Vision and Language
The Past, Present and Future

Jiebo Luo

University of Rochester

2021 IEEE International Conference on Multimedia and Expo

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

Vision-and-Language

A cat is sitting next to a pine tree, looking up

Vision — Language

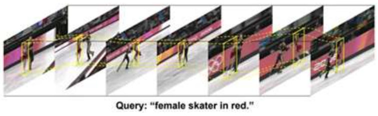
- The intersection of **computer vision** and **natural language processing**
- Multi-modal learning

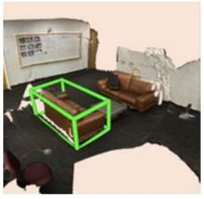
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

A cat is sitting next to a pine tree, looking up

Vision

Language





A cat is sitting next to a pine tree, looking up

Vision

Language

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

A cat is sitting next to a
pine tree, looking up

Vision

Language

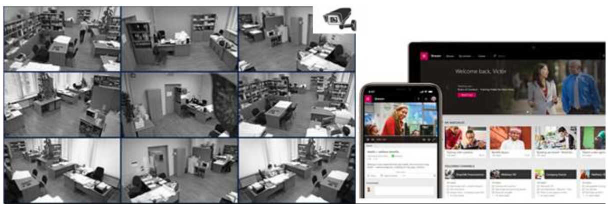
- Vision-language joint representation learning
- General multi-modal learning

Why Language in Vision?

- **Direct applications**

"the water bottle next to the green glass"

- Human-computer-interaction
  - Easy to generate

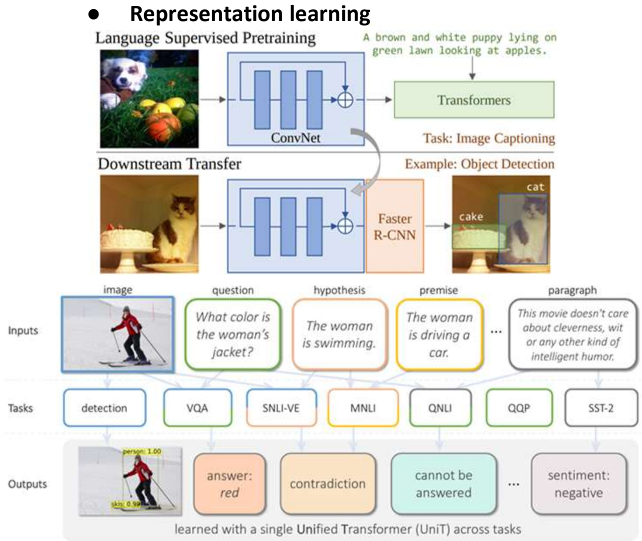


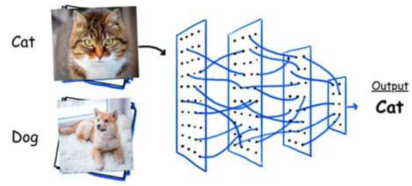
"person wearing black pants walking to the back chair"

- Language-based search
  - Clear specification

Why Language in Vision?

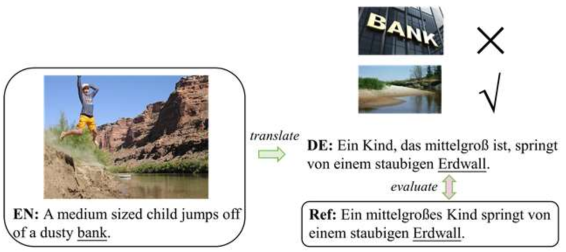
- **Representation learning**

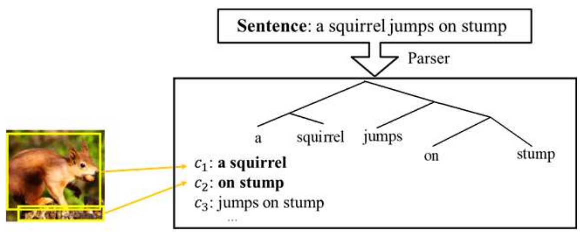




- Visual representation learning with language supervision
- Stronger vision, language, and VL models

Why Vision in Language?





- Multimodal Machine Translation
  - Help solve data sparsity and ambiguity
- Unsupervised Grammar Induction
  - Help induce syntactic structures

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Vision-and-Language Tasks

A cat is sitting next to a pine tree, looking up.

Understanding tasks

Generation tasks

Image retrieval

Visual grounding

A cat is sitting next to a pine tree, looking up.

A cat is sitting next to a pine tree, looking up.

Visual captioning

What is the cat doing

QA/ Reasoning

A cat is sitting next to a pine tree, looking up.

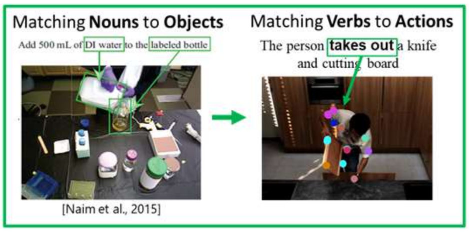
Text-to-image synthesis

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER

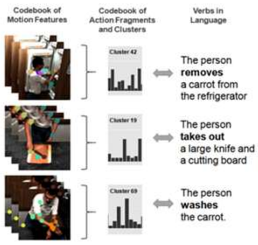
DEPARTMENT OF COMPUTER SCIENCE

Vision-and-Language Research in the Stone Age (Pre-DL)

- Unsupervised alignment of video with text



[Naim et al., 2015]



An overview of the text and video alignment framework

- Motivations
  - Generate labels from data (reduce burden of manual labeling)
  - Learn new actions from only parallel video+text
  - Extend noun/object matching to verbs and actions

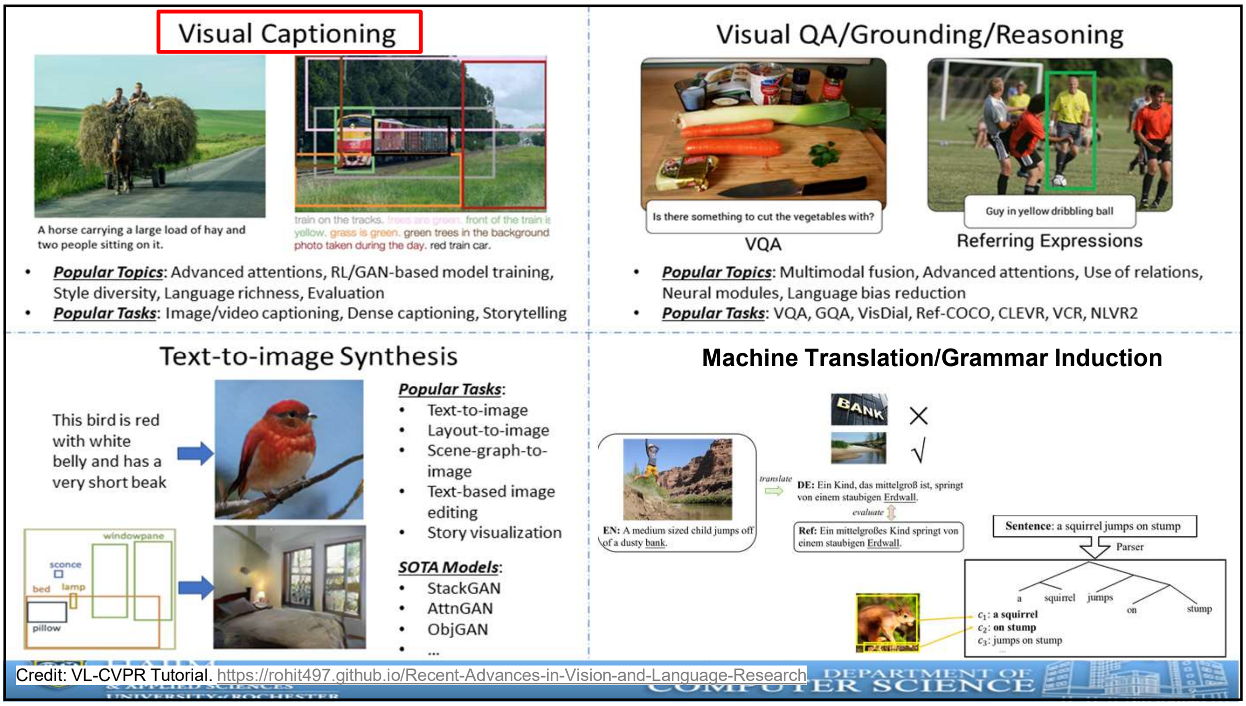
Unsupervised

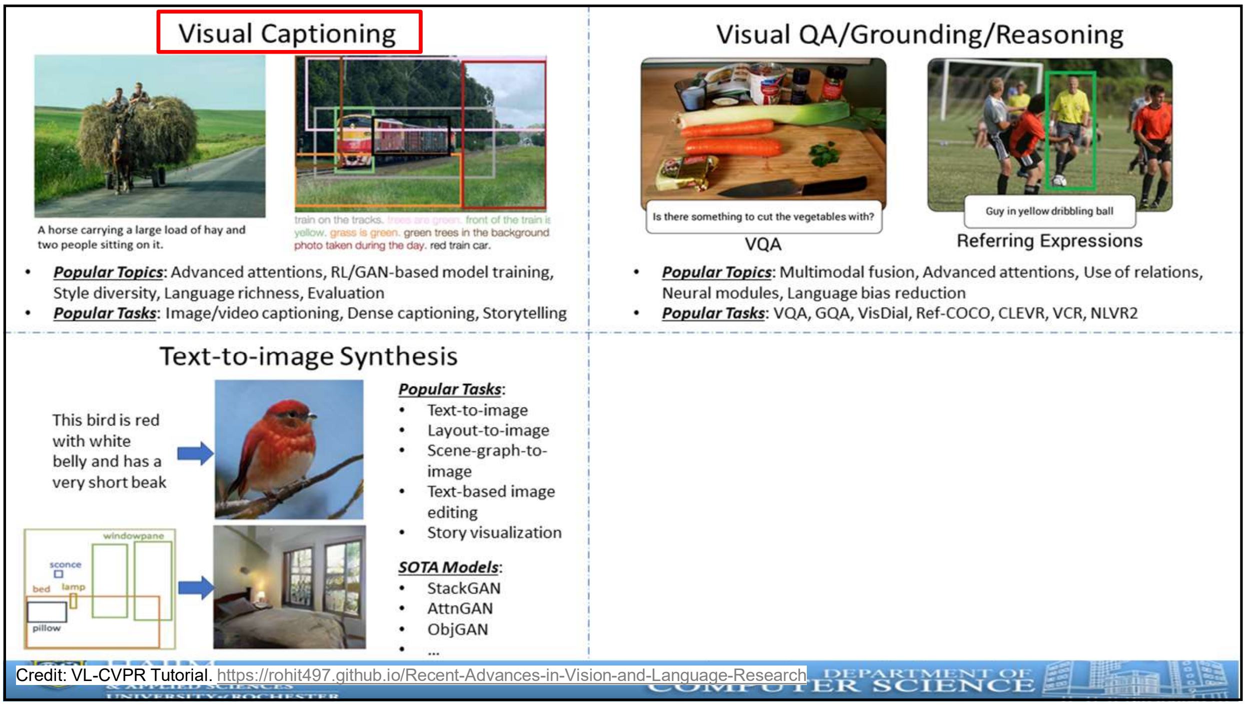
BoW, CRF, DTW, etc.

[1] * Iftekhar Naim, Young Song, Daniel Gildea, Qiguang Liu, Henry Kautz and Jiebo Luo, "Unsupervised Alignment of Natural Language Instructions with Video Segments," AAAI 2014.
[2] * Iftekhar Naim, Young Chol Song, Henry Kautz, Jiebo Luo, Qiguang Liu, Daniel Gildea, and Liang Huang, "Discriminative Unsupervised Alignment of Natural Language Instructions with Corresponding Video Segments," NAACL 2015.
[3] * Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea and Henry Kautz, "Unsupervised Alignment of Actions in Video with Text Descriptions," IJCAI 2016.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER

DEPARTMENT OF COMPUTER SCIENCE


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
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
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
windowpane
sconce
bed
lamp
pillow
Popular Tasks:
• Text-to-image
• Layout-to-image
• Scene-graph-to-image
• Text-based image editing
• Story visualization
SOTA Models:
• StackGAN
• AttnGAN
• ObjGAN
• ...
Machine Translation/Grammar Induction
BANK
EN: A medium sized child jumps off of a dusty bank.
translate
DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.
evaluate
Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.
Sentence: a squirrel jumps on stump
Parser
a squirrel jumps on stump
c1: a squirrel
c2: on stump
c3: jumps on stump
Credit: VL-CVPR Tutorial. https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research
DEPARTMENT OF COMPUTER SCIENCE

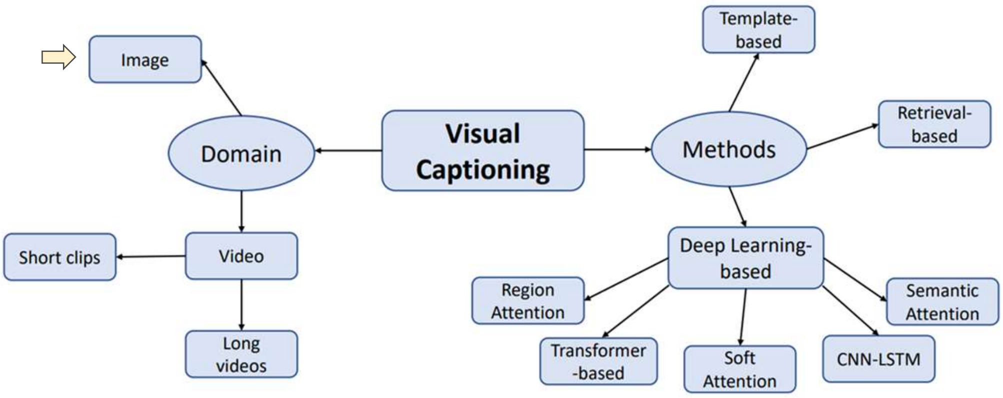

Visual Captioning
A horse carrying a large load of hay and two people sitting on it.
train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background photo taken during the day. red train car.
• Popular Topics: Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
• Popular Tasks: Image/video captioning, Dense captioning, Storytelling
Visual QA/Grounding/Reasoning
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
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
windowpane
sconce
bed
lamp
pillow
Popular Tasks:
• Text-to-image
• Layout-to-image
• Scene-graph-to-image
• Text-based image editing
• Story visualization
SOTA Models:
• StackGAN
• AttnGAN
• ObjGAN
• ...
Credit: VL-CVPR Tutorial. https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research
DEPARTMENT OF COMPUTER SCIENCE

Visual Captioning

Visual Captioning – Describe the content of an image or video with a natural language sentence

A cat is sitting next to a pine tree, looking up.

Visual Captioning Taxonomy



Credit: Luowei Zhou

Image Captioning

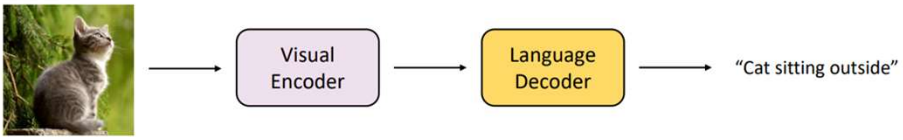
- Motivations
  - Real-world Usability
    - Help visually impaired people, learning-impaired
  - Improving Image Understanding
    - Classification, Objection detection
  - Image Retrieval

1. a young girl inhales with the intent of blowing out a candle
2. girl blowing out the candle on an ice cream

1. A shot from behind home plate of children playing baseball
2. A group of children playing baseball in the rain
3. Group of baseball players playing on a wet field

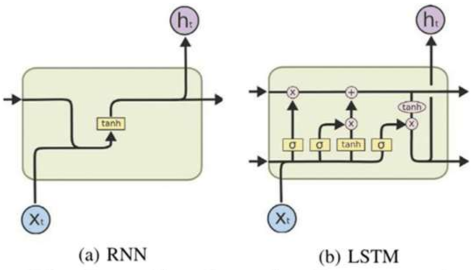
Image Captioning with CNN-LSTM

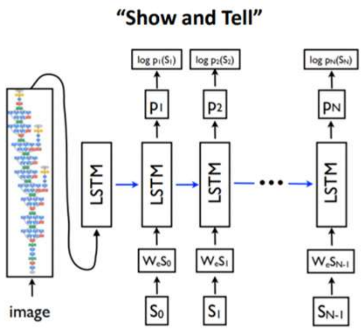
- The Encoder-Decoder framework



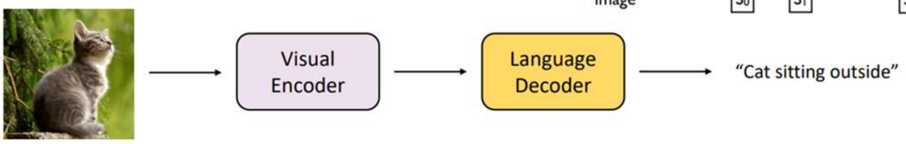
Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015

Image Captioning with CNN-LSTM





- The Encoder-Decoder framework

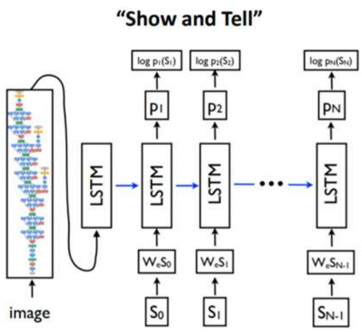


Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015

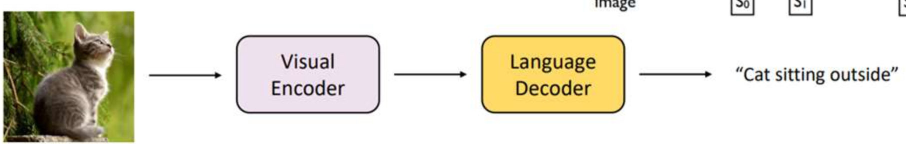
Image Captioning with CNN-LSTM

- Teacher forcing training

p_1 Cat sitting outside [END]
S_0 [Begin] cat sitting outside



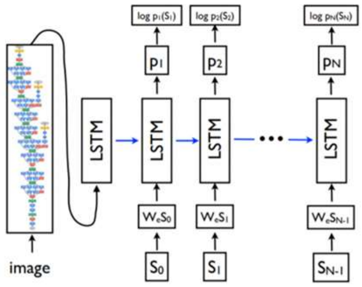
- The Encoder-Decoder framework



Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015

Image Captioning with CNN-LSTM

"Show and Tell"



- Problem Formulation

$$\theta^{\star} = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

- The Encoder-Decoder framework

→ Visual Encoder → Language Decoder → "Cat sitting outside"

Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

Image Captioning with Soft Attention

- Soft (self) Attention – Dynamically attend to input content based on query

→ Visual Encoder

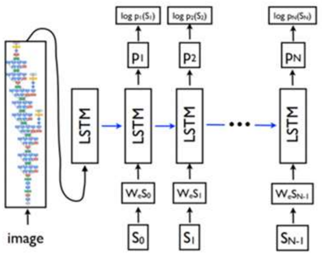
Entire image

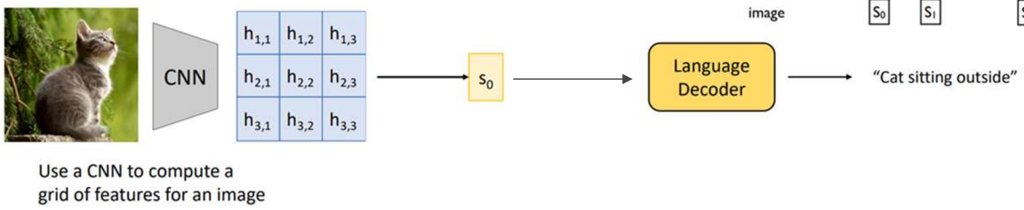
=> Informative *parts*

A stop sign is on a road with a mountain in the background.

