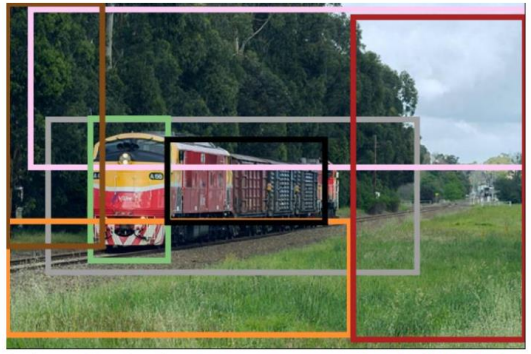
Recent Advances in Vision-and-Language Research

Zhe Gan, Licheng Yu, Yu Cheng, Luowei Zhou,
Linjie Li, Yen-Chun Chen, Jingjing Liu, Xiaodong He

Visual Captioning

A horse carrying a large load of hay and two people sitting on it.



train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background. photo taken during the day. red train car.

- ***Popular Topics***: Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
- ***Popular Tasks***: Image/video captioning, Dense captioning, Storytelling

Visual QA/Grounding/Reasoning

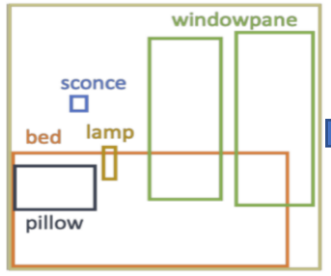
VQA

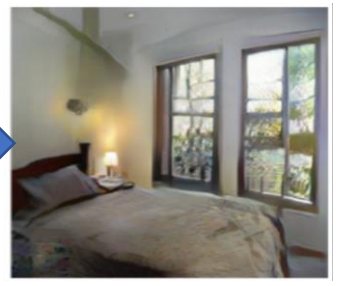
Referring Expressions

- ***Popular Topics***: Multimodal fusion, Advanced attentions, Use of relations, Neural modules, Language bias reduction
- ***Popular Tasks***: VQA, GQA, VisDial, Ref-COCO, CLEVR, VCR, NLVR2

Text-to-image Synthesis

This bird is red with white belly and has a very short beak





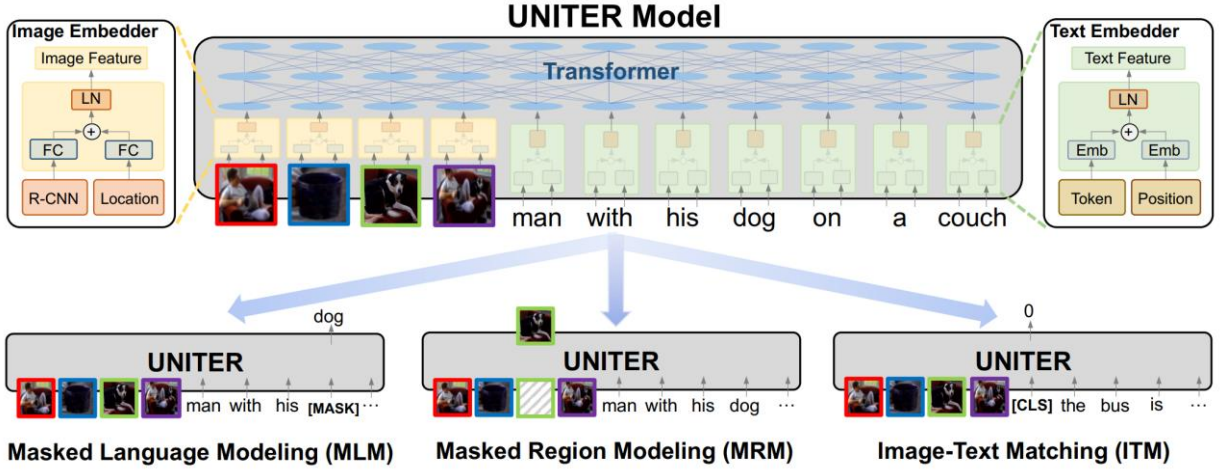
Popular Tasks:

- Text-to-image
- Layout-to-image
- Scene-graph-to-image
- Text-based image editing
- Story visualization

SOTA Models:

