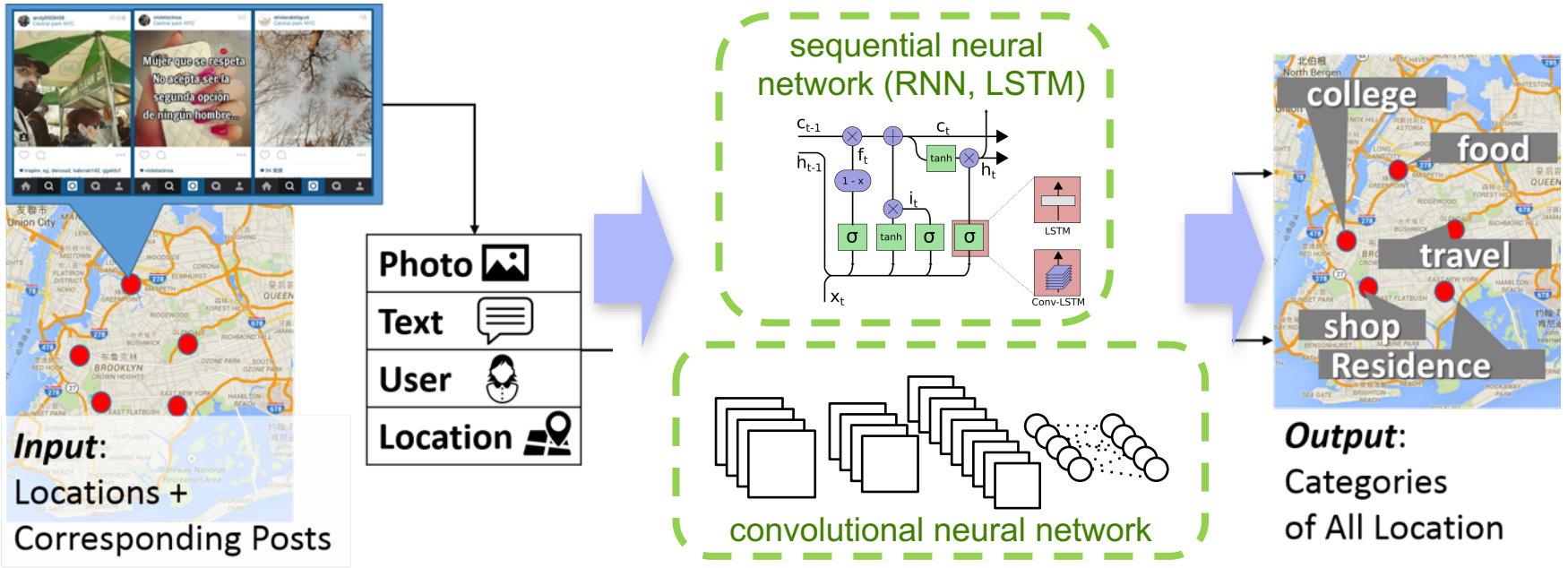
- Exploring and integrating multiple contents and sources for NYC life
- ACM Multimedia 2014 Grand Challenge Multimodal Award





Understand Human Activities from Social Media (e.g., Instagram): Time + Photos + Tags

Microsoft Research

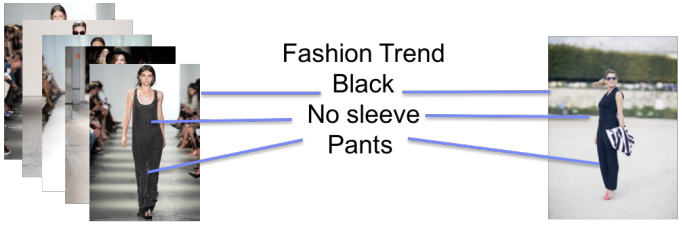


- **Why:** Huge needs in location-based services: advertisement, location understanding, recommendation, city planning, etc.
- **Problem Definition**: Location classification, provided a collection of photos and associated metadata
- **Location Categories (10):** Arts & Entertainment, College & University, Event, Food, Professional Places, Nightlife Spot, Outdoors & Recreation, Shop & Service, Travel & Transport, Residence

Fashion Mining from Social Media by Clothing Attributes – Huge Interest from Fashion Industry

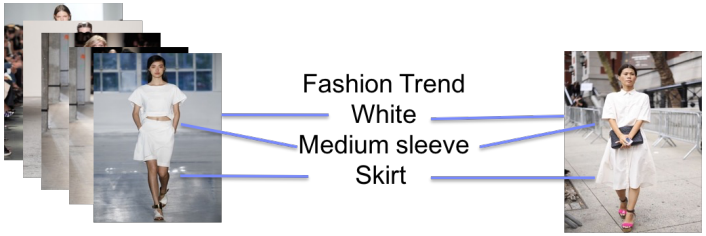
[Chen et al., ACMMM'15]

- Confirmed the influence of fashion shows in daily life
  - 60 clothing attributes
- Widely discussed in social media and news media (NY Post, MIT Tech. Review, Science News, etc.)