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

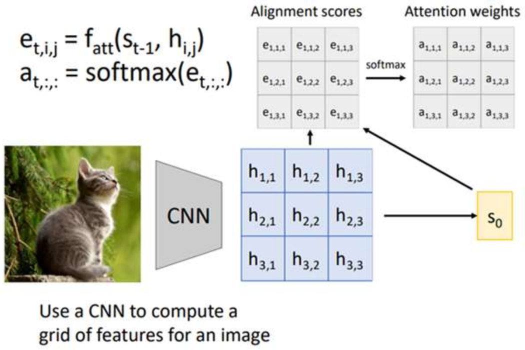
Review: Previous Image Captioning





Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

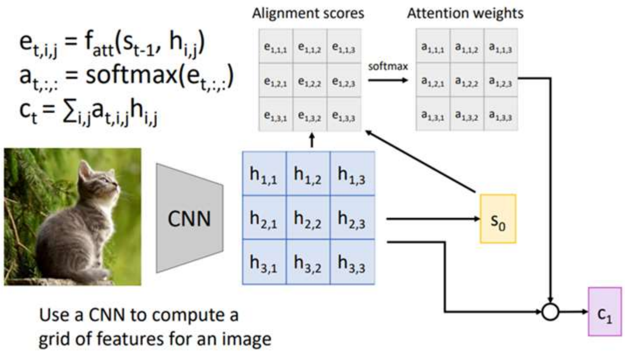
Image Captioning with Soft Attention



Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$



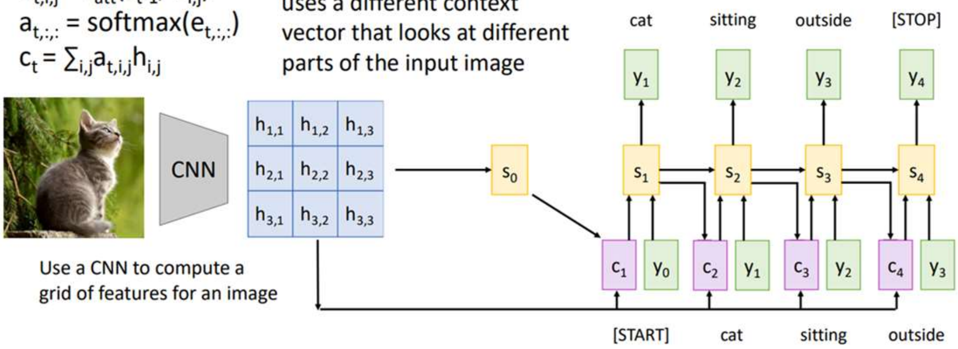
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

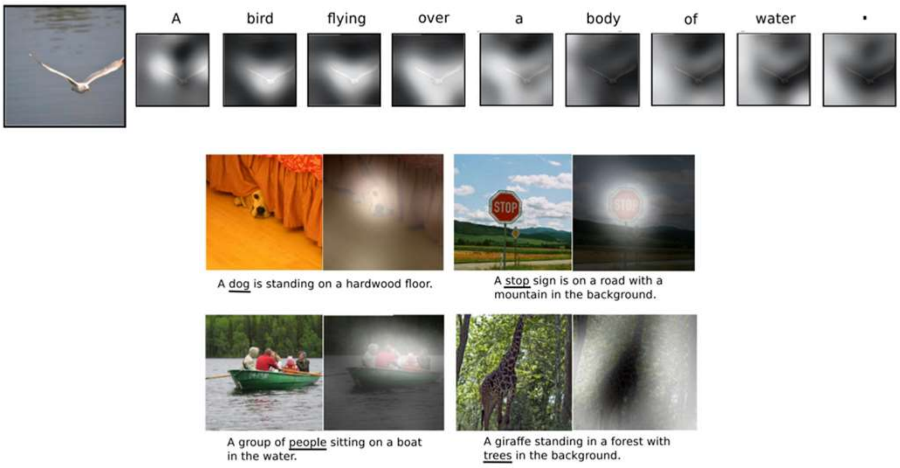
Each timestep of decoder uses a different context vector that looks at different parts of the input image



Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention



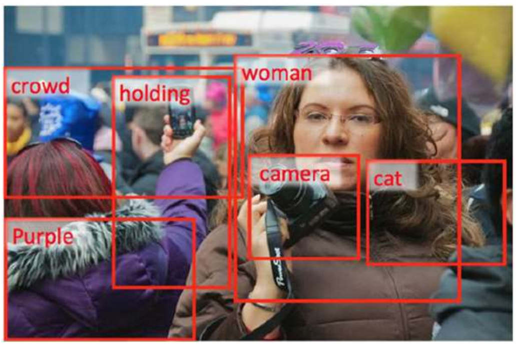
Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Semantic Attention

Additional textual information

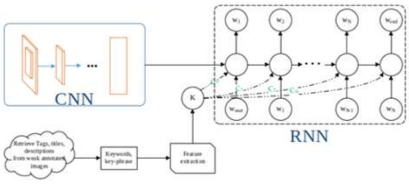
- Own noisy titles, tags or captions (Web)
- Visually similar nearest neighbor images

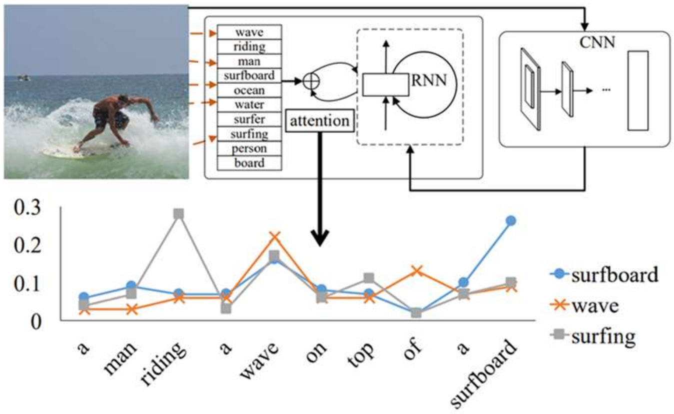
Exploit stronger vision models



*You, Jin, Wang, Fang, Luo. "Image captioning with semantic attention." CVPR 2016.

Image Captioning with Semantic Attention





HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Visual Attributes

k-NN

vase flowers bathroom table glass sink blue small white clear

Multi-label Ranking

sitting table small many little glass different flowers vase shown

FCN

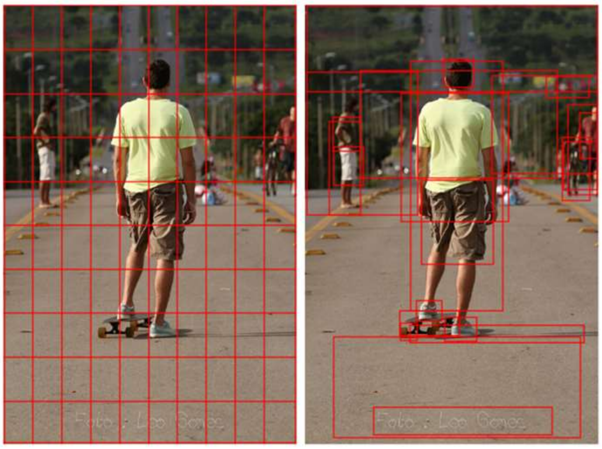
vase flowers table glass sitting kitchen water room white filled

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Google NIC	a white plate topped with a variety of food.	a baby is eating a piece of paper.	a close up of a plate of food on a table.	a teddy bear sitting on top of a chair .	a person is holding colorful umbrella.	a woman is holding a cell phone in her hand .
Top-5 visual attributes	plate broccoli fries food french	teeth brushing toothbrush holding baby	cake table plate sitting birthday	teddy cat bear stuffed white	umbrella beach water sitting boat	woman bathroom her scissors man
ATT-FCN	a plate with a sandwich and french fries.	a baby with a toothbrush in its mouth.	a table topped with a cake with candles on it.	a white teddy bear sitting next to a stuffed animal .	a black umbrella sitting on top of a sandy beach .	a woman holding a pair of scissors in her hands .

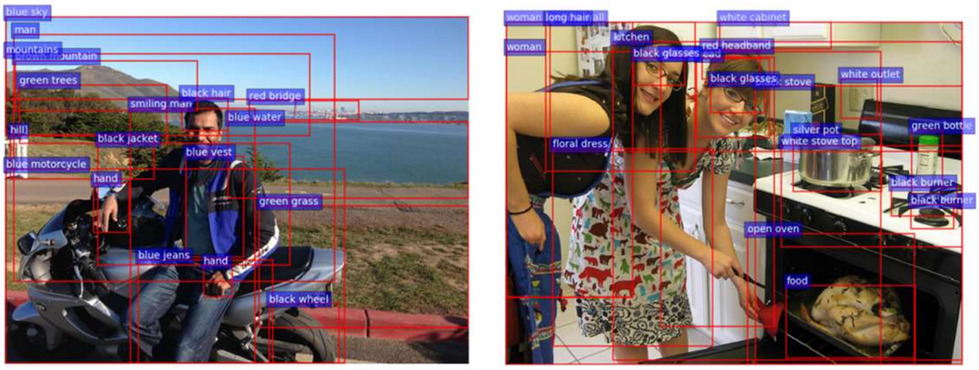
Image Captioning with “Fancier” Attention

- **Region based attention**



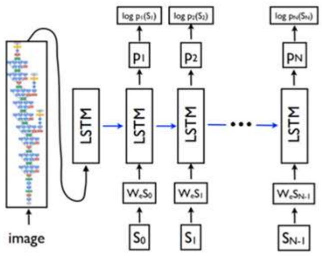
Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering.", CVPR 2018

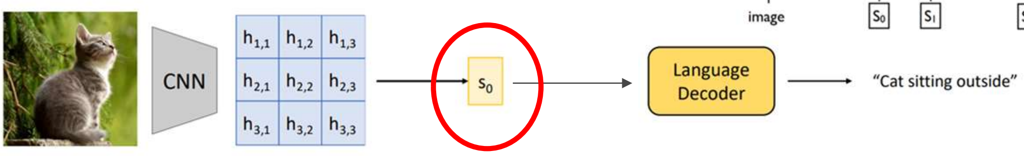
Image Captioning with “Fancier” Attention



Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering.", CVPR 2018

Review: Previous Image Captioning



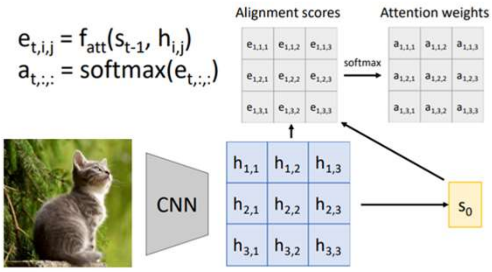


Use a CNN to compute a
grid of features for an image

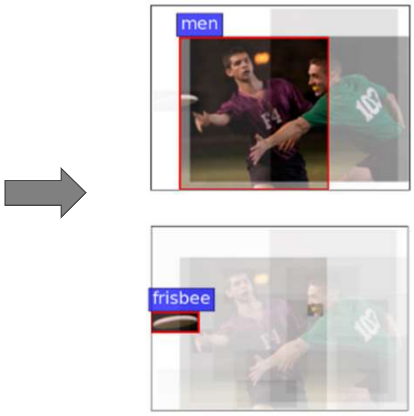
Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with “Fancier” Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$

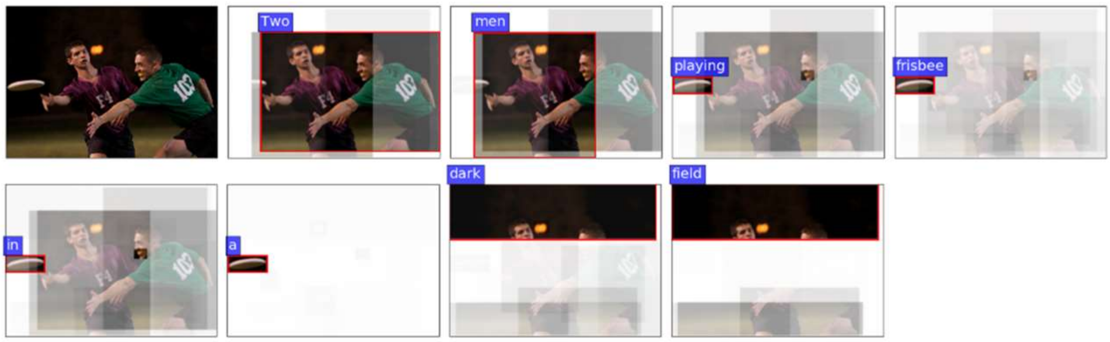


Use a CNN to compute a grid of features for an image



- Attention (weighted sum) over:
  N*N patches -> K object regions

Image Captioning with “Fancier” Attention

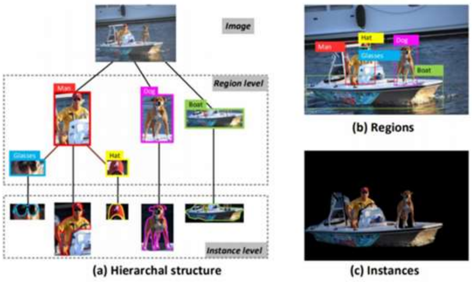


Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering.", CVPR 2018

Image Captioning with “Fancier” Attention

Hierarchy Parsing and GCNs

- Hierarchal tree structure in image



Auto-Encoding Scene Graphs

- Scene Graphs in image and text

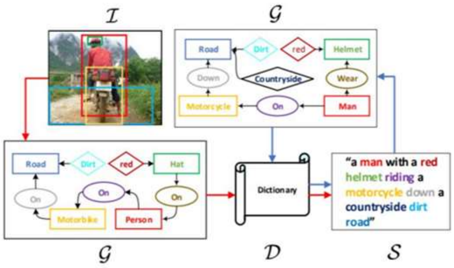
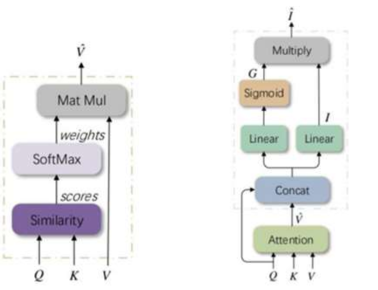


Image Captioning with “Fancier” Attention

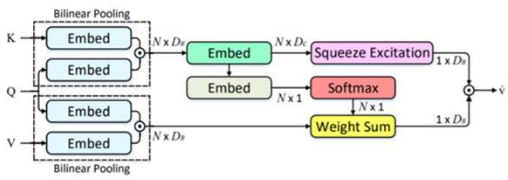
Attention on Attention



(a) Attention (b) Attention on Attention

X-Linear Attention

- Spatial and channel-wise bilinear attention



Evaluation – Benchmark Datasets

COCO Captions

- Train / val / test: 113k / 5k / 5k
- Hidden test (leaderboard): 40k
- Vocabulary (≥ 5 occurrences): 9,587

Flirckr30K

- Train / val / test: 29k / 1k / 1k
- Vocabulary (≥ 5 occurrences): 6,864

Evaluation – Benchmark Datasets

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

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

Bunk bed with a narrow shelf sitting underneath it.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Evaluation – Metrics

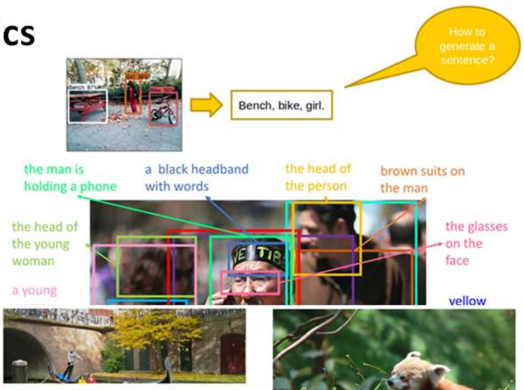
Most commonly-used: BLEU / METEOR / CIDEr / SPICE

- BLEU: based on n-gram based precision
- METEOR: ordering sensitive through unigram matching
- CIDEr: gives more weight-age to important n-grams through TF-IDF
- SPICE: F1-score over caption scene-graph tuples

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Image Captioning – Advanced Topics

- **Un-/weakly-supervised**
- Dense captioning
- Novel object captioning
- **Stylized captioning**
- **Captioning with reading comprehension**
- Grounded captioning



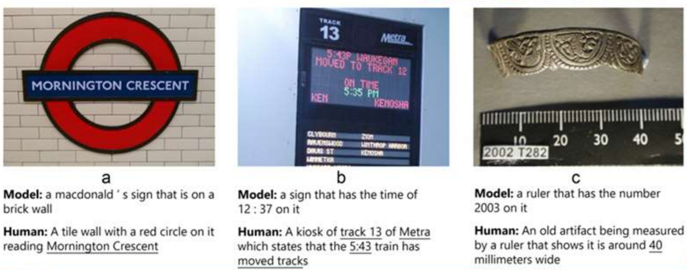


Image Captioning with Unpaired Data

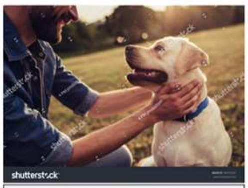
Bench, bike, girl.

How to generate a sentence?

Key: disentangle **sentence quality** and **image-text relevance**

* Yang Feng, Lin Ma, Wei Liu, Jiebo Luo. "Unsupervised image captioning." In CVPR 2019.