- StackGAN
- AttnGAN
- ObjGAN
- …

Self-supervised Learning

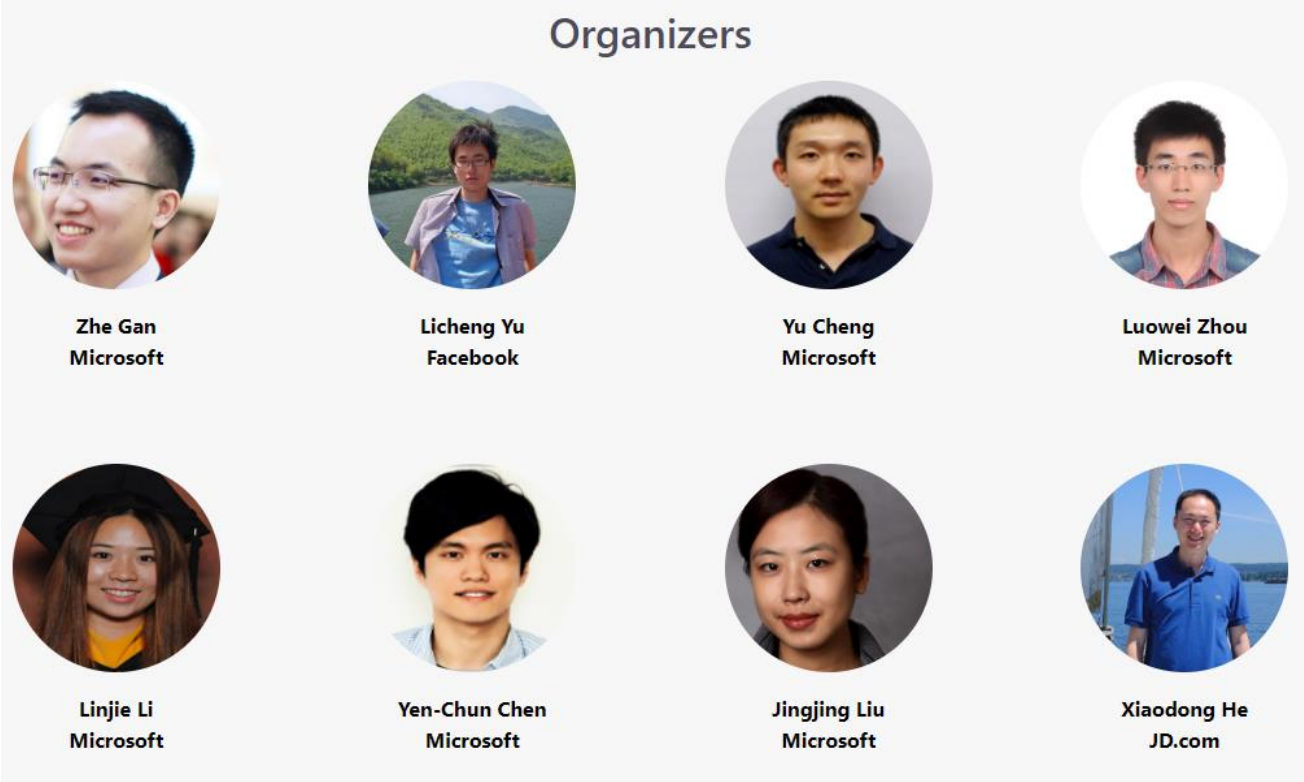


SOTA Models:

- ***Image+Text:*** ViLBERT, LXMERT, Unicoder-VL,UNITER, etc.
- ***Video+Text***: Video-BERT, CBT, UniViLM, etc.

Tutorial Agenda

- 1:15 – 1:25 **Opening Remarks**
- 1:25 – 2:15 **Visual QA/Reasoning**
- 2:15 – 2:30 ***Coffee Break***
- 2:30 – 3:10 **Visual Captioning**
- 3:10 – 3:40 **Text-to-image Generation**
- 3:40 – 4:00 ***Coffee Break***
- 4:00 – 5:00 **Self-supervised Learning**



Tutorial Website: https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/

Session 1: **Visual QA and Reasoning**

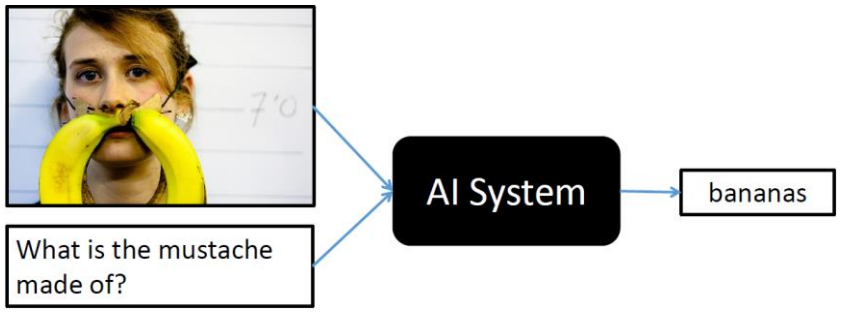
Time:
1:25 – 2:15 PM (50 mins)

Presenter:
Zhe Gan (Microsoft)

Zhe Gan is a Senior Researcher at Microsoft Dynamic 365 AI Research. His current research interests include Vision-and-Language Pre-training and Self-supervised Learning. Zhe obtained his Ph.D. degree from Duke University in 2018, and Master's and Bachelor's degrees from Peking University in 2013 and 2010, respectively. He is an Area Chair for NeurIPS 2020 and 2019, and received AAAI-2020 Outstanding Senior Program Committee Award.