2014 New York Fashion Show

2014 Street-chic



2015 New York Fashion Show

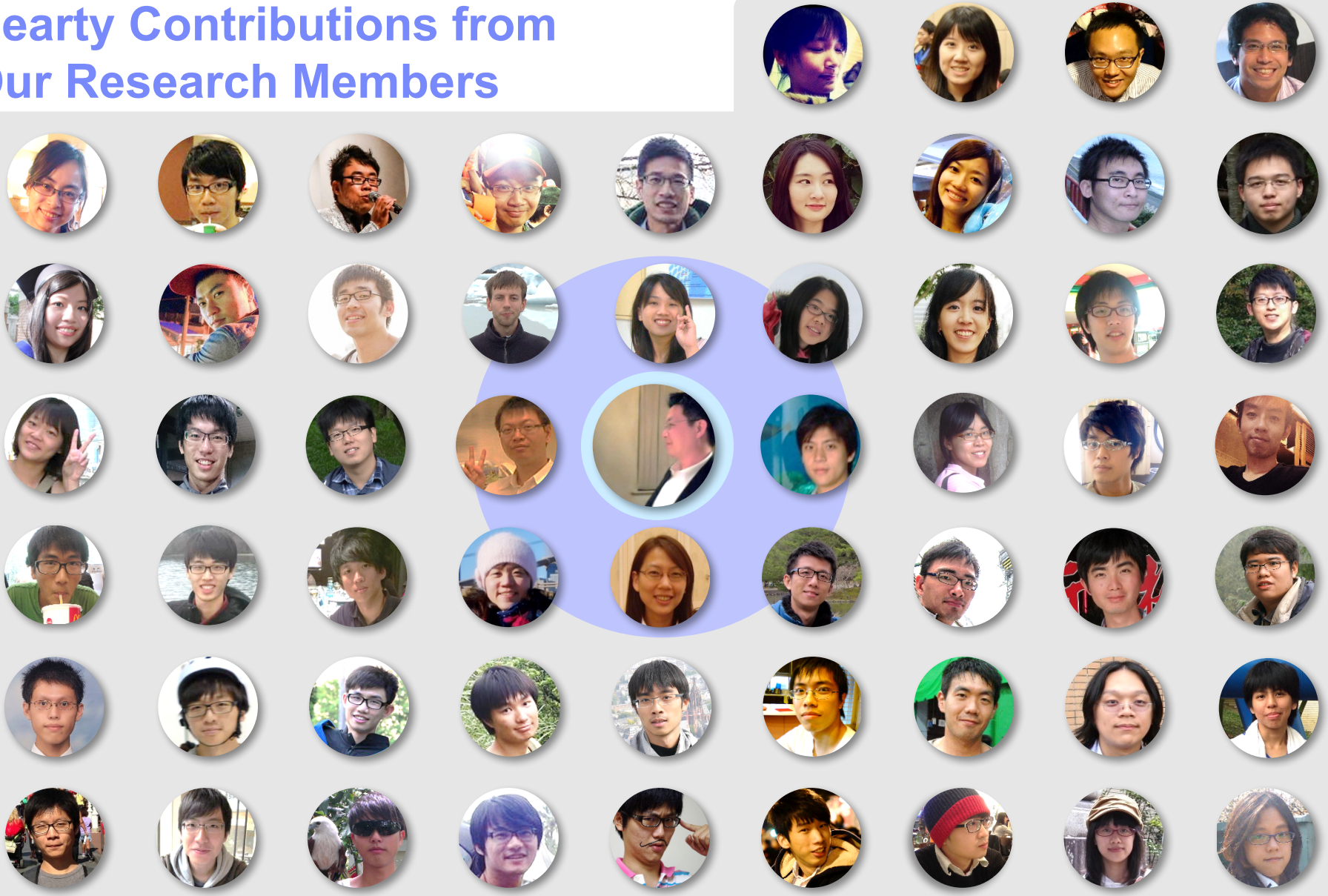
2015 Street-chic

Drone AR – Understanding the Context from Drone Views (Ongoing Project)

Recent Student Awards (selected)
– Working on Essential and Emerging Problems

- FIRST PLACE in MSR-Bing Image Retrieval Challenge 2013
- First Prize for ACM Multimedia Grand Challenge 2011
- ACM Multimedia 2013 Grand Challenge Multimodal Award
- 陳殷盈ACM Multimedia 2012 Doctoral Symposium Best Paper Award
- 郭盈希Microsoft Research Asia Fellowship 2012
- 朱冠宇榮獲「中國電機工程學會102年青年論文獎」第三名
- 博士班學生陳冠婷(102)、陳殷盈(101)、林彥良(101)獲得 「補助博士生赴國外研究（千里馬）」獎助
- 陳柏村榮獲101年度中華民國人工智慧學會碩士論文獎
- 中華電信2011電信創新應用大賽雲端應用校園組亞軍
- 鄭安容榮獲「中國電機工程學會100年青年論文獎」第二名
- 李文瑜榮獲頂尖國際會議SIGIR 2011 Google Fellowship for Women
- 陳殷盈榮獲頂尖國際會議WWW 2011 Google Fellowship for Women
- 郭盈希同學榮獲「中國電機工程學會99年青年論文獎」第二名
- 學生榮獲中華電信2010電信奧斯卡—花博應用組冠軍

Hearty Contributions from Our Research Members



Acknowledgements for Research Sponsors