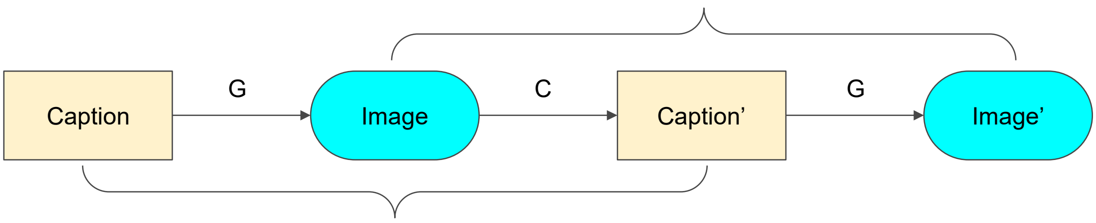
Image Captioning with Unpaired Data



Cropped image of handsome young man with labrador outdoors. Man on a green grass with dog. Cynologist

British Shorthair cat isolated on white. Smiling expression, happy

- Image-text relevance
  - Cycle consistency



HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Image Captioning with Unpaired Data

- Sentence quality

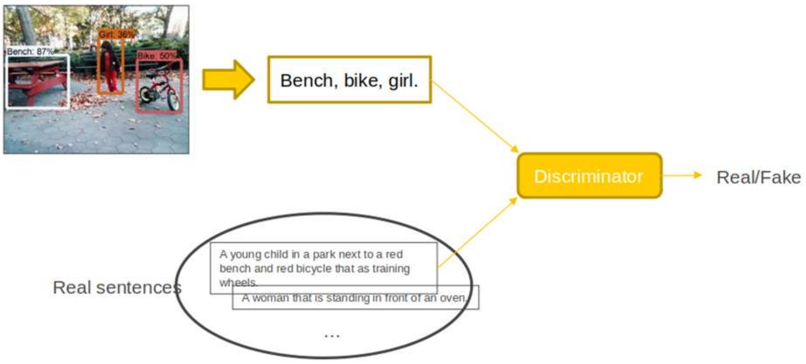
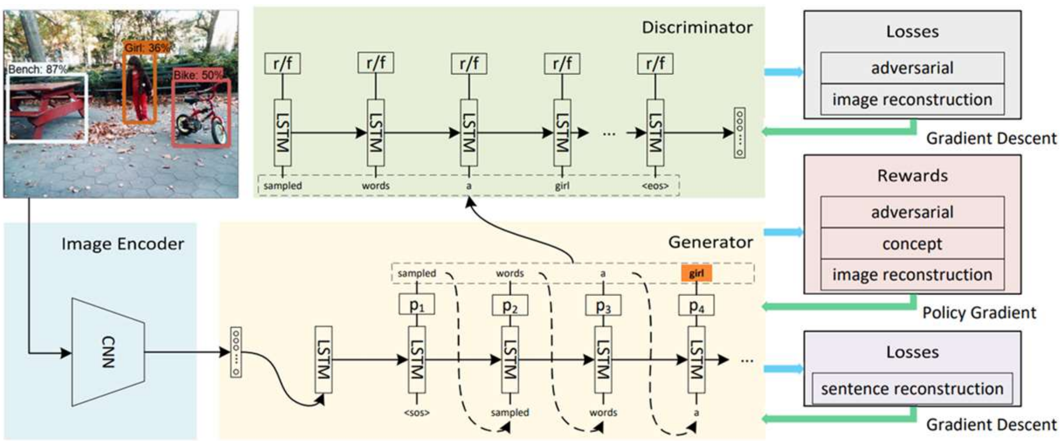
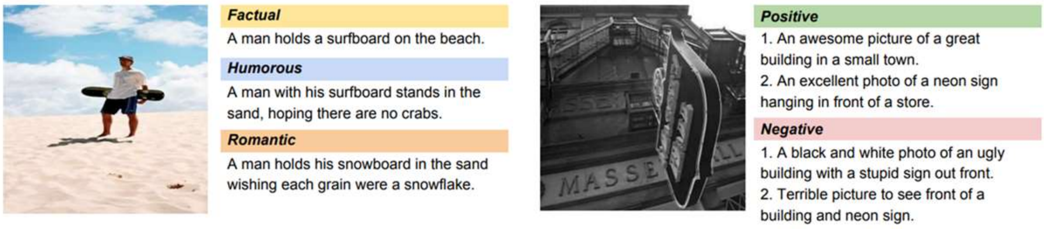


Image Captioning with Unpaired Data



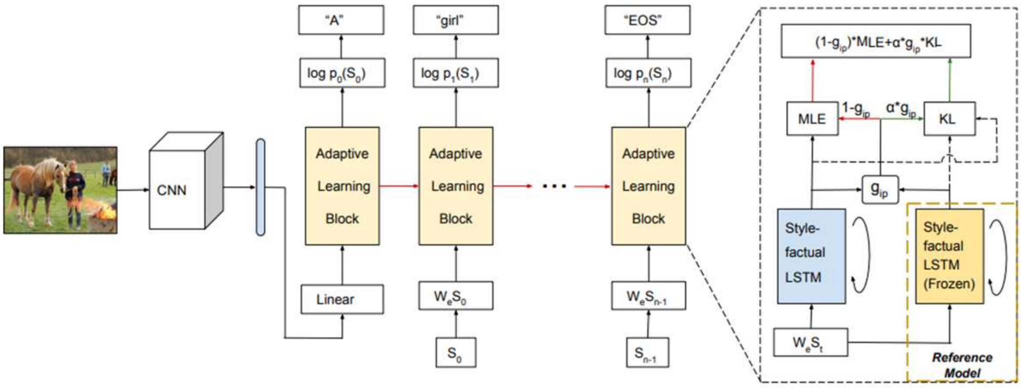
Stylized Captioning



Style word and factual word

* Chen, Zhang, You, Fang, Wang, Jin, Luo. "``Factual''or``Emotional'': Stylized Image Captioning with Adaptive Learning and Attention." In ECCV 2018.

Stylized Captioning



Captioning with Reading Comprehension

a

Model: a macdonald ' s sign that is on a brick wall

Human: A tile wall with a red circle on it reading Mornington Crescent

b

Model: a sign that has the time of 12 : 37 on it

Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks

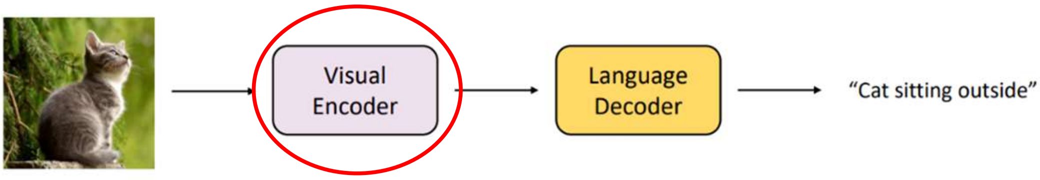
c

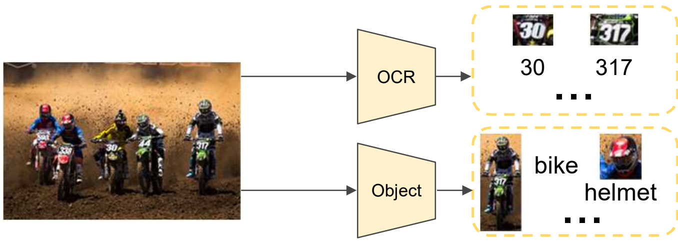
Model: a ruler that has the number 2003 on it

Human: An old artifact being measured by a ruler that shows it is around 40 millimeters wide

* Yang, Lu, Yin, Florencio, Wang, Zhang, Zhang, Luo. "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption." In CVPR 2021.

Captioning with Reading Comprehension

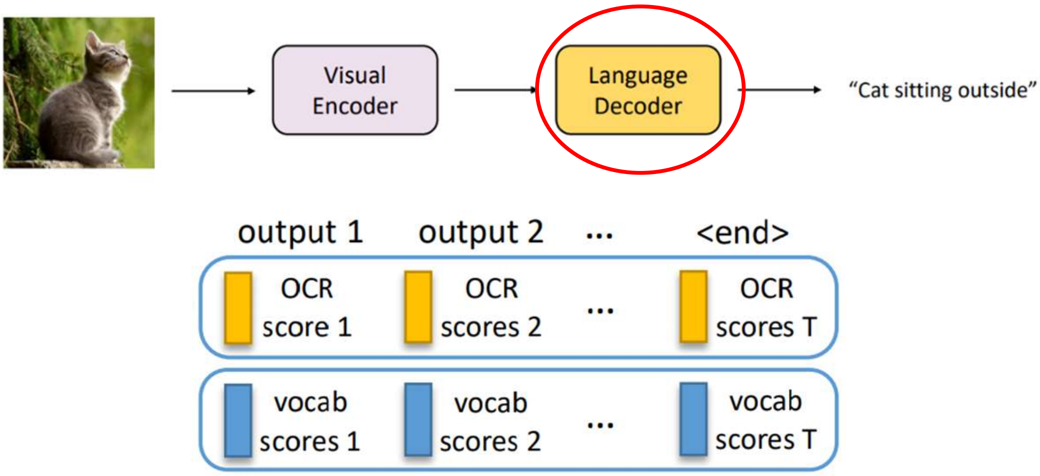




* Yang, Lu, Yin, Florencio, Wang, Zhang, Zhang, Luo. "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption." In CVPR 2021.

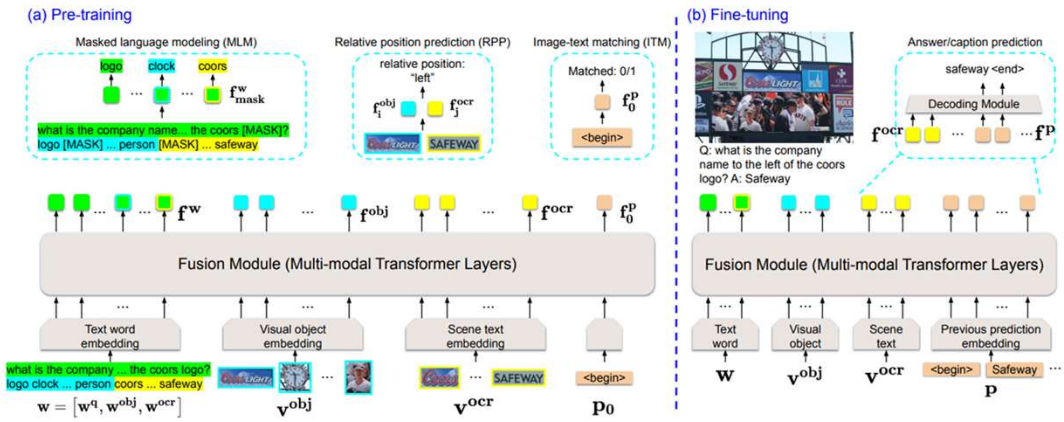
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Captioning with Reading Comprehension

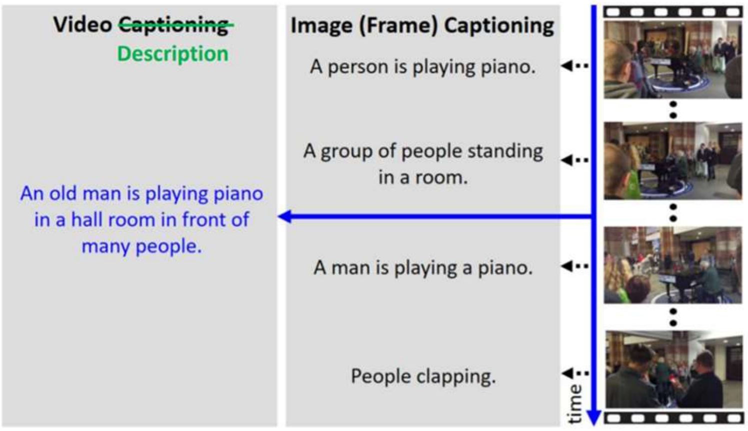


* Yang, Lu, Yin, Florencio, Wang, Zhang, Zhang, Luo. "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption." In CVPR 2021.

Captioning with Reading Comprehension

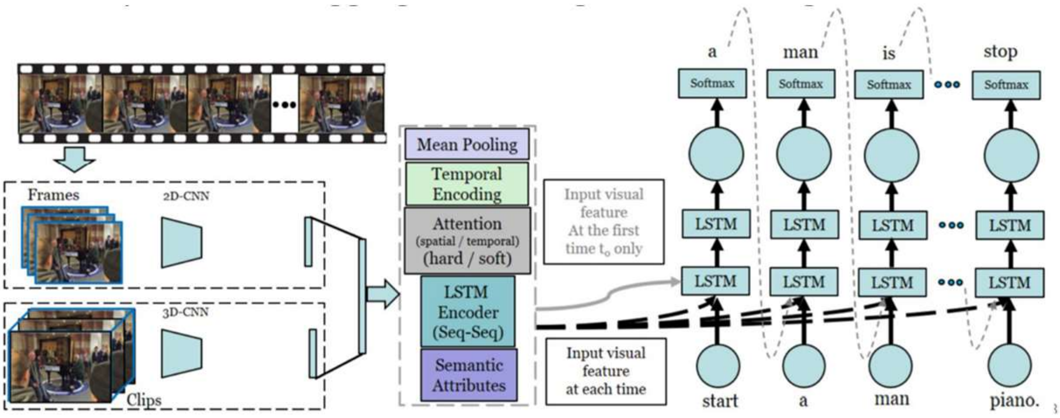


From Image to Video

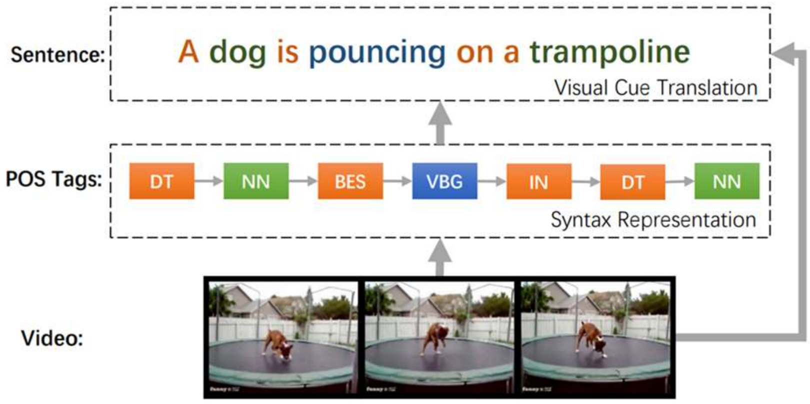


Video Captioning

- **Encoder-Decoder Network**



Assisted by POS Tags

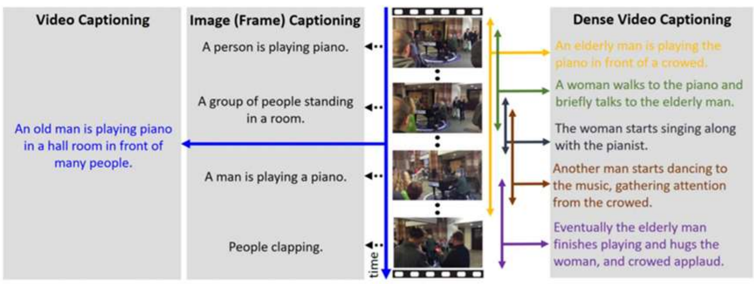


* Hou, Wu, Zhao, Luo. "Joint syntax representation learning and visual cue translation for video captioning." In ICCV 2019.

Video Captioning

- Dense video captioning
- Video paragraph description
- Other special domains

Add chopped bacon to a hot pan and stir. Remove the bacon from the pan. Place the beef into a hot pan to brown. Add onion and carrots to the pan. Pour the meat back into the pan and add flour. Place the pan into the oven. Add bay leaves thyme red wine beef stock garlic and tomato paste to the pan and boil. Add pearl onions to a hot pan and add beef stock bay leaf and thyme. Add mushrooms to a hot pan. Add the mushrooms and pearl onions to the meat...



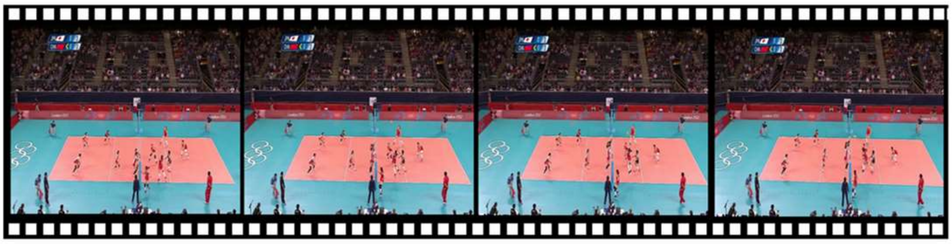
GIF, Sports, etc.

Animated GIF Captioning (TGIF dataset)

* Li, Song, Cao, Tetreault, Goldberg, Luo. "TGIF: A new dataset and benchmark on animated GIF description." In CVPR 2016.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

Sports Video Captioning



Conventional Captioning:
Two teams of players are playing a volleyball match in the gym.

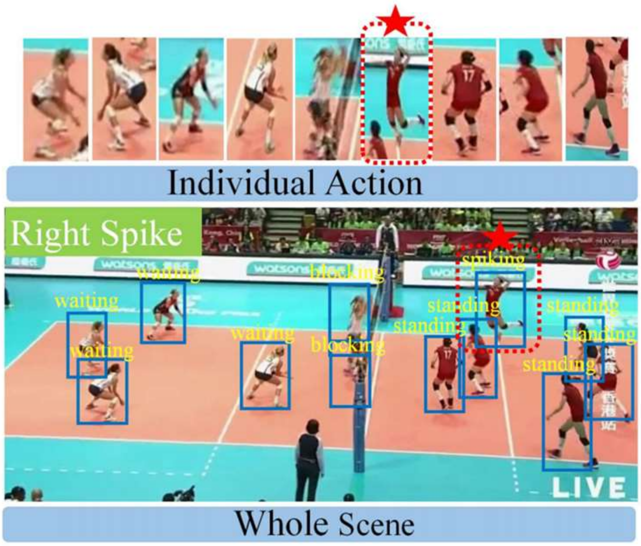
Our Captioning:
Now the team on the left side is defending, while the team on the right side is attacking. On the left team, a player is jumping and blocking. A player is digging, a player is waiting, and other teammates are standing. On the right team, a player is passing the ball to her teammate. A player is jumping and spiking, while other teammates are standing.

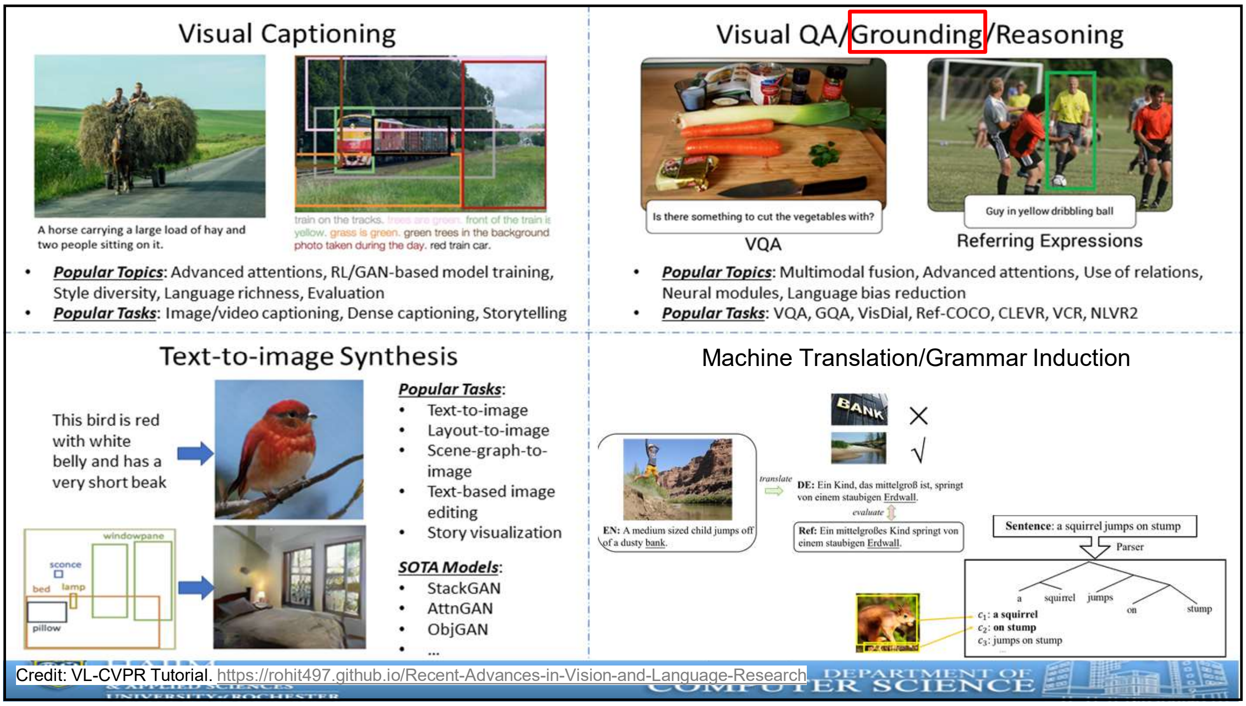
* Qi, Qin, Li, Wang, Luo. "Sports video captioning by attentive motion representation based hierarchical recurrent neural networks." ACMMW 2018.
* Qi, Qin, Li, Wang, Luo, Van Gool. "stagNet: an attentive semantic RNN for group activity and individual action recognition." IEEE TCSVT.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

Sports Video Captioning

Sports Video Captioning




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
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
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
windowpane
sconce
bed
lamp
pillow
Popular Tasks:
• Text-to-image
• Layout-to-image
• Scene-graph-to-image
• Text-based image editing
• Story visualization
SOTA Models:
• StackGAN
• AttnGAN
• ObjGAN
• ...
Machine Translation/Grammar Induction
BANK
EN: A medium sized child jumps off of a dusty bank.
translate
DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.
evaluate
Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.
Sentence: a squirrel jumps on stump
Parser
a squirrel jumps on stump
c1: a squirrel
c2: on stump
c3: jumps on stump
Credit: VL-CVPR Tutorial. https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research
DEPARTMENT OF COMPUTER SCIENCE