Visual QA/Reasoning/Grounding

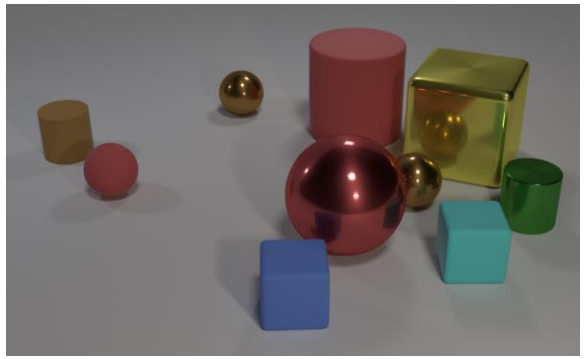
GQA



VQA

VCR

Referring Expressions

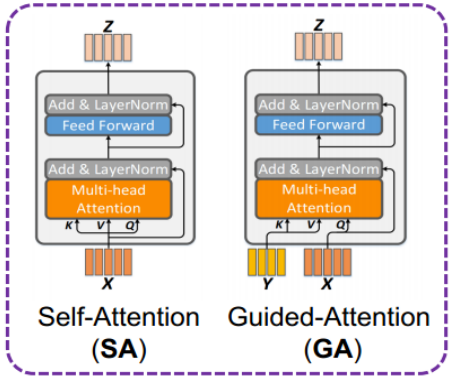


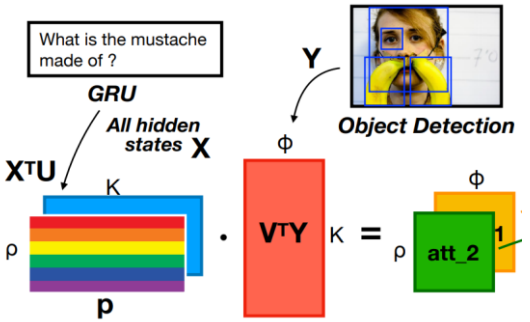
CLEVR

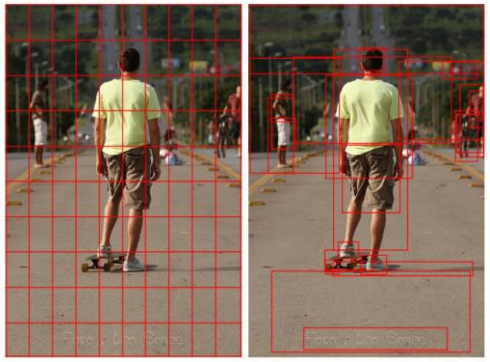
NLVR2

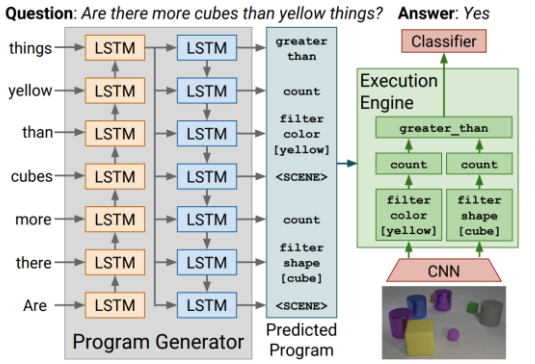
Main Topics

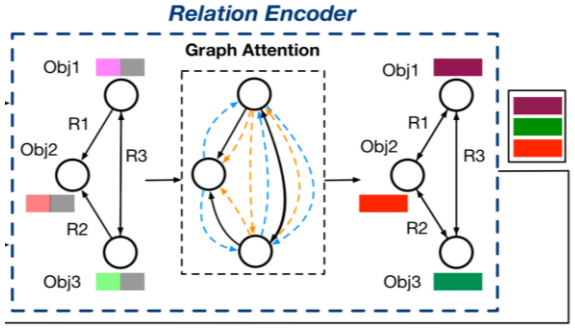
- Advanced attention mechanism
- Enhanced multimodal fusion
- Better image feature preparation
- Multi-step reasoning
- Incorporation of object relations
- Neural module networks
- Language bias reduction
- Multimodal pre-training

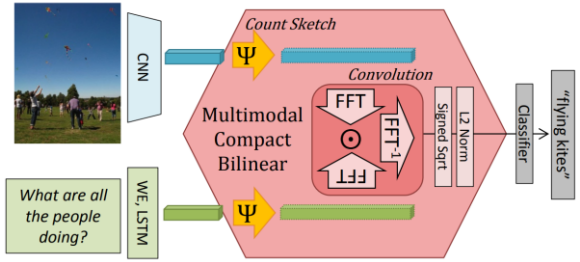












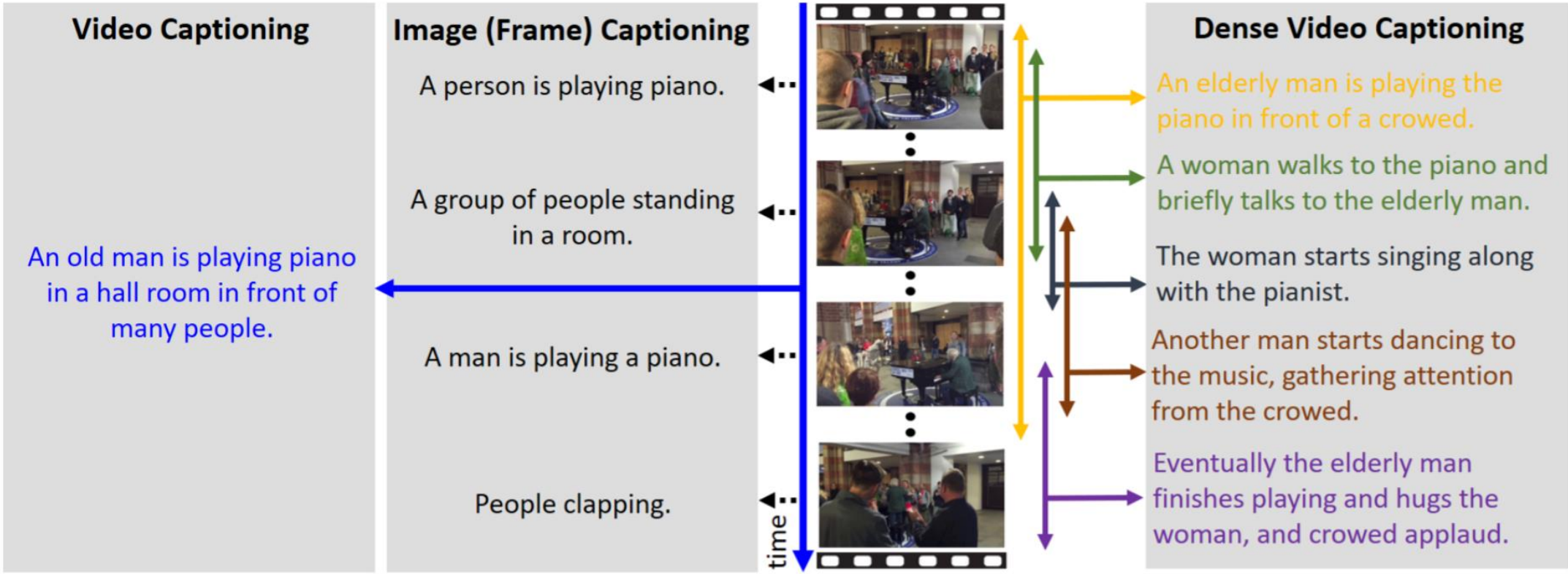
Session 2: **Visual Captioning**

Time:
2:30 – 3:10 PM (40 mins)

Presenter:
Luowei Zhou (Microsoft)

Luowei Zhou is a Researcher at Microsoft. He received his Ph.D. degree in Robotics from the University of Michigan in 2020 and Bachelor's degree in Automation from Nanjing University in 2015. His research interests include computer vision and deep learning, in particular, the intersection of vision and language. He is a PC member/reviewer for TPAMI, IJCV, CVPR, ICCV, ECCV, ACL, EMNLP, NeurIPS, AAAI, ICML etc. and actively organizes affiliated workshops and tutorials.