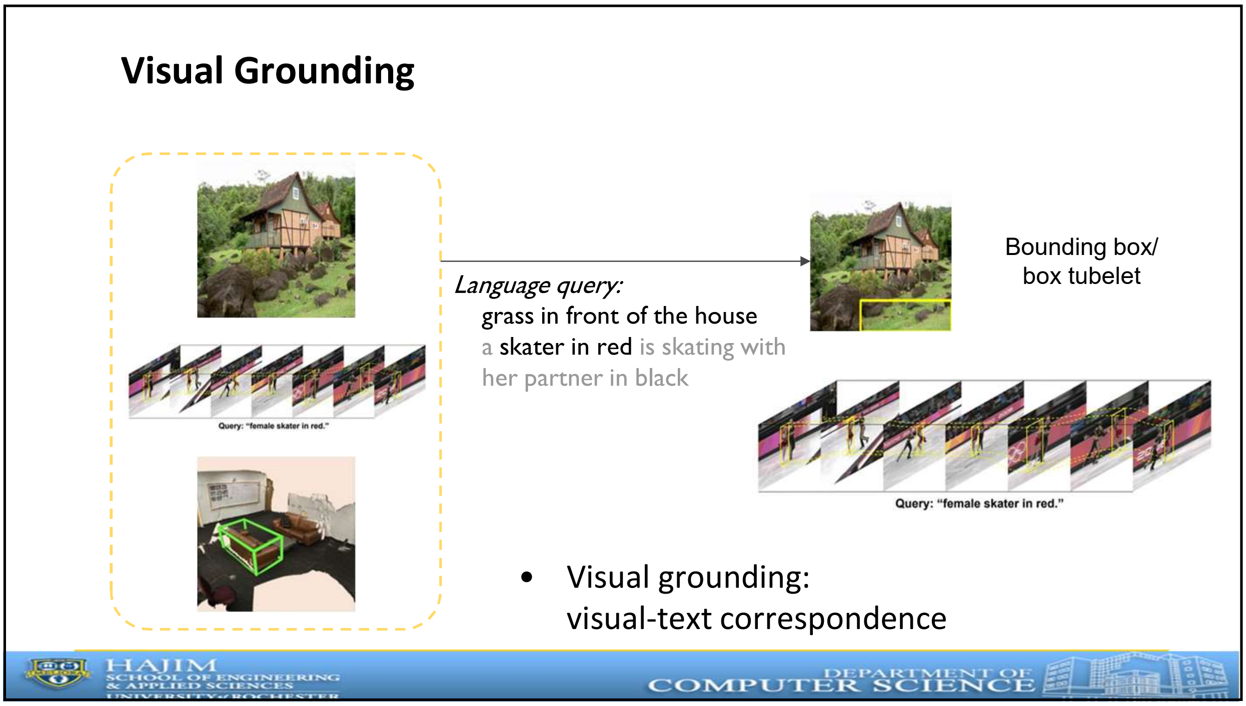
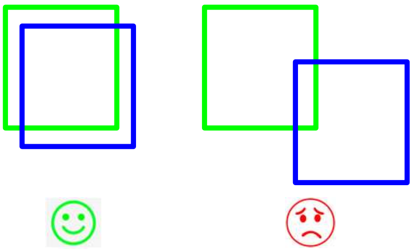

Visual Grounding
Query: "female skater in red."
Language query:
grass in front of the house
a skater in red is skating with her partner in black
Bounding box/
box tubelet
Query: "female skater in red."
• Visual grounding:
visual-text correspondence
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Image Grounding

Language query:
grass in front of the house

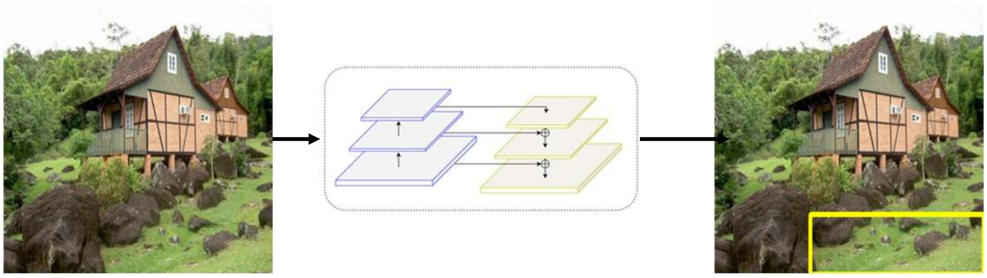
Bounding box



- Image visual grounding

Image Visual Grounding

- Language query => region of the image

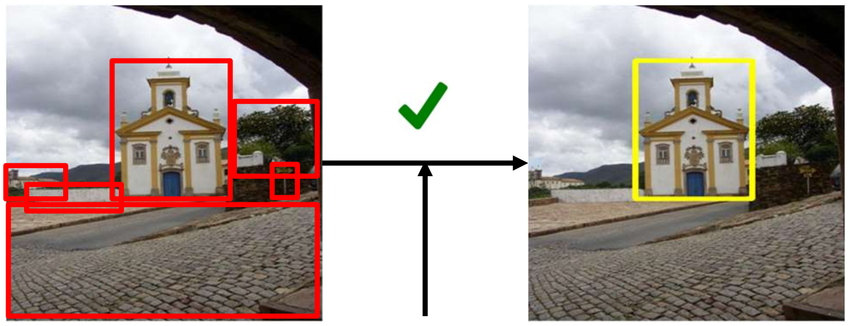


Query: grass in front of the house

* Yang, Gong, Wang, Huang, Yu, Luo. "A fast and accurate one-stage approach to visual grounding." In ICCV 2019. (oral)

Existing Framework

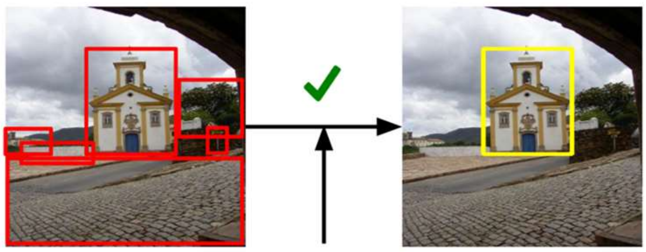
- Two-stage framework



Query: center building

Existing Framework

- Two-stage framework
  – Region proposal
  – Similarity ranking

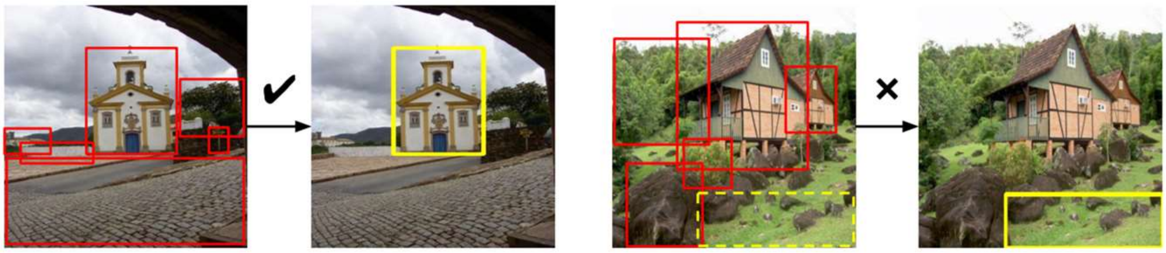


Query: center building

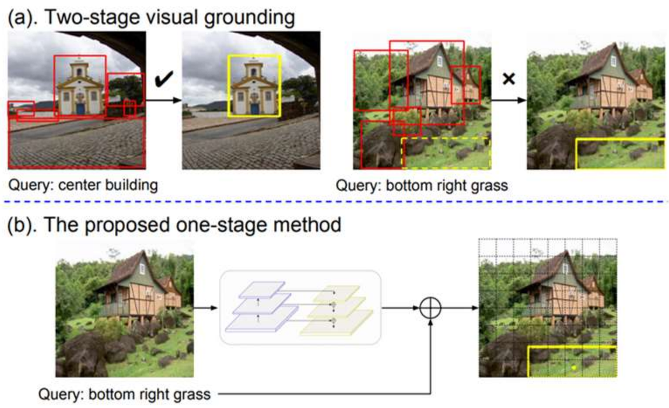
Existing Framework

- **Propose-and-Rank**

- Limited candidates
- Slow in speed



One-stage Visual Grounding

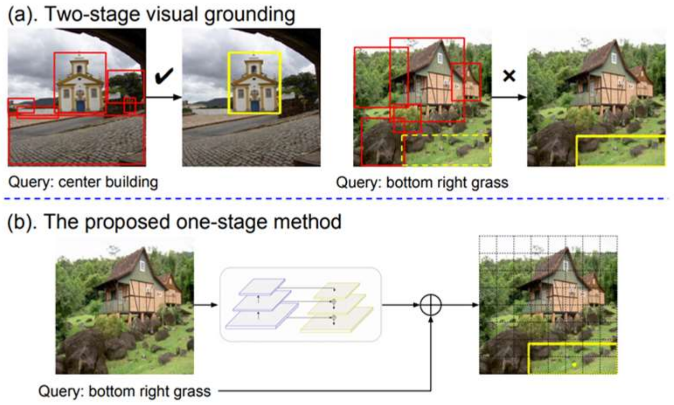


- Propose-and-rank => Grounding-by-detection

* Yang, Gong, Wang, Huang, Yu, Luo. "A fast and accurate one-stage approach to visual grounding." In ICCV 2019. (oral)

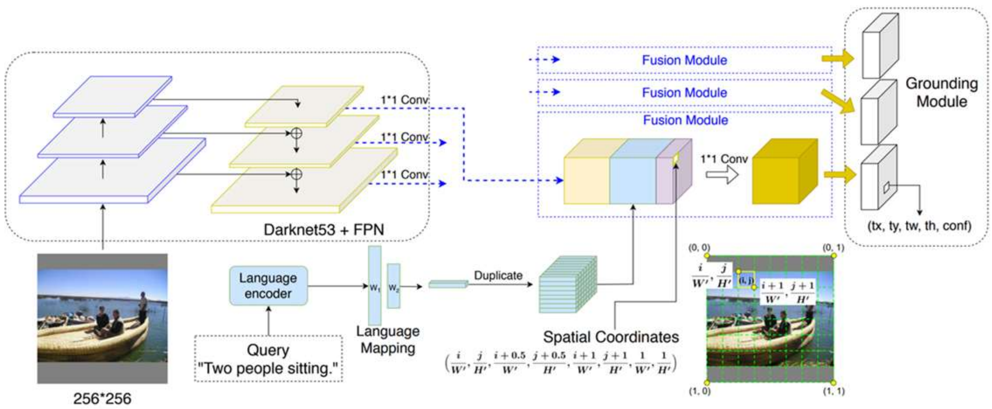
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

One-stage Visual Grounding



- Accurate: +7-20% absolute
- Fast: 10x

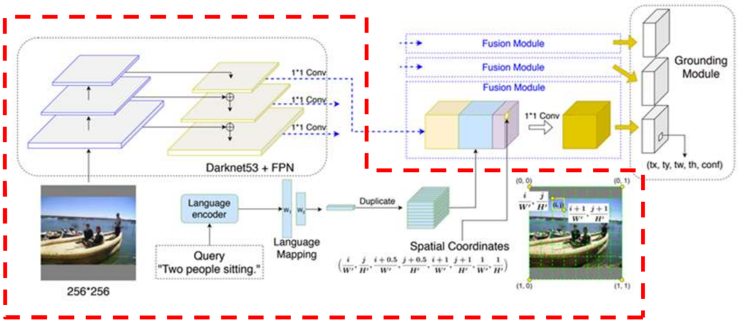
Architecture



- Encoder module
- Fusion module
- Grounding module

Architecture

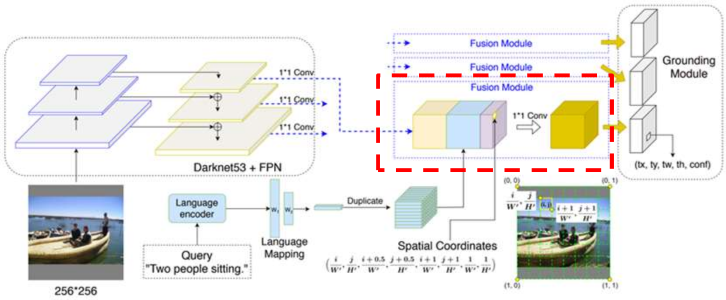
- Encoder
- Fusion module
- Grounding module



- Visual encoder
- Language encoder
- Spatial encoder

Architecture

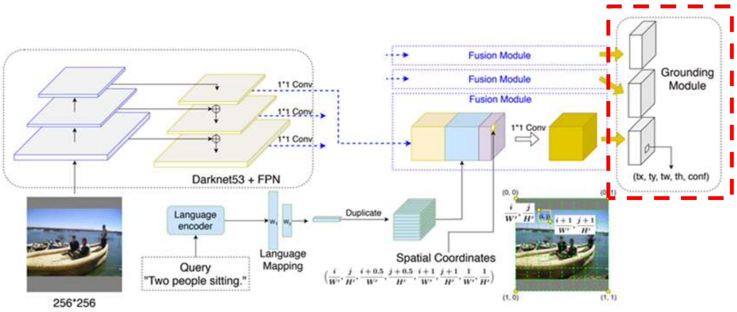
- Encoder
- Fusion module
- Grounding module



- Image-level fusion

Architecture

- Encoder
- Fusion module
- Grounding module



- Output format: box + confidence

Datasets

- For phrase localization: Flickr 30K Entities
- For referring expression comprehension: ReferItGame
- Acc@0.5 IoU

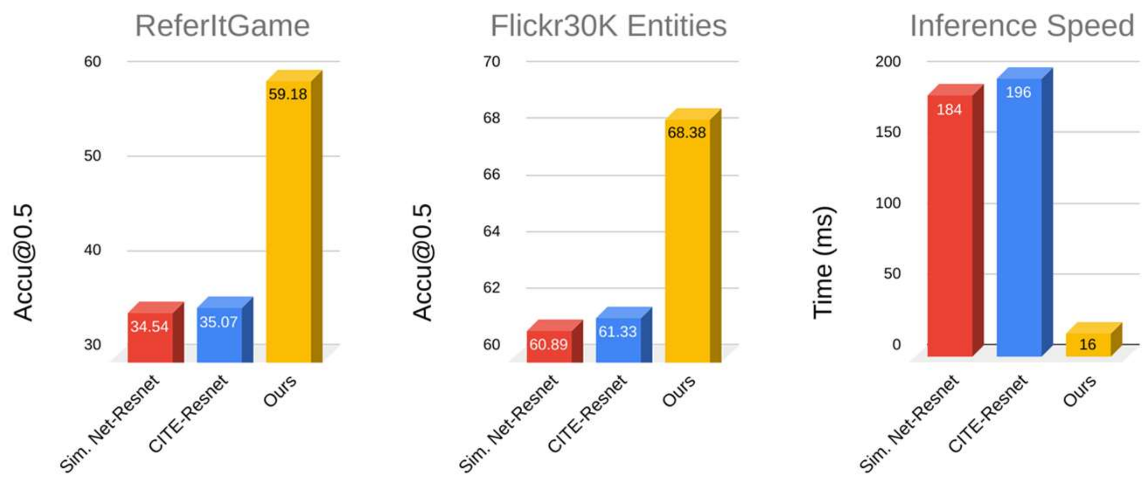
An older man in a white jacket works at a stand featuring a wide variety of colorful food.

Phrase localization

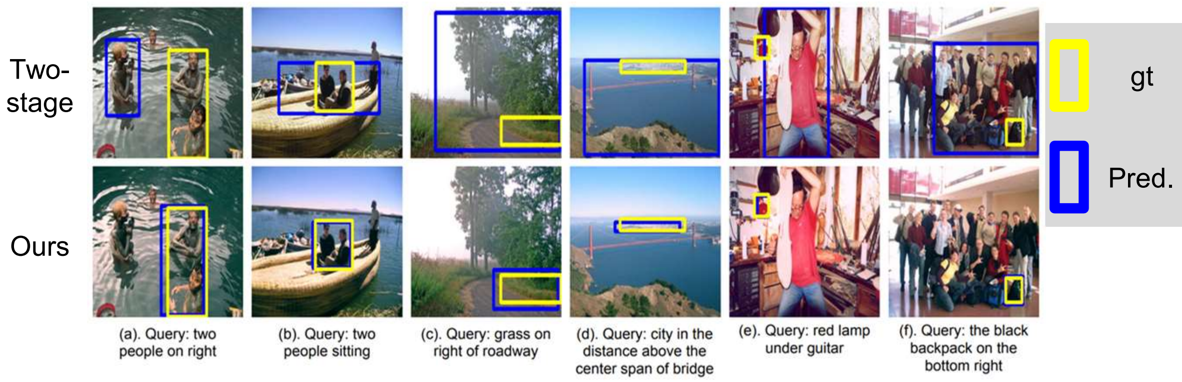
the black backpack on the bottom right

Referring expression comprehension

Comparison to other methods

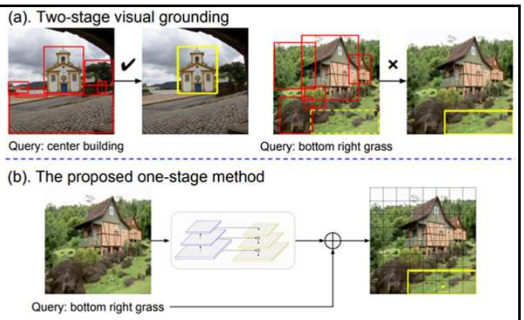


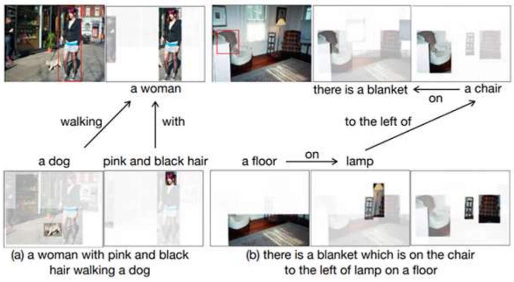
Qualitative Results

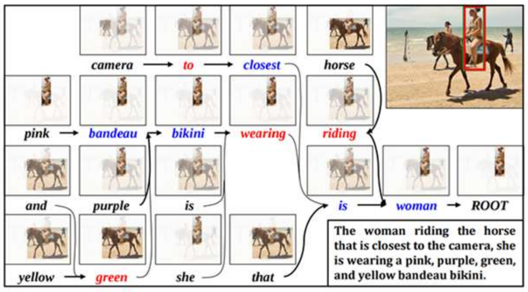


Understanding *Complex* Queries

- External VL graph/ tree
- Recursive modeling

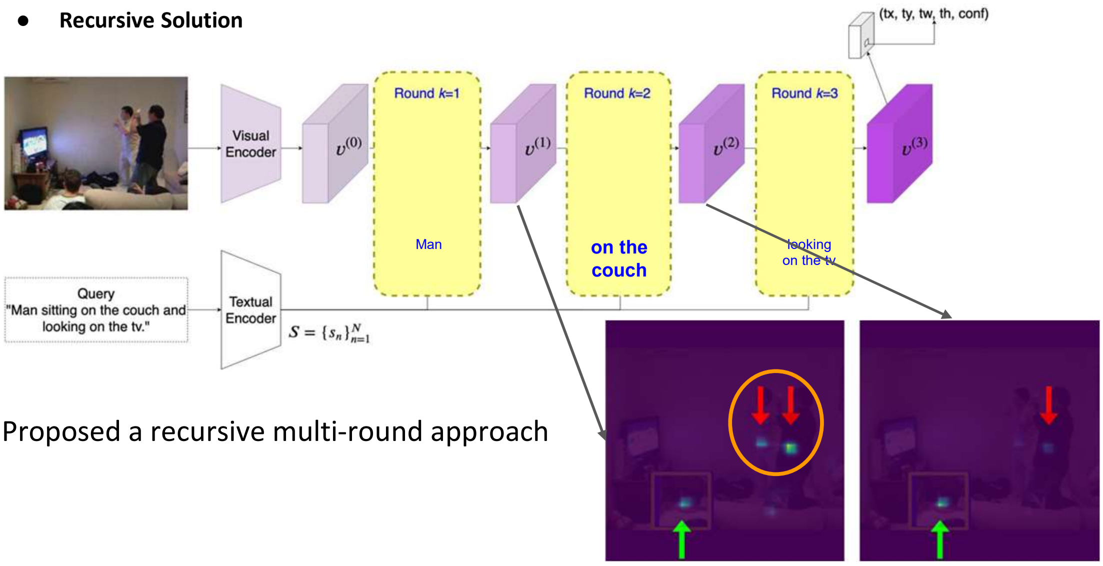






Understanding Complex Queries

- **Recursive Solution**

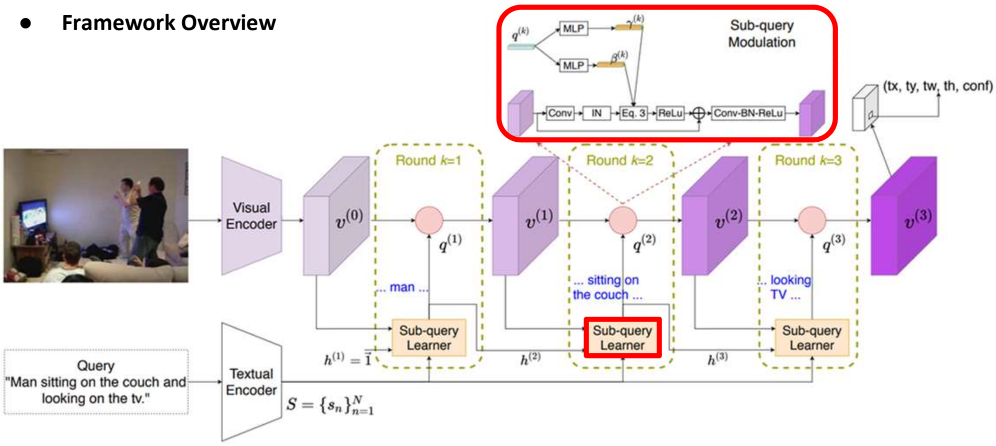


- Proposed a recursive multi-round approach

* Yang, Chen, Wang, Luo. "Improving one-stage visual grounding by recursive sub-query construction." In ECCV 2020