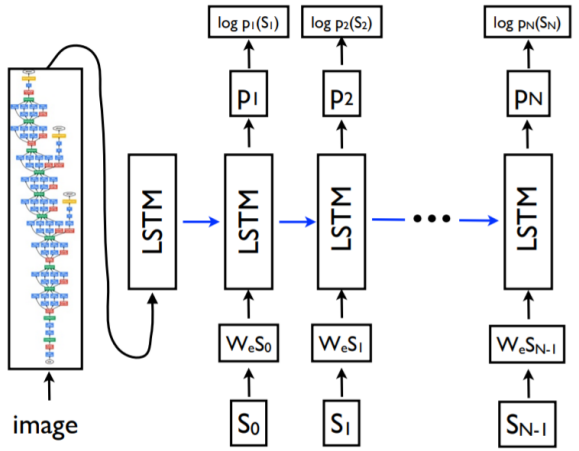
From Images to Videos and Beyond

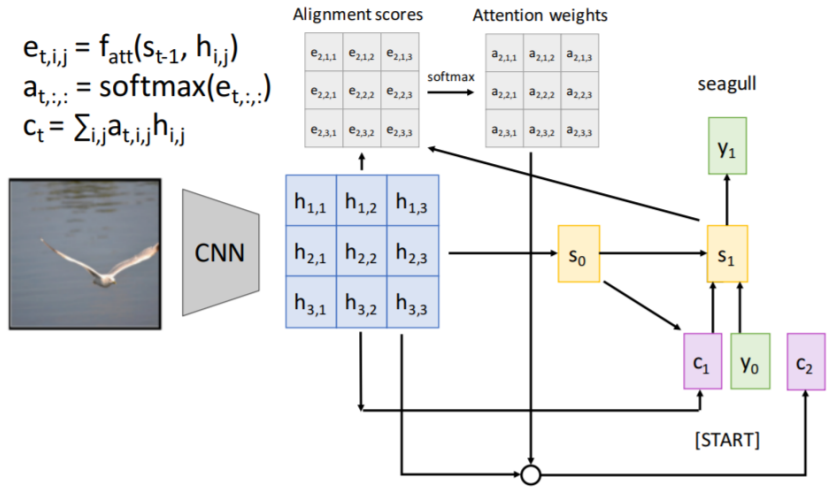


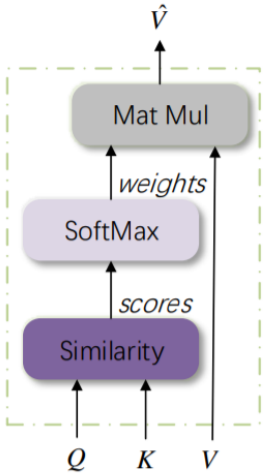
[Figure credit: Aafaq et al., 2019]

Main Topics

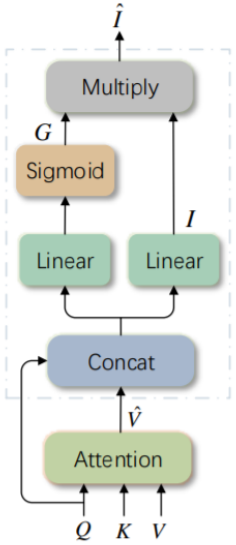
- Show and Tell
- Attention-based
- “Fancier” Attention
- Transformer-based
- Pre-training



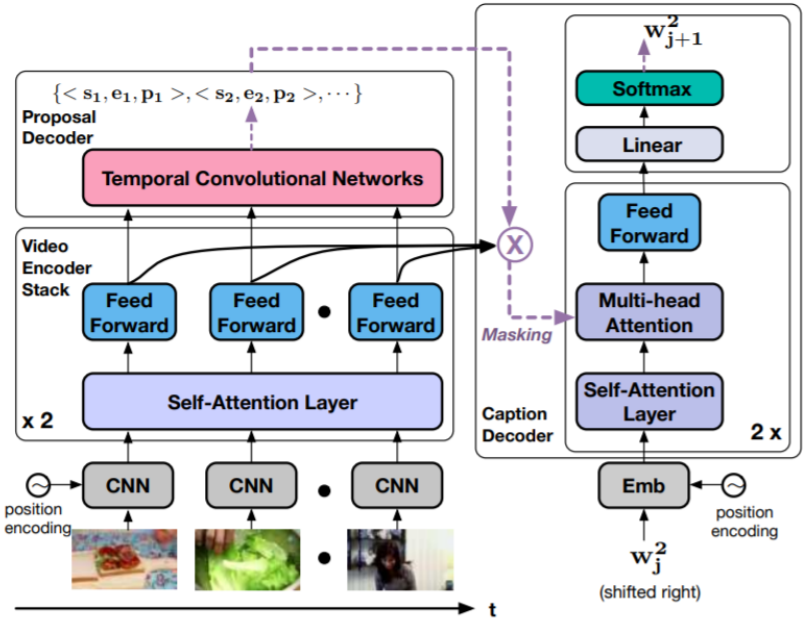




(a) Attention



(b) Attention on Attention



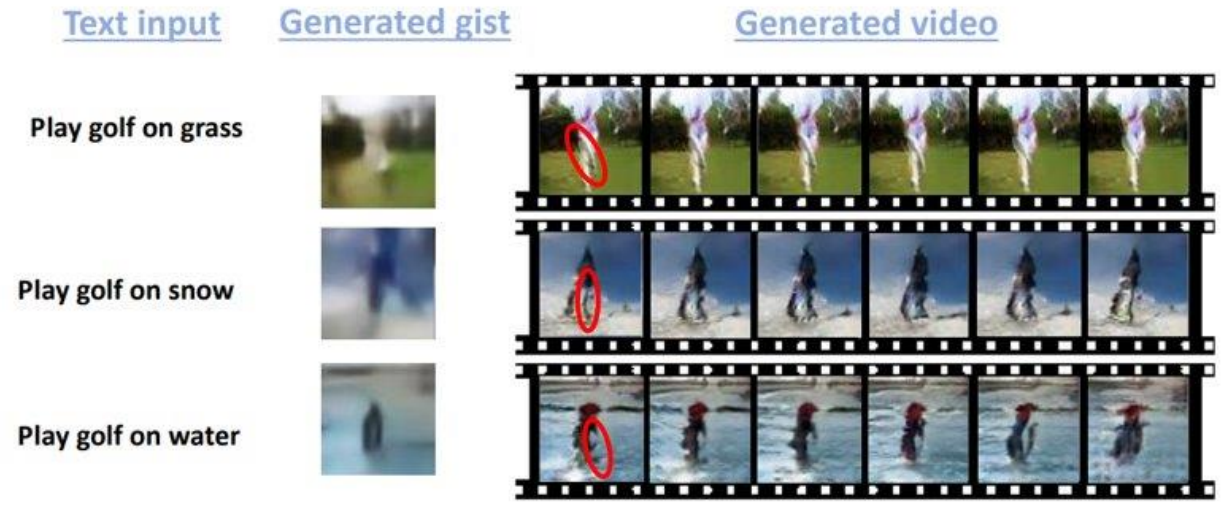
Session 3: Text-to-Image Synthesis

Time:
3:10 – 3:40 PM (30 mins)

Presenter:
Yu Cheng (Microsoft)

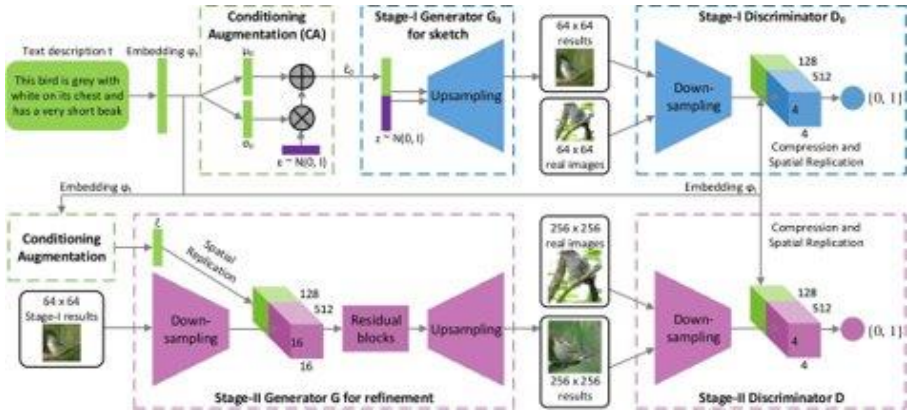
Yu Cheng is a Senior Researcher at Microsoft. Before that, he was a Research Staff Member at IBM Research/MIT-IBM Watson AI Lab. Yu got his Ph.D. from Northwestern University in 2015 and bachelor from Tsinghua University in 2010. His research is in deep learning in general, with specific interests in model compression, deep generative model and adversarial learning. Currently he focuses on using these techniques to solve real-world problems in computer vision and NLP.