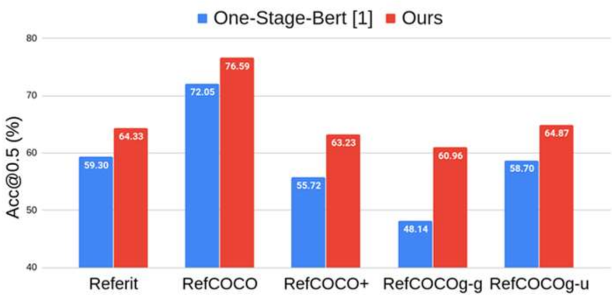
Method

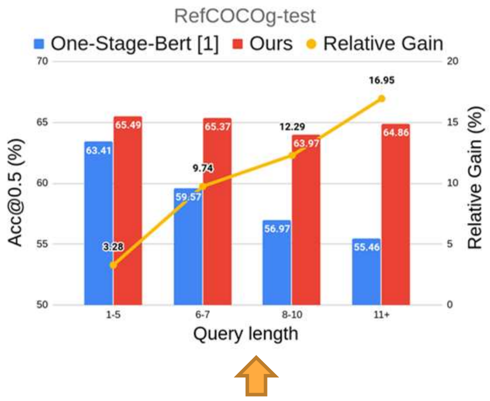
- **Framework Overview**



- Sub-query learner (to construct the sub-queries)
- Sub-query modulation (to refine the fused feature with sub-queries)

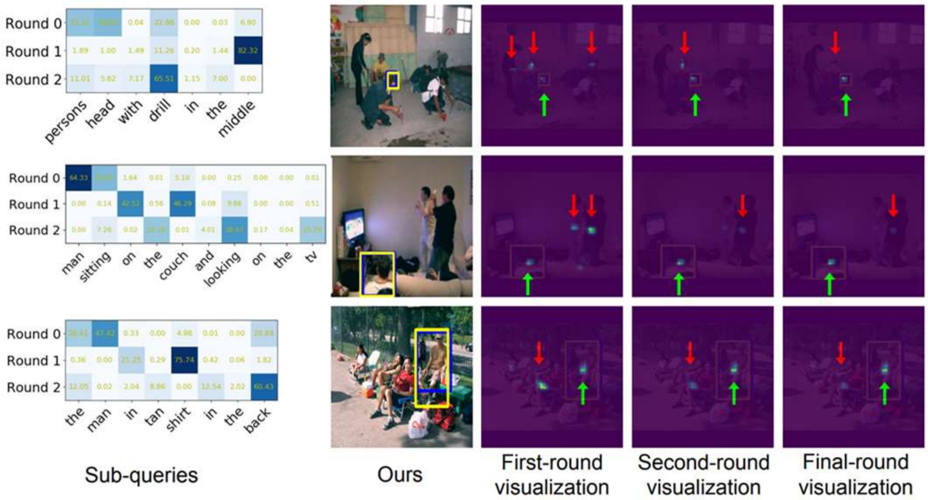
Experiments





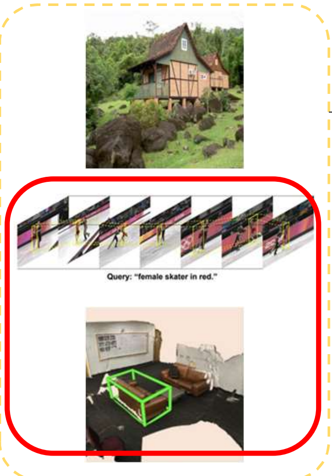
Experiments

- **Recursive Disambiguation**



- Recursive dis-ambiguous procedures

Visual Grounding beyond Images

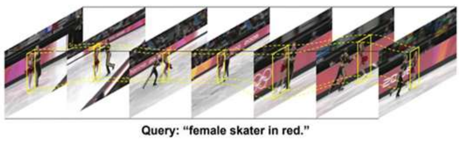


Language query:

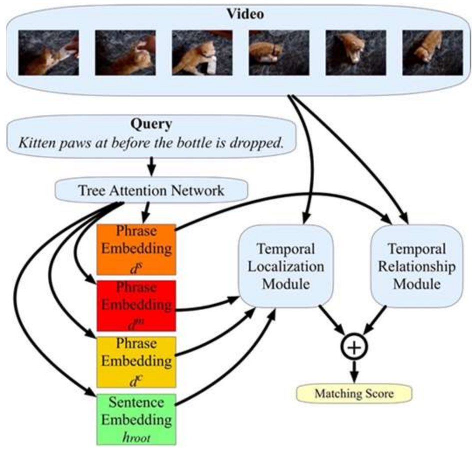
grass in front of the house

a skater in red is skating with her partner in black

Bounding box/
box tubelet

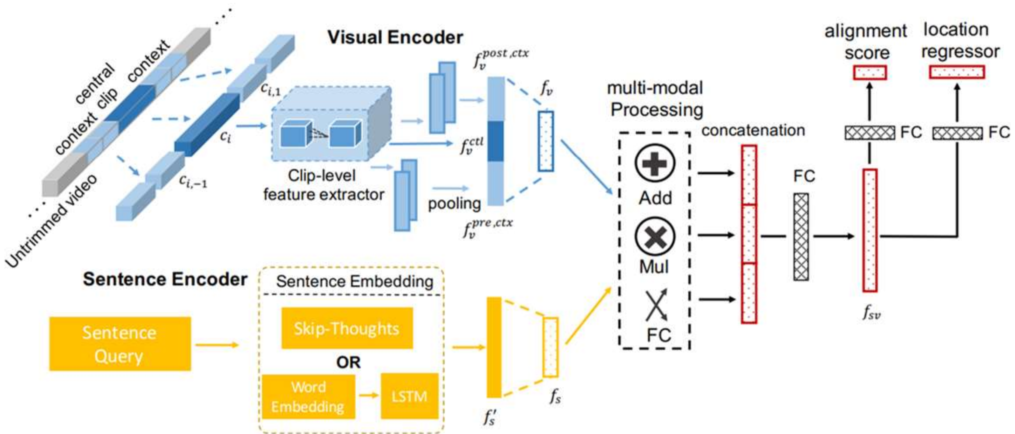


Video Temporal Grounding

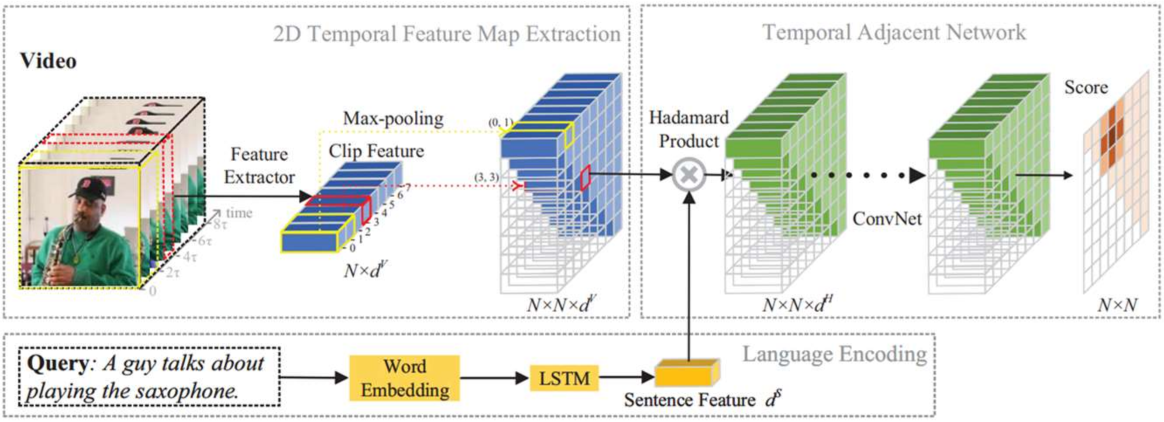


[1] Gao, Jiyang, et al. "Tall: Temporal activity localization via language query." In ICCV 2017.
[2] *Songyang Zhang, Jinsong Su, Jiebo Luo. "Exploiting Temporal Relationships in Video Moment Localization with Natural Language." MM 2019.

Video Temporal Grounding



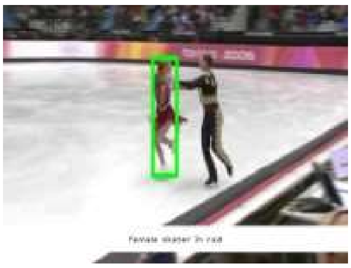
Video Temporal Grounding (2D TAN)

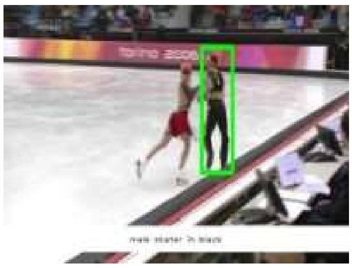


* Zhang, Peng, Fu, Luo. "Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language." In AAAI 2020.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Video Object Grounding



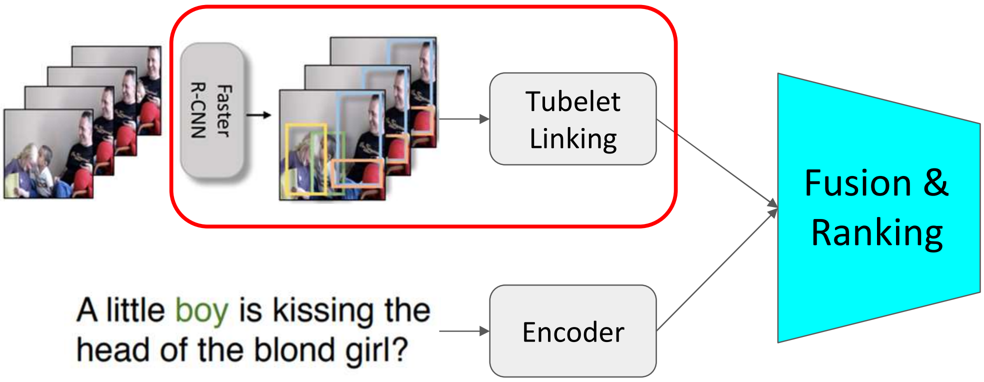


- From image to video

* Real-time (~40 fps) with a single NVIDIA 1080TI GPU
** Results of Yang, Kumar, Chen, Su, Luo. "Grounding-Tracking-Integration." *In IEEE T-CSVT*.

Video Object Grounding

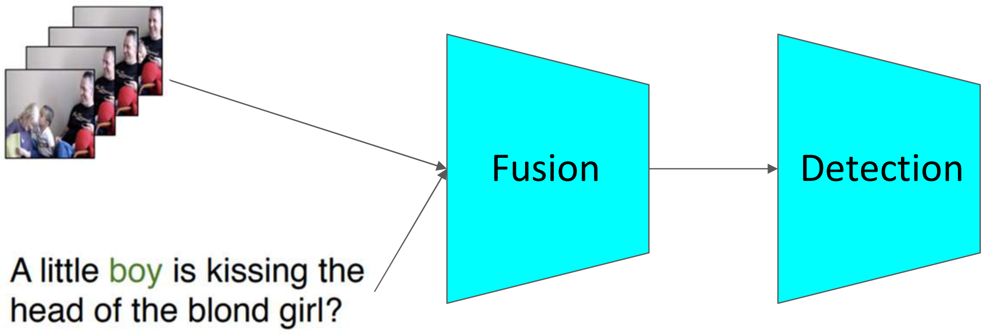
- **Two-stage Approach**



Zhang, Zhu, et al. "Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences." In CVPR 2020.

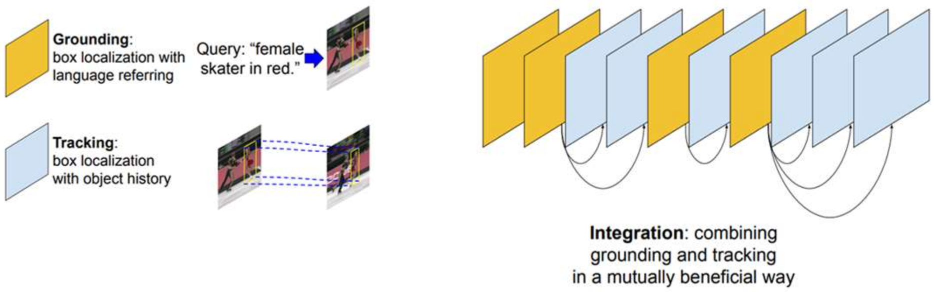
Video Object Grounding

- **One-stage Approach**



Video Object Grounding

- **One-stage Approach**



- Self-evaluate
- First real-time video object grounding framework

3D Visual Grounding

"When facing the wall of bookshelves, choose the box to the left."

"Choose the bookcase on the far left."

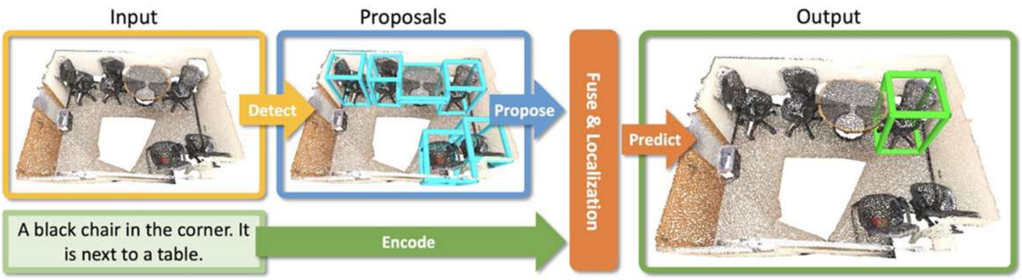
"the long bookshelf near a window"

"The tall bookshelf furthest from the door."

"When facing the green chalk board this bookshelf is the one on the right."

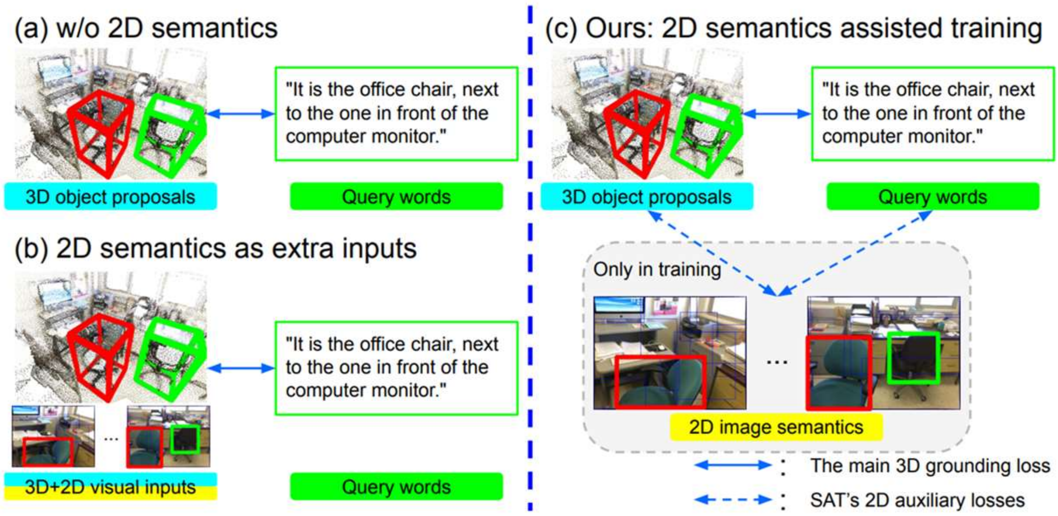
[1] Panos, Achlioptas, et al. "ReferIt3D: Neural Listeners for Fine-Grained Object Identification in Real-World 3D Scenes." In ECCV 2020 Oral.
[2] Chen, Dave, et al. "ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language." In ECCV 2020.

Challenges

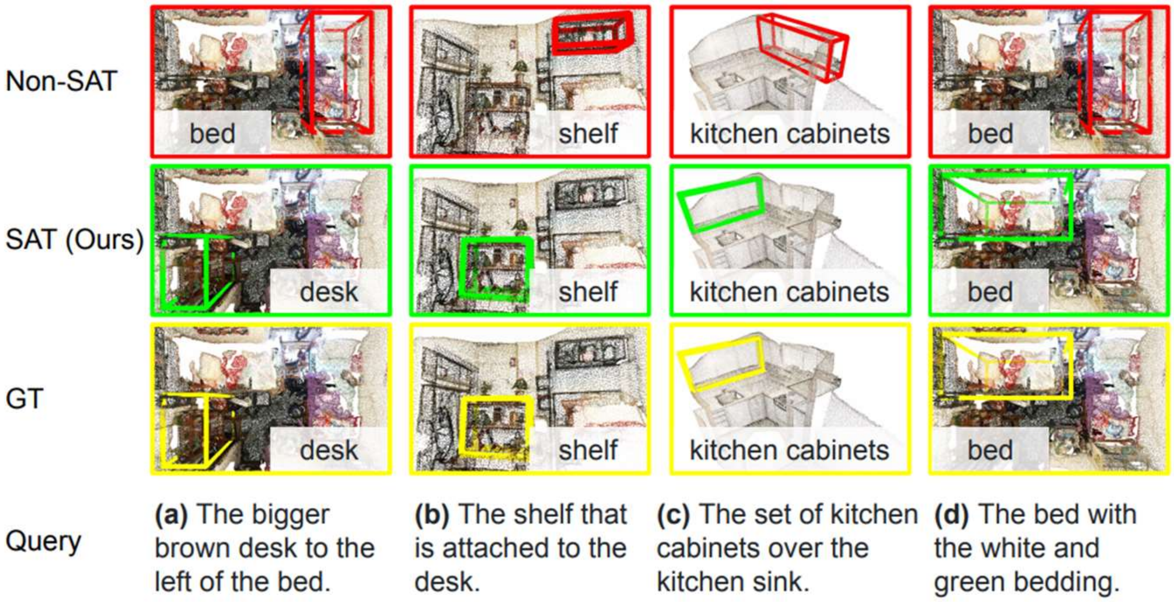


- Propose and rank
- Ranking is the key: point-cloud and language joint representation learning

[1] Panos, Achlioptas, et al. "ReferIt3D: Neural Listeners for Fine-Grained Object Identification in Real-World 3D Scenes." In ECCV 2020 Oral.
[2] Chen, Dave, et al. "ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language." In ECCV 2020.