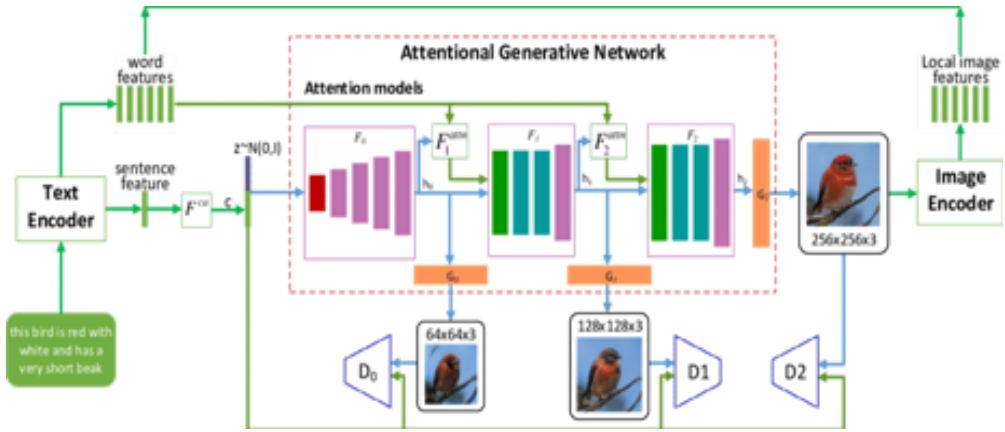
Image and Video Synthesis from Text



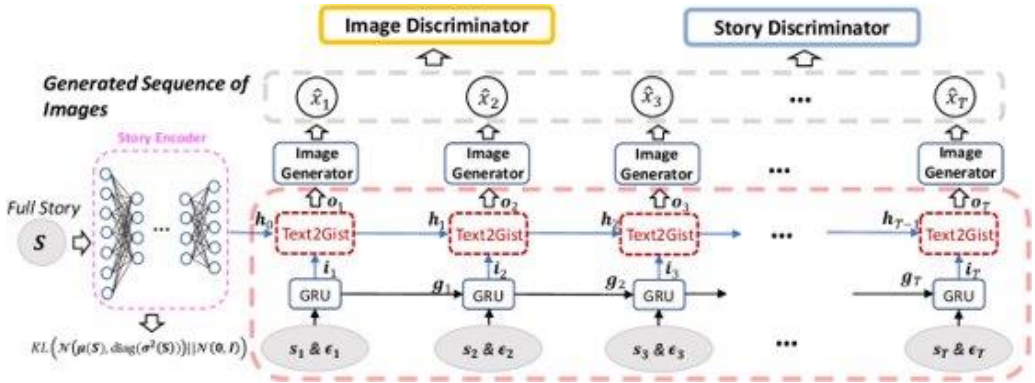
[Figure credits: Zhang et al, 2017; Li et al., 2018]

Main Topics

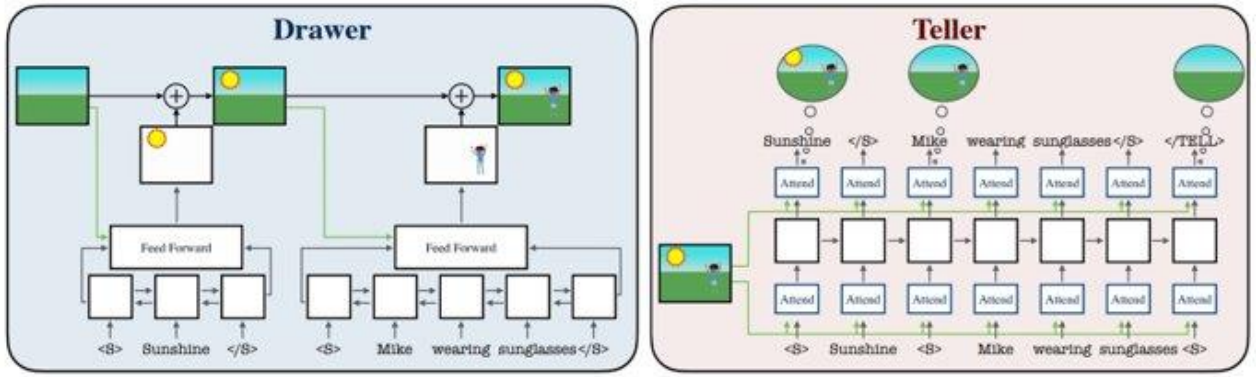




Text-to-Image Synthesis *(StackGAN, AttnGAN, TAGAN, Obj-GAN)*



Text-to-Video Synthesis *(GAN-based, VAE-based)*



Dialogue-based Image Synthesis *(ChatPainter, CoDraw, SeqAttnGAN)*

Session 4: **Self-supervised Learning**

Time:
4:00 – 5:00 PM (60 mins)

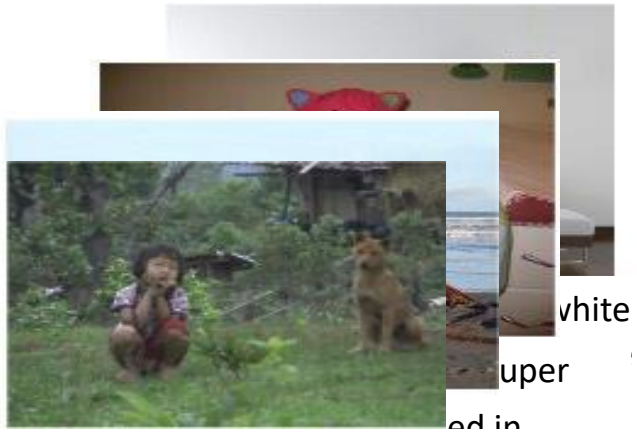
Presenters:
Licheng Yu (Facebook), Yen-Chun Chen (Microsoft), Linjie Li (Microsoft)

Dr. Licheng Yu is a Research Scientist at Facebook AI. Before then, he was at Microsoft Dynamics 365 AI Research. Licheng completed his PhD from University of North Carolina at Chapel Hill in 2019, and got his B.S degree from Shanghai Jiaotong University (SJTU) and M.S degrees from both SJTU and Georgia Tech. During his PhD study, he did summer internships at eBay Research, Adobe Research and Facebook AI Research.

Linjie Li is a Research SDE at Microsoft Dynamic 365 AI Research. Her current research interests include Vision-and-Language pre-training and self-supervised learning. Linjie obtained her Master's degree in computer science from Purdue University in 2018. She also holds a Master's degree in Electrical Engineering from UC, San Diego.

Yen-Chun Chen is a Research SDE at Microsoft. He received his M.S. in computer science from UNC Chapel Hill in 2017, where he focused on NLP and text summarization. He got his bachelor degree in electrical engineering from NTU in 2014. His current research focus is large-scale self-supervised pre-training and its applications.