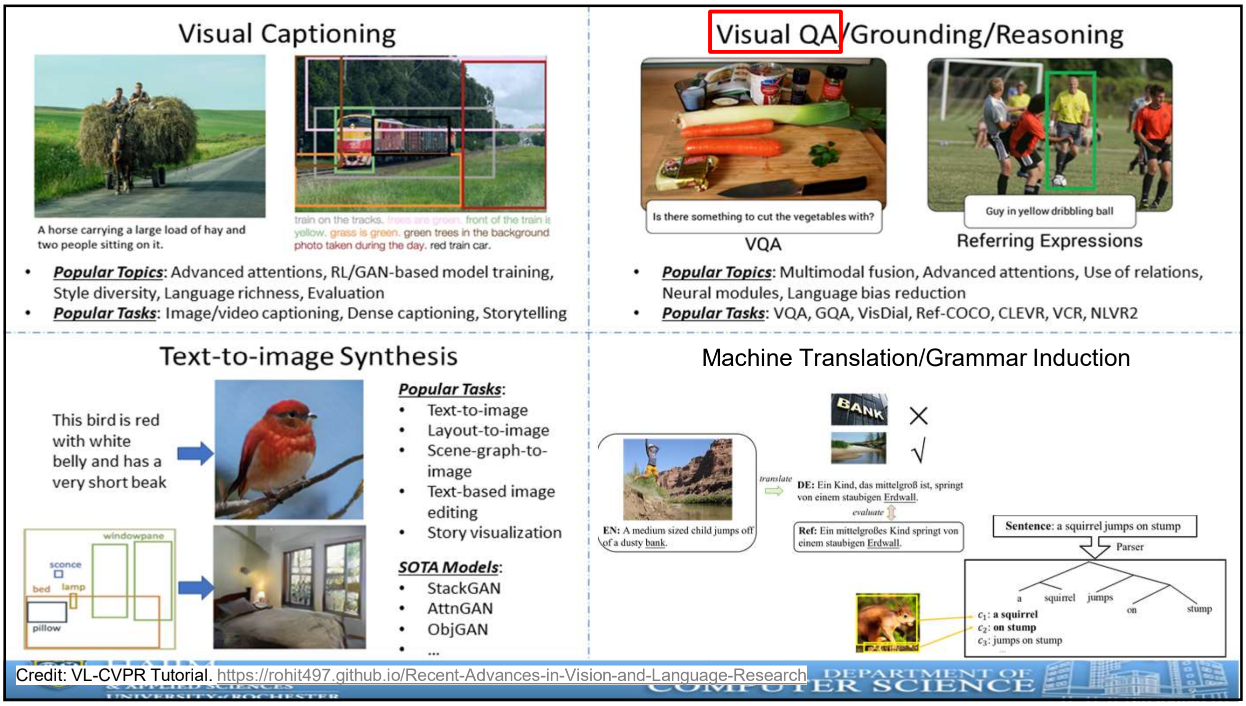
* Zhengyuan Yang, Songyang Zhang, Liwei Wang, Jiebo Luo. "SAT: 2D Semantics Assisted Training for 3D Visual Grounding." arxiv.

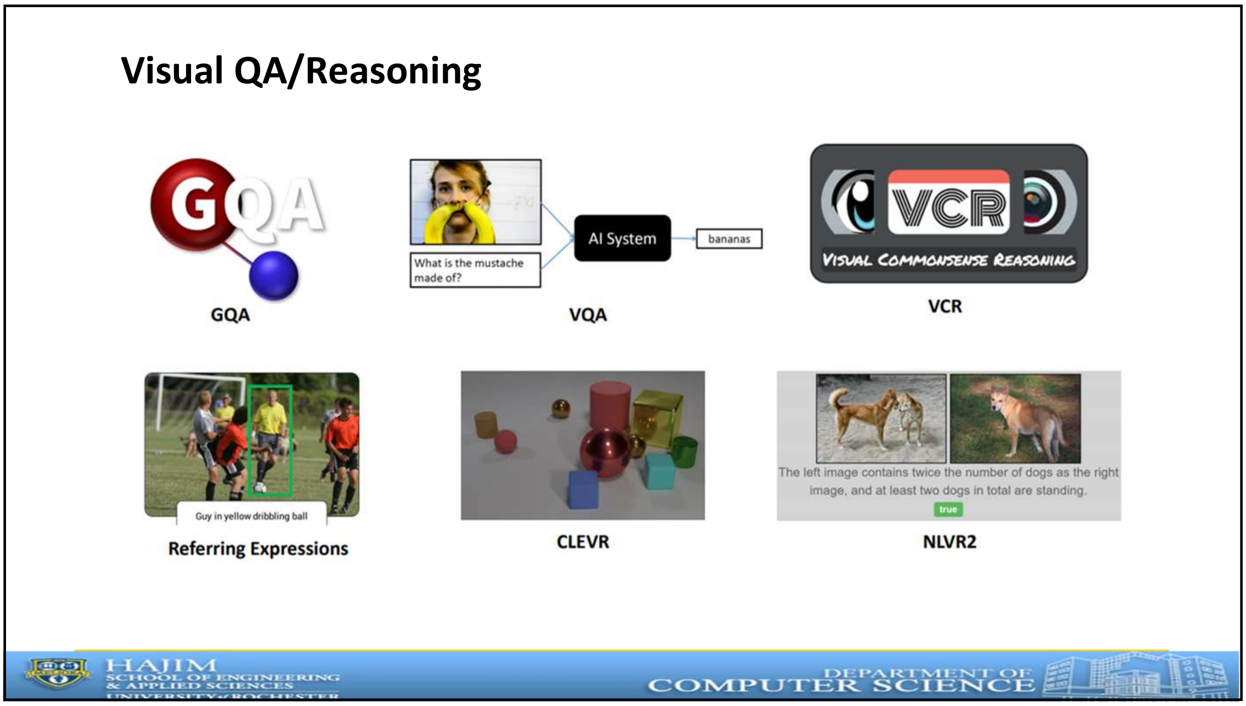


Nr3D Challenge Winner (CVPR 2021)

Nr3D Challenge

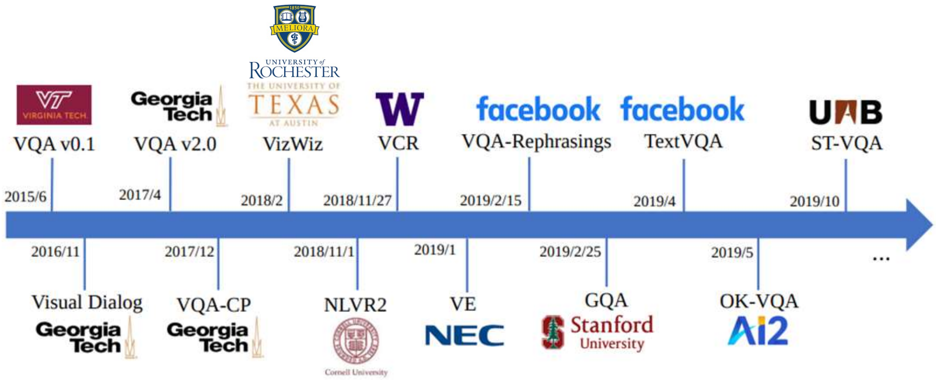
Paper	Overall	Easy	Hard	View-Dependent	View-Independent
SAT	**49.2%**	**56.3%**	**42.4%**	**46.9%**	**50.4%**
TransRefer3D	42.1%	48.5%	36.0%	36.5%	44.9%
FFL-3DOG	41.7%	48.2%	35.0%	37.1%	44.7%
InstanceRefer	38.8%	46.0%	31.8%	34.5%	41.9%
Text-Guided-GNNs	37.3%	44.2%	30.6%	35.8%	38.0%
ReferIt3D	35.6%	43.6%	27.9%	32.5%	37.1%


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
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
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
windowpane
sconce
bed
lamp
pillow
Popular Tasks:
• Text-to-image
• Layout-to-image
• Scene-graph-to-image
• Text-based image editing
• Story visualization
SOTA Models:
• StackGAN
• AttnGAN
• ObjGAN
• ...
Machine Translation/Grammar Induction
BANK
EN: A medium sized child jumps off of a dusty bank.
translate
DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.
evaluate
Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.
Sentence: a squirrel jumps on stump
Parser
a squirrel jumps on stump
c_1: a squirrel
c_2: on stump
c_3: jumps on stump
Credit: VL-CVPR Tutorial. https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research
DEPARTMENT OF COMPUTER SCIENCE


Visual QA/Reasoning
GQA
GQA
What is the mustache made of?
AI System
bananas
VQA
VCR
VISUAL COMMONSENSE REASONING
VCR
Guy in yellow dribbling ball
Referring Expressions
CLEVR
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.
true
NLVR2
HAJIM
SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

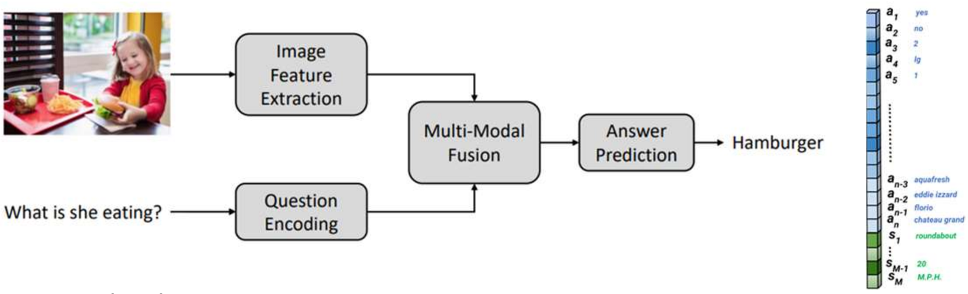
Datasets

- Large-scale annotated datasets have driven tremendous progress in this field



Slide credit: Zhe Gan. CVPR tutorial

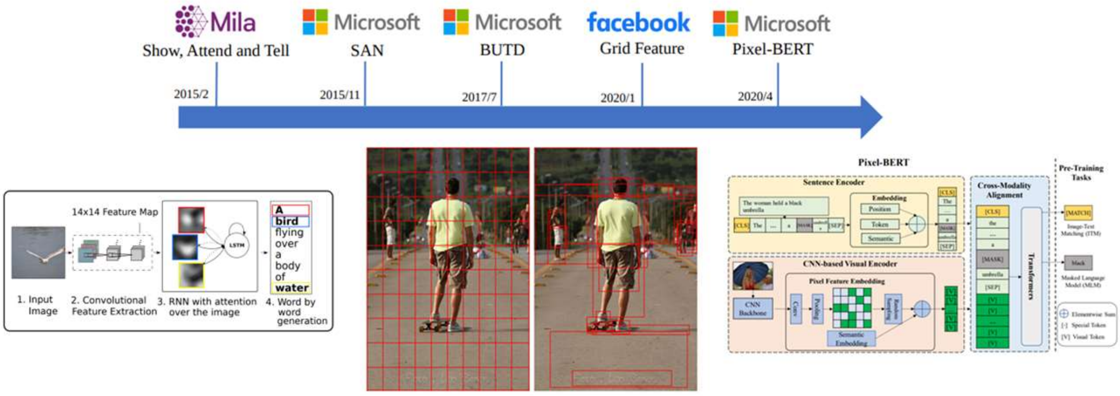
- What a typical system looks like



- Image captioning
- Visual grounding

Image Feature

- From grid features to region features, and to grid features again



HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

Multi-modal Fusion

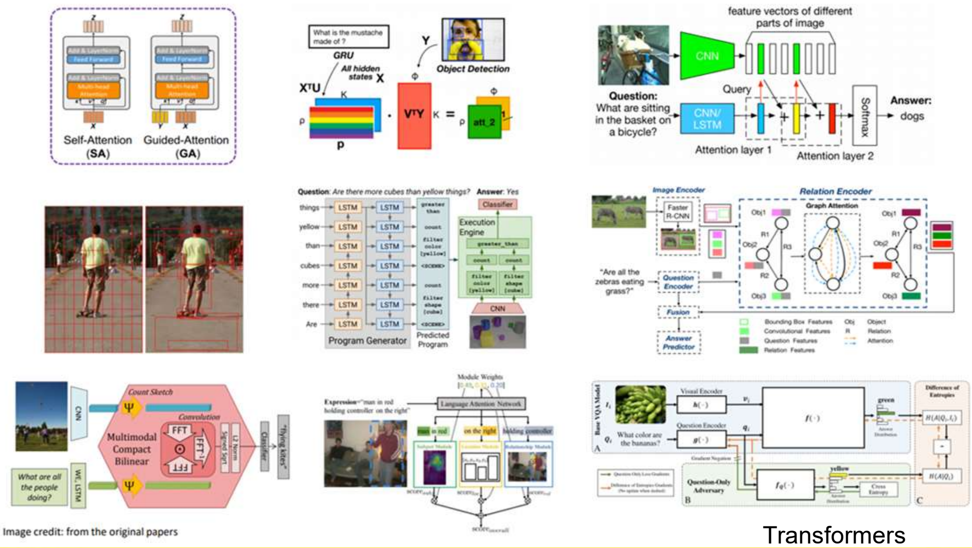
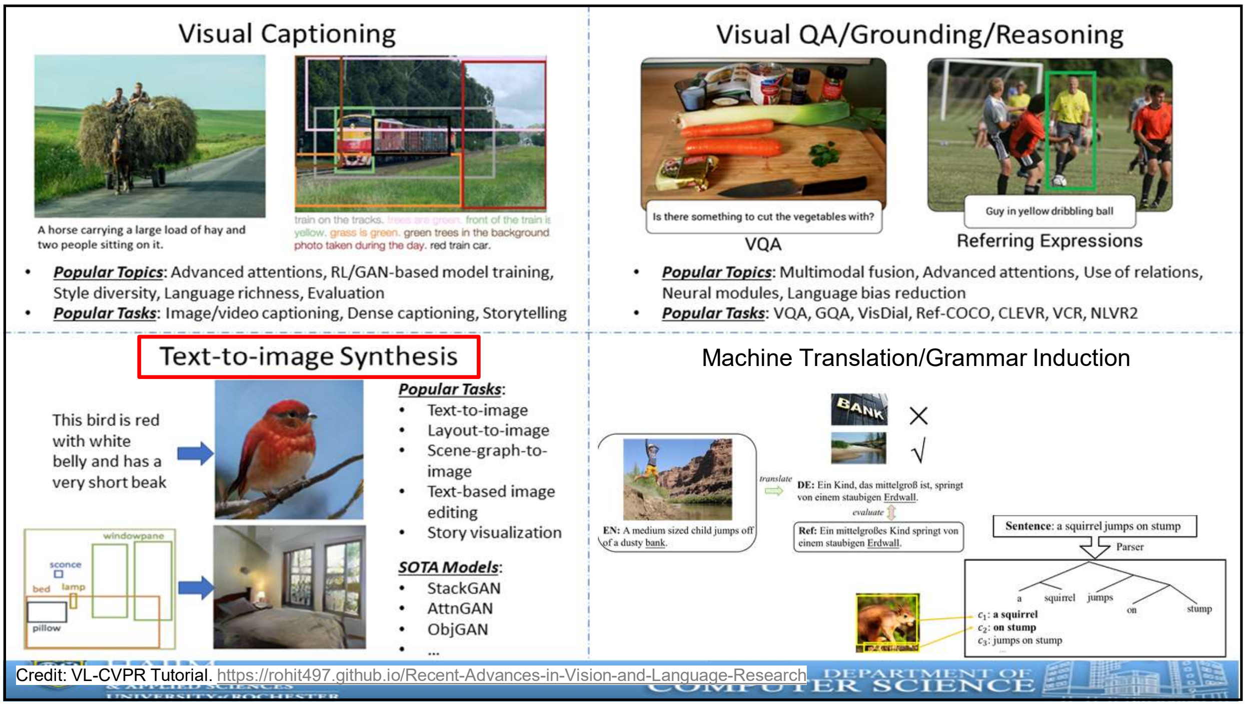


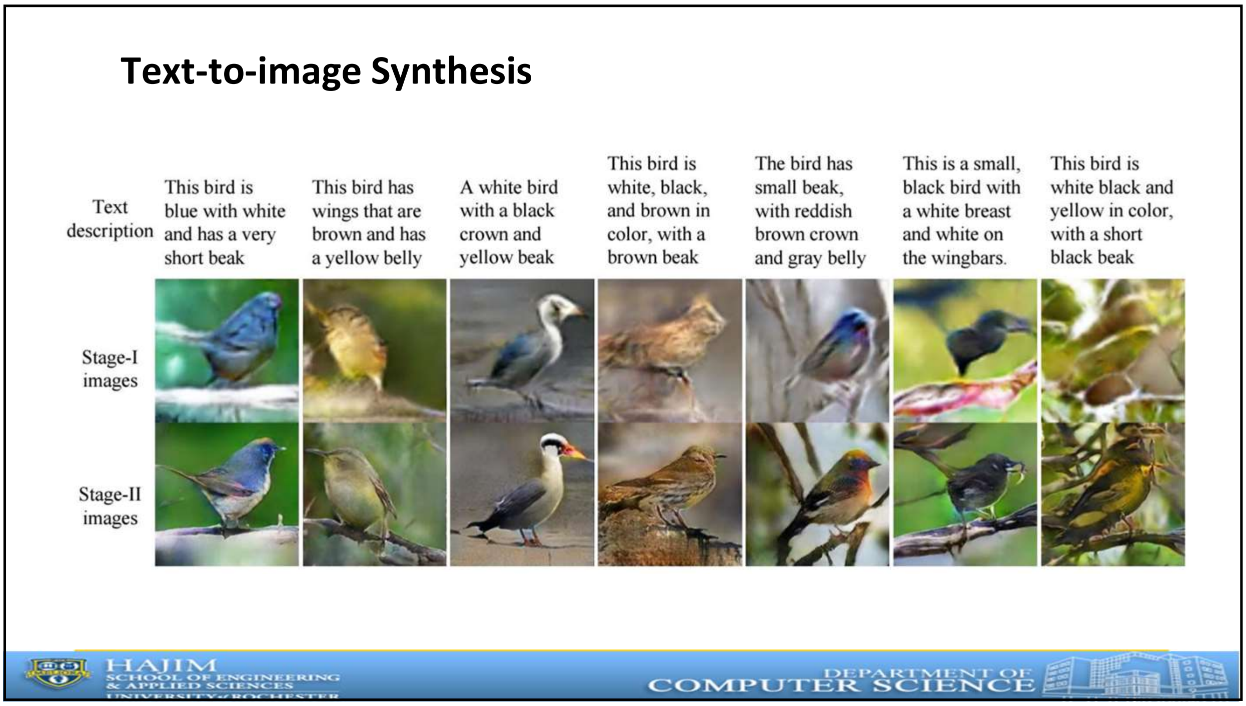
Image credit: from the original papers

Transformers

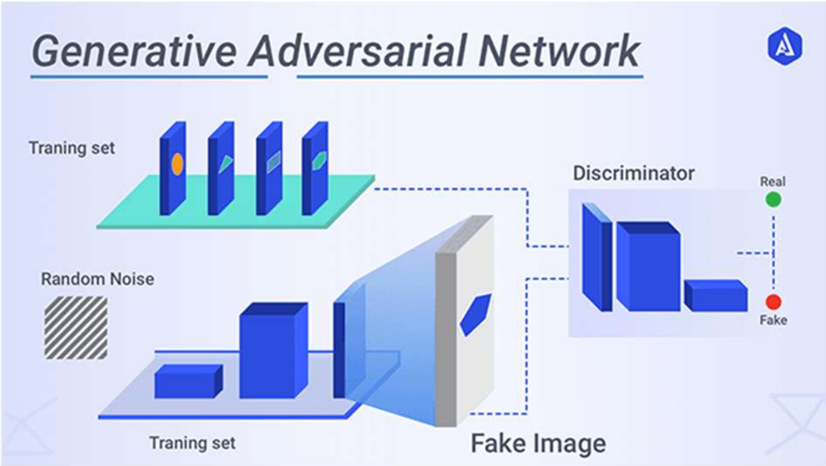
Slide credit: Zhe Gan. CVPR tutorial

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER — DEPARTMENT OF COMPUTER SCIENCE

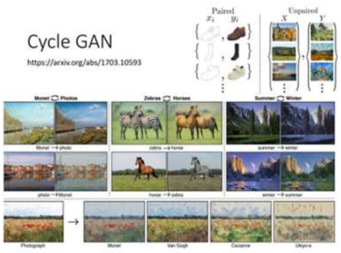

Visual Captioning
A horse carrying a large load of hay and two people sitting on it.
train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background photo taken during the day. red train car.
Popular Topics: Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
Popular Tasks: Image/video captioning, Dense captioning, Storytelling
Visual QA/Grounding/Reasoning
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
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
windowpane
sconce
bed
lamp
pillow
Popular Tasks:
Text-to-image
Layout-to-image
Scene-graph-to-image
Text-based image editing
Story visualization
SOTA Models:
StackGAN
AttnGAN
ObjGAN
...
Machine Translation/Grammar Induction
BANK
×
√
EN: A medium sized child jumps off of a dusty bank.
translate
DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.
evaluate
Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.
Sentence: a squirrel jumps on stump
Parser
a squirrel jumps on stump
c_1: a squirrel
c_2: on stump
c_3: jumps on stump
Credit: VL-CVPR Tutorial. https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research
DEPARTMENT OF COMPUTER SCIENCE

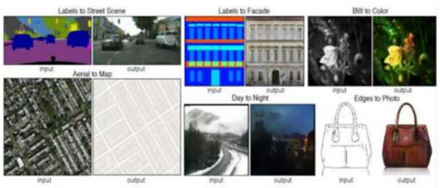

Text-to-image Synthesis
Text description
This bird is blue with white and has a very short beak
This bird has wings that are brown and has a yellow belly
A white bird with a black crown and yellow beak
This bird is white, black, and brown in color, with a brown beak
The bird has small beak, with reddish brown crown and gray belly
This is a small, black bird with a white breast and white on the wingbars.
This bird is white black and yellow in color, with a short black beak
Stage-I images
Stage-II images
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Generative Adversarial Networks (GAN)

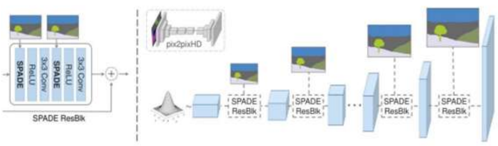


Conditional Image Synthesis



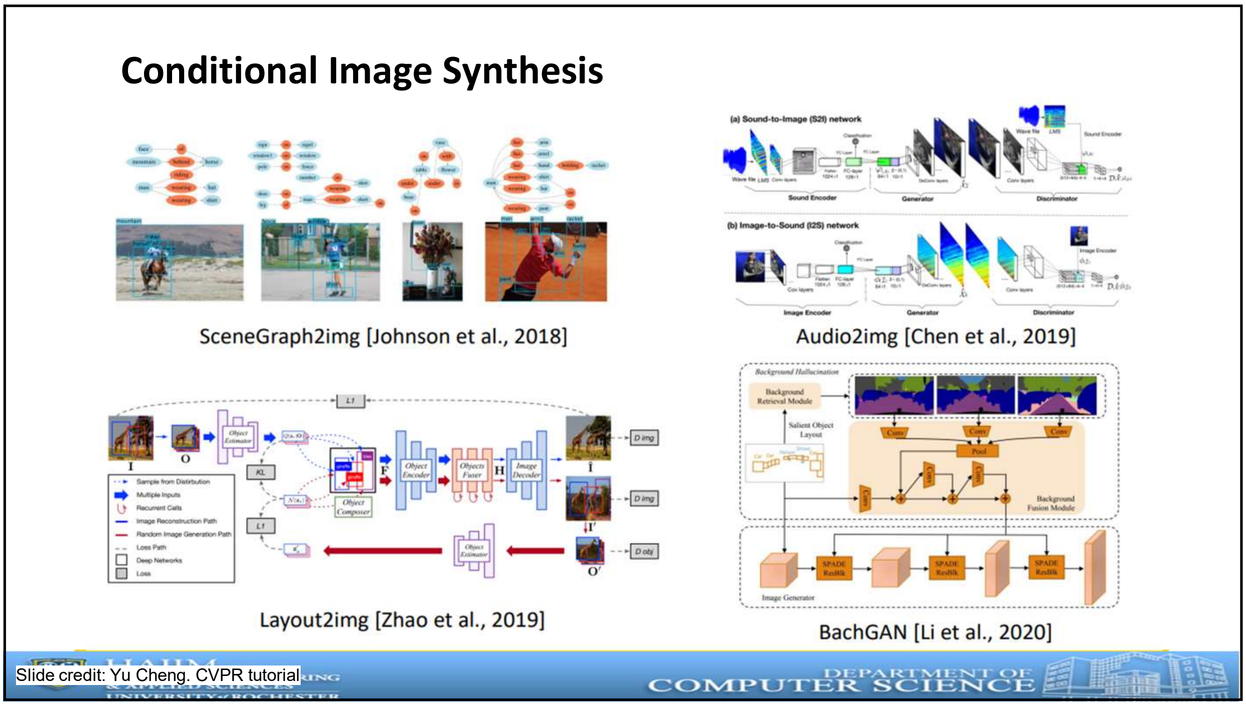


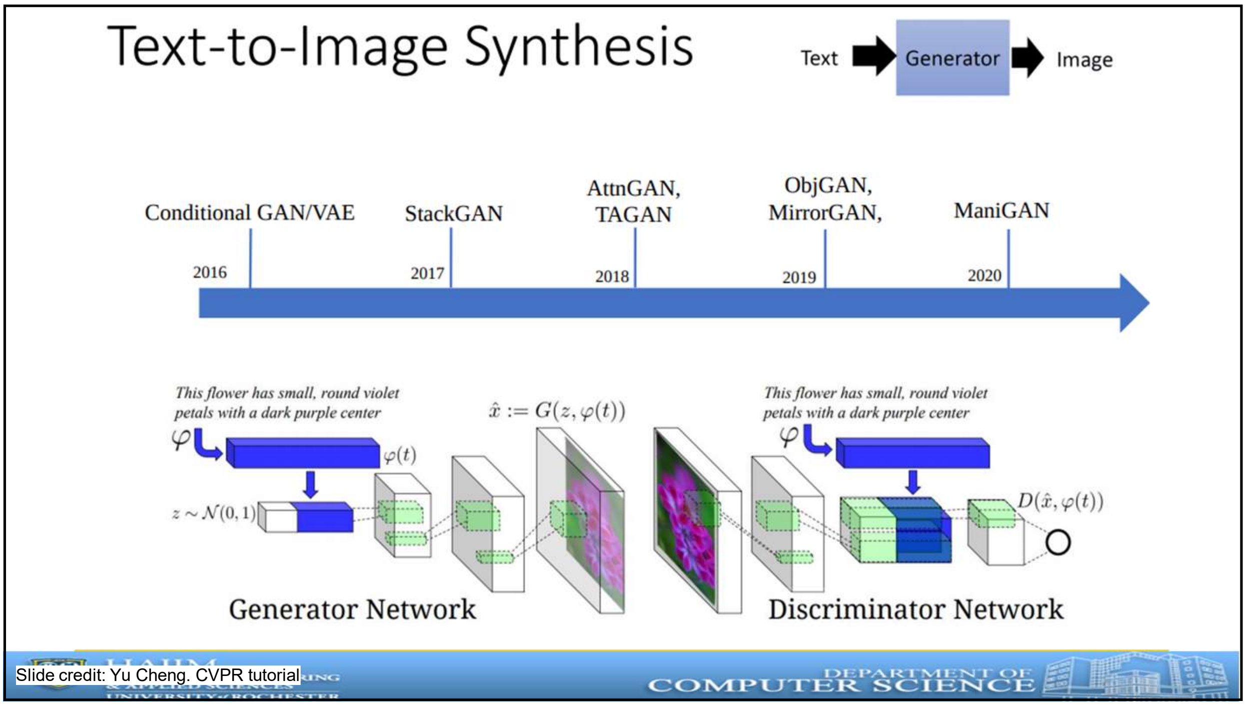
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", arXiv preprint, 2016



SPADE [Park et al., 2019]

Slide credit: Yu Cheng. CVPR tutorial


Conditional Image Synthesis
SceneGraph2img [Johnson et al., 2018]
Audio2img [Chen et al., 2019]
Layout2img [Zhao et al., 2019]
BachGAN [Li et al., 2020]
Slide credit: Yu Cheng. CVPR tutorial
DEPARTMENT OF
COMPUTER SCIENCE


Text-to-Image Synthesis
Text
Generator
Image
Conditional GAN/VAE
StackGAN
AttnGAN,
TAGAN
ObjGAN,
MirrorGAN,
ManiGAN
2016
2017
2018
2019
2020
This flower has small, round violet petals with a dark purple center
$\hat{x} := G(z, \varphi(t))$
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network
Slide credit: Yu Cheng. CVPR tutorial
DEPARTMENT OF
COMPUTER SCIENCE

Dall·e

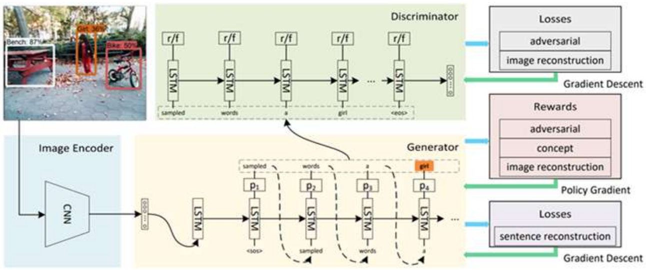
a male mannequin dressed in an orange and black flannel shirt

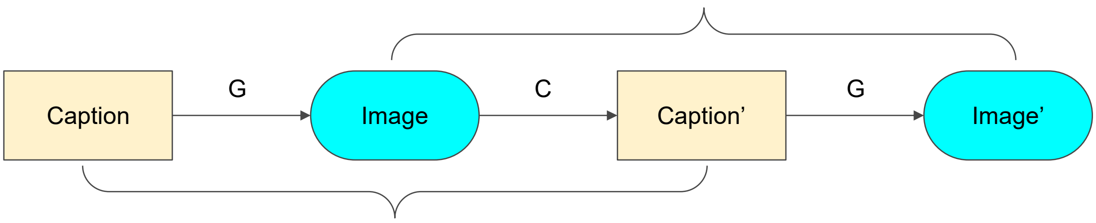
a female mannequin dressed in a black leather jacket and gold pleated skirt

a living room with two white armchairs and a painting of the colosseum. the painting is mounted above a modern fireplace.

a loft bedroom with a white bed next to a nightstand. there is a fish tank beside the bed.

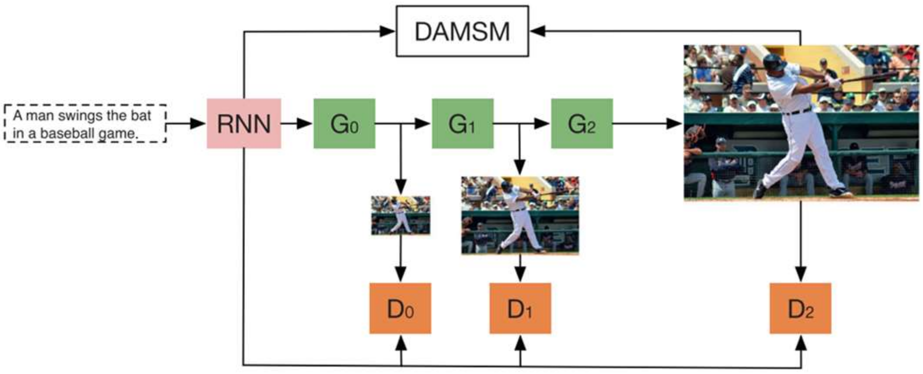
Unsupervised Text-to-Image Synthesis





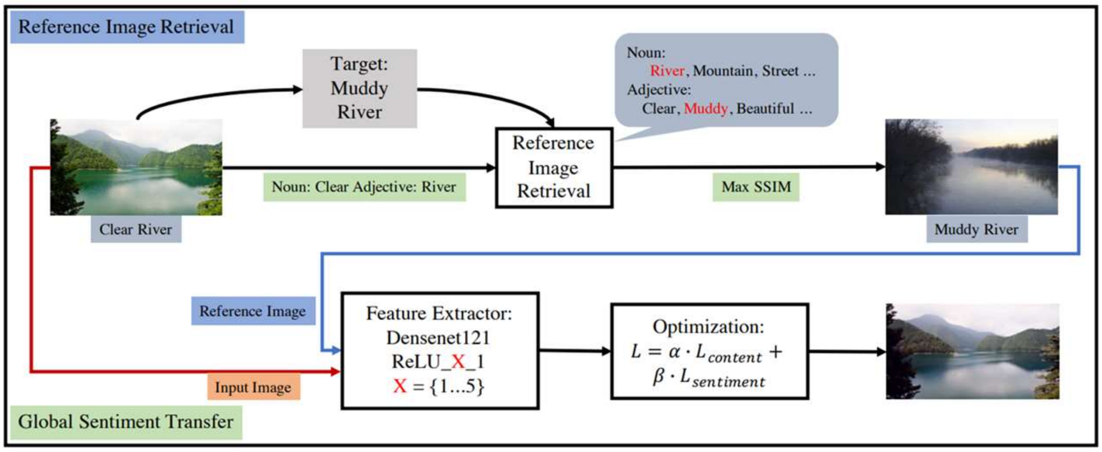
* Yanlong Dong, Ying Zhang, Lin Ma, Zhi Wang, Jiebo Luo, "Unsupervised text-to-image synthesis," Pattern Recognition.

Unsupervised Text-to-Image Synthesis

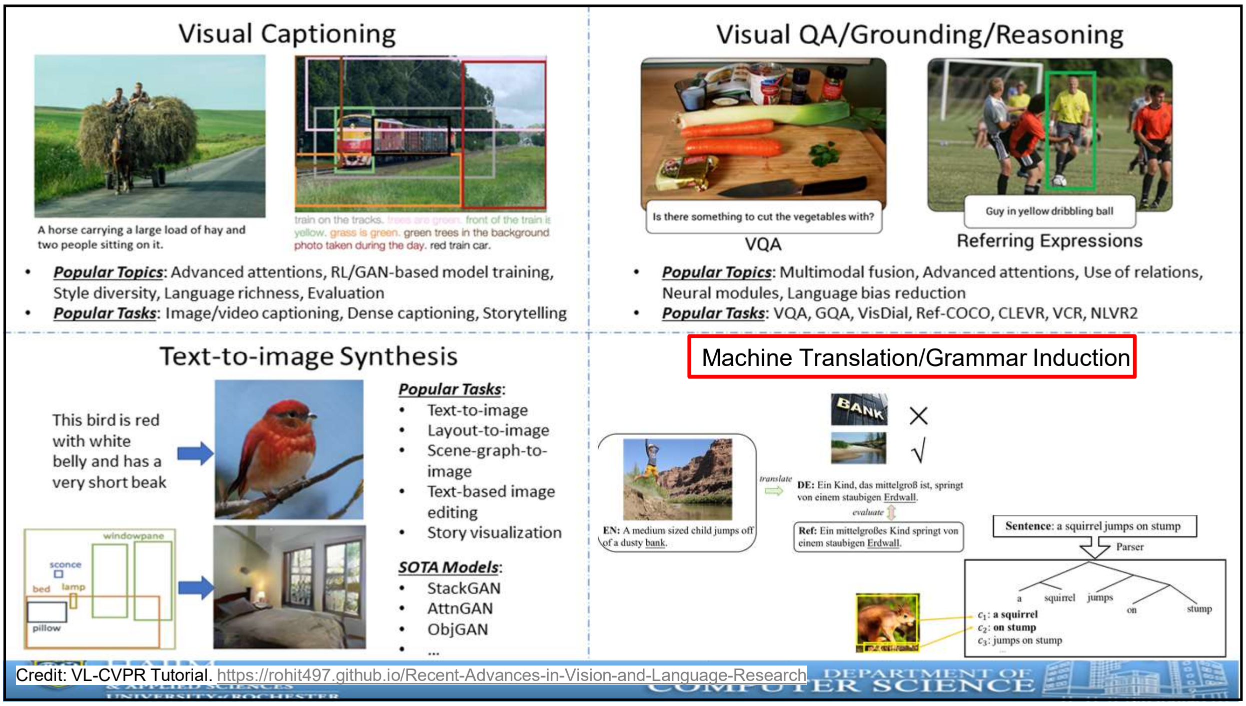


* Yanlong Dong, Ying Zhang, Lin Ma, Zhi Wang, Jiebo Luo, "Unsupervised text-to-image synthesis," Pattern Recognition.

Text to Sentiment



[1] * Jie An, Tianlang Chen, Songyang Zhang, Jiebo Luo, "Global Image Sentiment Transfer," ICPR 2020.
[2] * Tianlang Chen, Wei Xiong, Haitian Zheng, Jiebo Luo, "Image Sentiment Transfer," ACM MM 2020.



Grammar Induction

Goal: Grammar induction aims to capture syntactic information in sentences in the form of constituency parsing trees.

- **Supervised Grammar Induction**
  - Annotating syntactic trees is expensive and time-consuming
  - Limited to the newswire domain in several major languages
- **Unsupervised Grammar Induction**
  - Learn from large-scale unlabeled text
  - Provide evidence for statistical learning

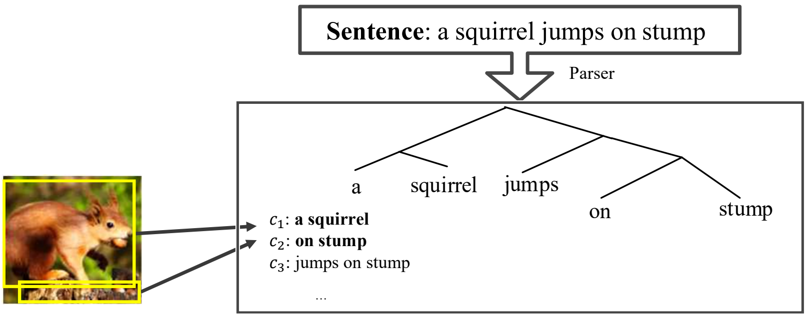
* Songyang Zhang, Linfeng Song, Lifeng Jin, Dong Yu,Jiebo Luo, "Video-aided Unsupervised Grammar Induction," In NAACL 2021 **(Best Long Paper).**

Image-aided Unsupervised Grammar Induction

Images can help us induce syntactic structure. [Shi et al. *ACL'19*]

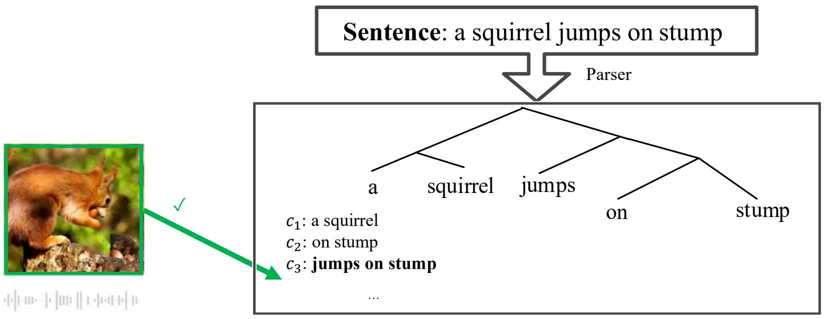
Intuition:

Exploiting regularities between text spans and images.