Self-supervised Learning for Vision-and-Language



Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

Large, Noisy, Free Data

Model

Pre-training Tasks

- Masked Language Modeling
- Masked Region Modeling
- Image-Text Matching
- Word-Region Alignment

VQA | VCR | NLVR2 | Img-Txt Retrieval | Txt-Img Retrieval | Referring Expressions | GQA | Visual Entailment | Image Captioning

Main Topics

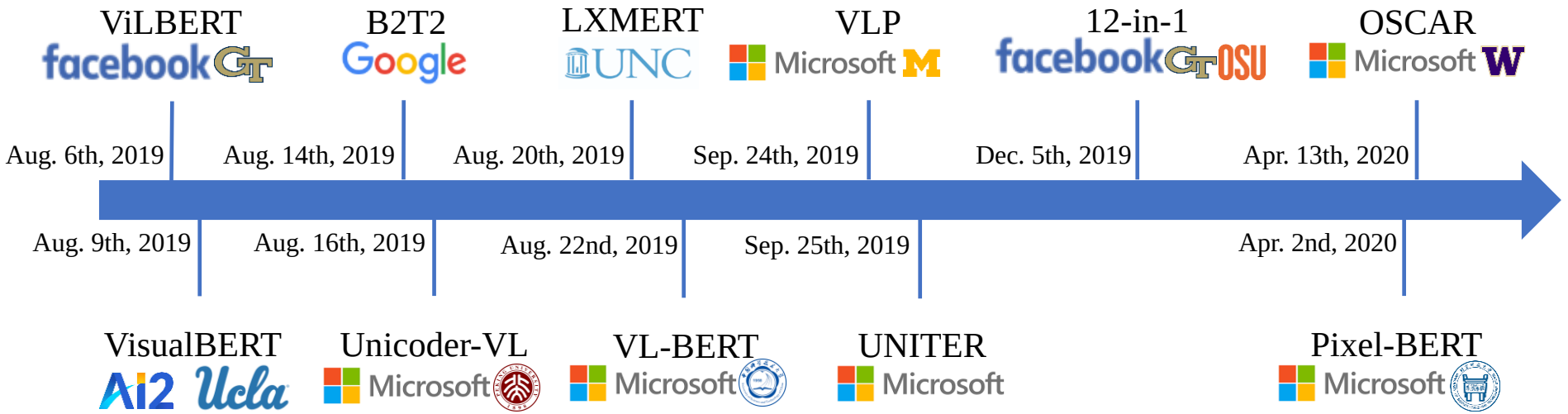
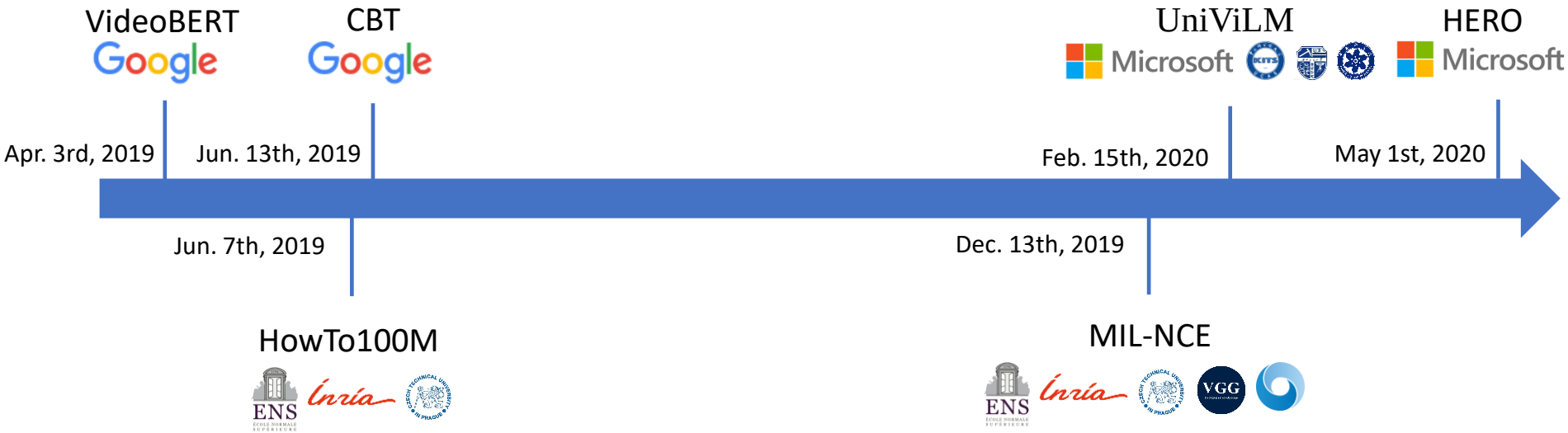


Image Downstream Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning



Video Downstream Tasks

- Video QA
- Video-and-Language Inference
- Video Captioning
- Video Moment Retrieval