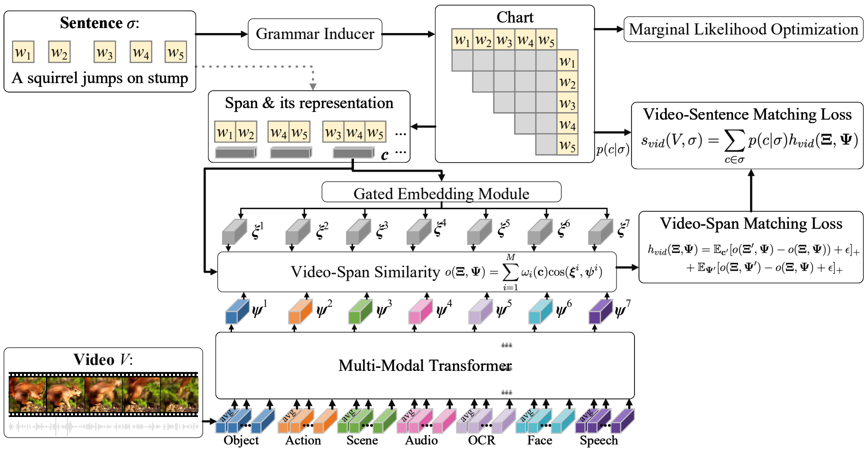
Video-aided Unsupervised Grammar Induction

Motivation: Videos include not only static objects but also actions and state changes useful for inducing *verb phrases (VP)*.



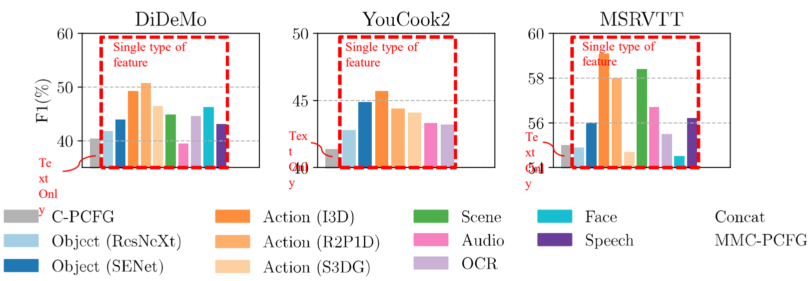
Our Approach

Multi-Modal Compound PCFG (MMC-PCFG) where PCFG stands for probabilistic context-free grammar



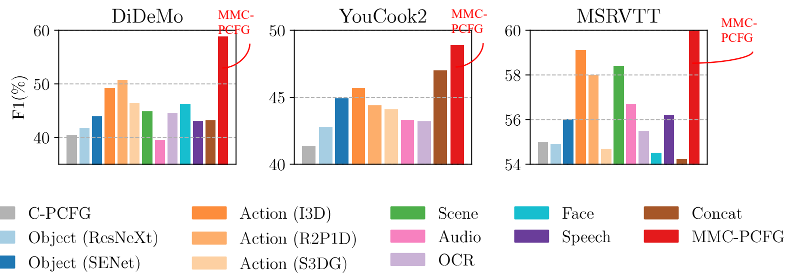
[Gabeur et al. ECCV 2020]

Main Results



Sentence-level F1 scores by singular features on three benchmark datasets.

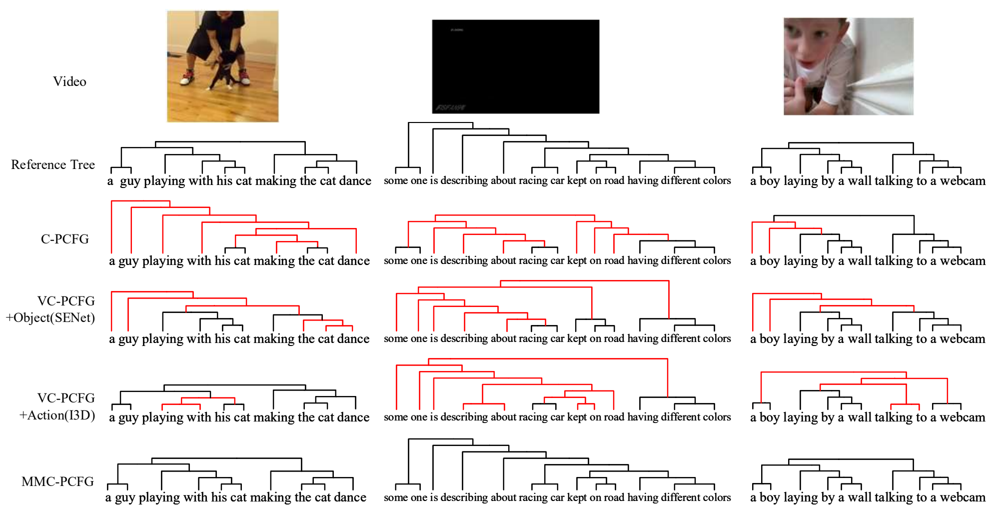
Main Results



Sentence-level F1 scores on three benchmark datasets.

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Qualitative Analysis



HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Multimodal Machine Translation (MMT)

✖

√

EN: A medium sized child jumps off of a dusty bank.

translate

DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.

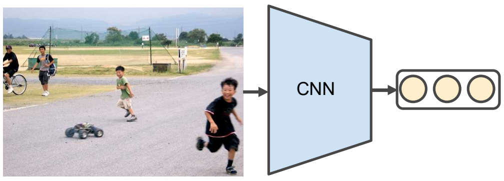
evaluate

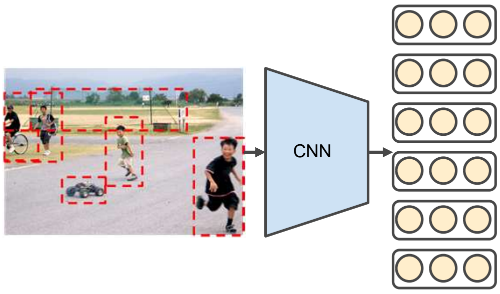
Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.

* Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, Jiebo Luo, "Dynamic Context-guided Capsule Network for Multimodal Machine Translation," *ACM Multimedia Conference*, 2020.

Visual Features used in MMT

- Global visual features (Huang et al.,2016; Calixto et al., 2017, Elliott et al., 2017; Zhou et al., 2018; Calixto et al., 2019)
- Object-based visual features (Huang et al., 2016, Ive et al., 2019)



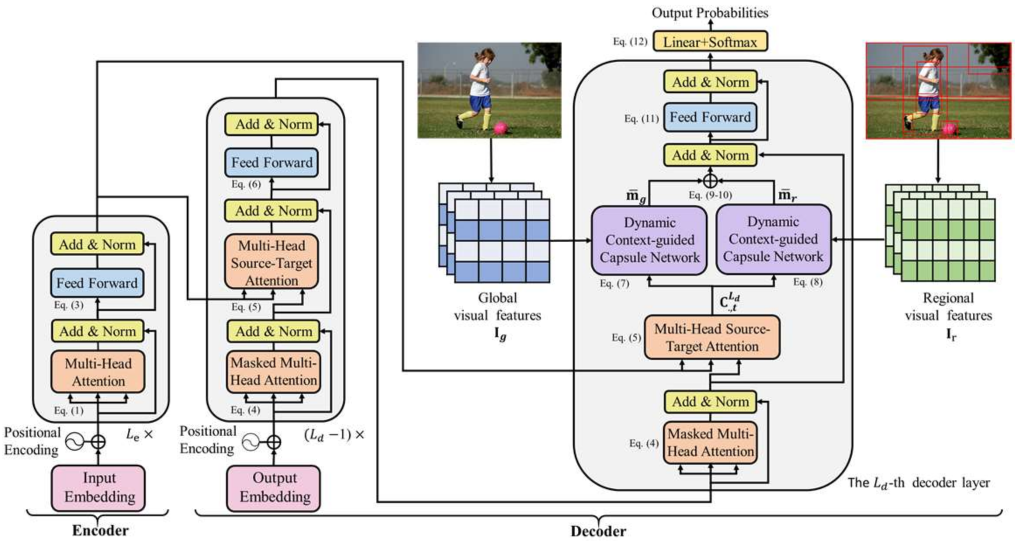


HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

How to Effectively Utilize Visual Features in MMT?

- Exploit visual features as **global visual context** (Huang et al., 2016; Calixto et al., 2017a; Grönroos et al., 2018) ← Lack variability！
- Apply **attention mechanism** to extract visual context (Calixto et al., 2017b; Delbrouck et al., 2017; Helcl et al., 2018; Arslan et al., 2018)← Too many parameters！
- Learn **multimodal joint representations** (Calixto et al., 2019; Elliott et al., 2017; Zhou et al., 2018)← Lack variability！

Dynamic Context-guided Capsule Network (DCCN)



Experiments

Table 1: Experimental results on the En-De translation task.

#	Model	#Params	EN⇒DE					
			Test2016		Test2017		MSCOCO	
			BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
	Existing MMT Systems							
1	Stochastic attention [15]	–	38.2	55.4	–	–	–	–
2	Imagination [21]	–	36.8	55.8	–	–	–	–
3	Fusion-conv [6]	–	37.0	57.0	29.8	51.2	25.1	46.0
4	Trg-mul [6]	–	37.8	57.7	30.7	52.2	26.4	47.4
5	Latent Variable MMT [10]	–	37.7	56.0	30.1	49.9	25.5	44.8
6	Deliberation Network [28]	–	38.0	55.6	–	–	–	–
	Our MMT Systems							
7	Transformer [48]	16.1M	38.4	56.0	29.4	48.8	25.3	44.4
8	Encoder-attention [16]	+1.1M	39.0	56.6	29.9	49.5	26.0	45.5
9	Doubly-attention [26]	+4.0M	38.7	56.4	30.4	49.1	25.5	44.7
10	DCCN	+1.0M	**$39.7^{\ddagger*\triangle\triangle}$**	**$56.8^{\ddagger\triangle}$**	**$31.0^{\ddagger**\triangle}$**	**$49.9^{\ddagger*\triangle\triangle}$**	**$26.7^{\ddagger*\triangle\triangle}$**	**$45.7^{\ddagger\triangle\triangle}$**

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Experiments

Table 1: Experimental results on the En-De translation task.

#	Model	#Params	EN⇒DE					
			Test2016		Test2017		MSCOCO	
			BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
	Existing MMT Systems							
1	Stochastic attention [15]	–	38.2	55.4	–	–	–	–
2	Imagination [21]	–	36.8	55.8	–	–	–	–
3	Fusion-conv [6]	–	37.0	57.0	29.8	51.2	25.1	46.0
4	Trg-mul [6]	–	37.8	57.7	30.7	52.2	26.4	47.4
5	Latent Variable MMT [10]	–	37.7	56.0	30.1	49.9	25.5	44.8
6	Deliberation Network [28]	–	38.0	55.6	–	–	–	–
	Our MMT Systems							
7	Transformer [48]	16.1M	38.4	56.0	29.4	48.8	25.3	44.4
8	Encoder-attention [16]	+1.1M	39.0	56.6	29.9	49.5	26.0	45.5
9	Doubly-attention [26]	+4.0M	38.7	56.4	30.4	49.1	25.5	44.7
10	DCCN	+1.0M	**$39.7^{\ddagger*\triangle\triangle}$**	**$56.8^{\ddagger\triangle}$**	**$31.0^{\ddagger**\triangle}$**	**$49.9^{\ddagger*\triangle\triangle}$**	**$26.7^{\ddagger*\triangle\triangle}$**	**$45.7^{\ddagger\triangle\triangle}$**

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Case Study

➢ Global visual features

Objects: [shorts, man, woman, sign, sidewalk, umbrella, dress, glasses, girl]

EN: A girl wearing a mask rides on a man's shoulders through a crowded sidewalk.

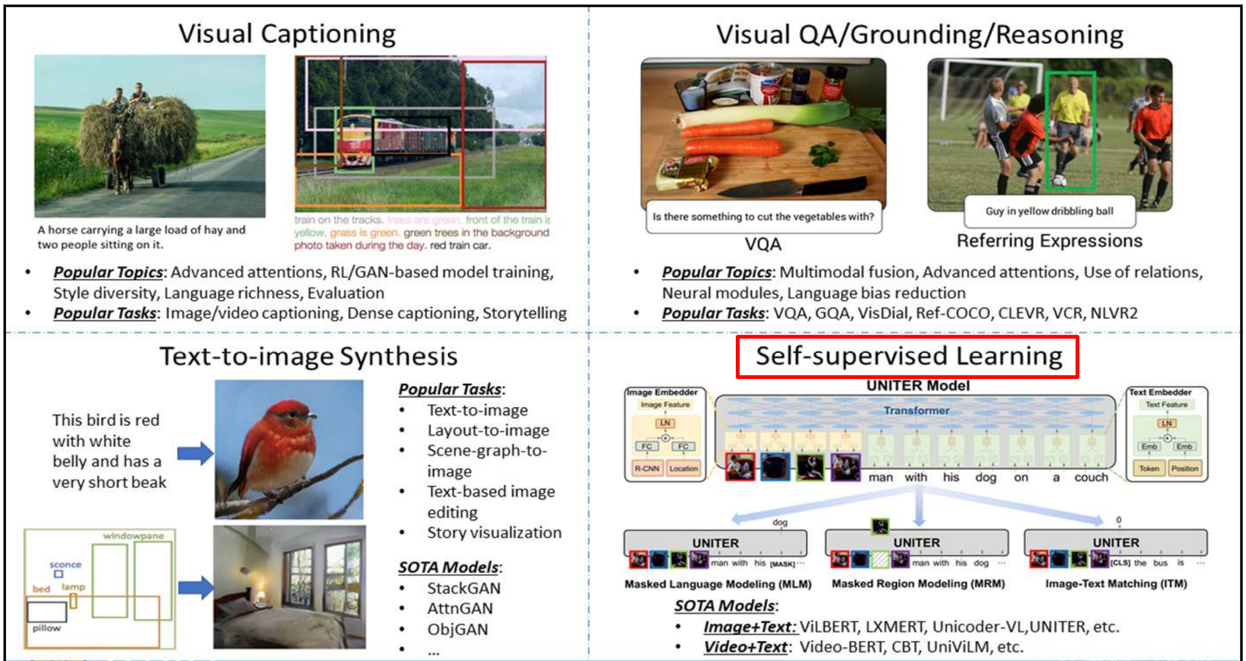
Ref (DE): ... reitet auf den Schultern eines Mannes ...

Transformer: ... fährt auf den Schultern eines Mannes ...

Encoder-attention: ... fährt auf den Schultern eines Mannes ...

Doubly-attention : ... fährt auf den Schultern eines Mannes ...

DCCN: ... reitet auf den Schultern eines Mannes ...



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

Supervised Learning

Datasets + Labels

- MS COCO's Image Captioning:
  - 120,000 images
  - 5 sentences / image
  - 15 cents / sentence
  - +20% AWS processing fee

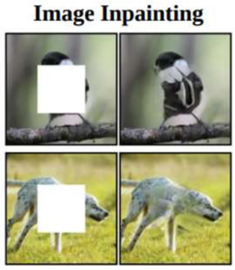

$108,000


Datasets ~~+ Labels~~: Self-Supervised Learning for Vision

Image Colorization

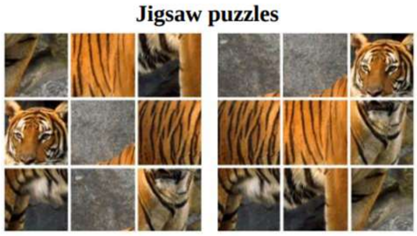
[Zhang et al. ECCV 2016]

Image Inpainting



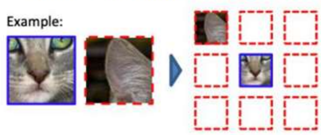
[Pathak et al. CVPR 2016]

Jigsaw puzzles



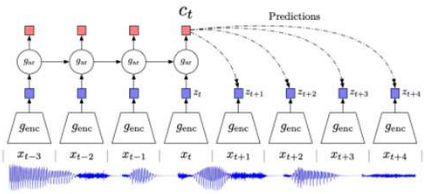
[Noroozi et al. ECCV 2016]

Relative Location Prediction

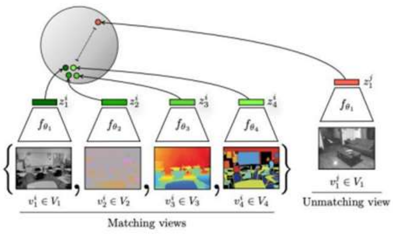


[Doersch et al. ICCV 2015]

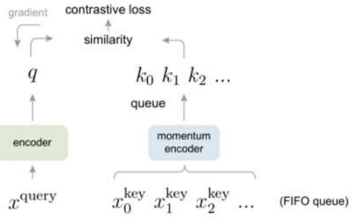
Datasets ~~+ Labels~~: Self-Supervised Learning for Vision



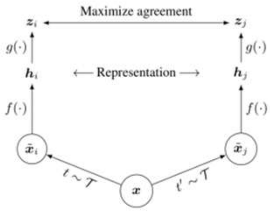
CPC; Ord et al, 2019



CMC; Tian et al, 2019



MOCO; He et al, 2019



SimCLR; Chen et al, 2020

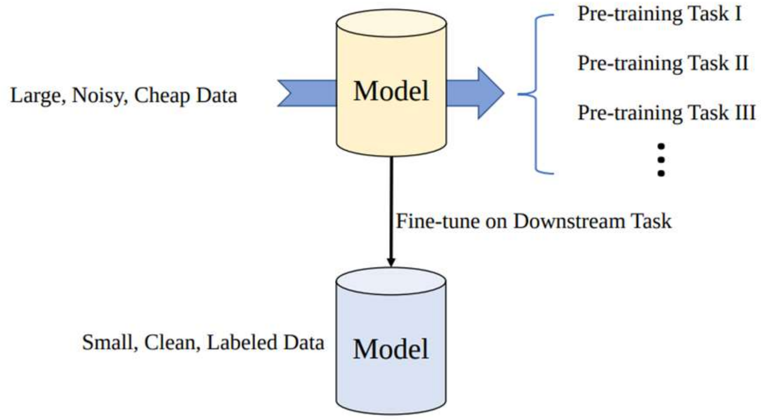
Datasets ~~+ Labels~~: Self-Supervised Learning for NLP

[Devlin et al. NAACL 2019]

[Radford et al. 2019]

1

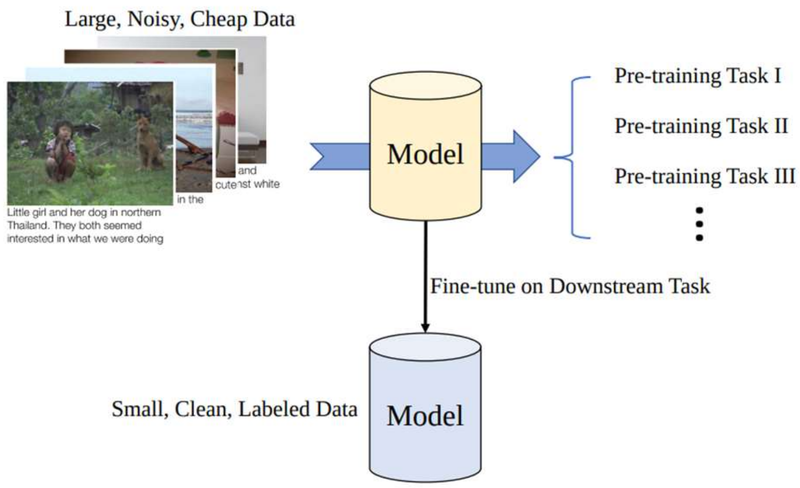
Pre-training + Finetuning



Slide credit: Licheng Yu , Linjie Li and Yen-Chun Chen CVPR tutorial

DEPARTMENT OF COMPUTER SCIENCE

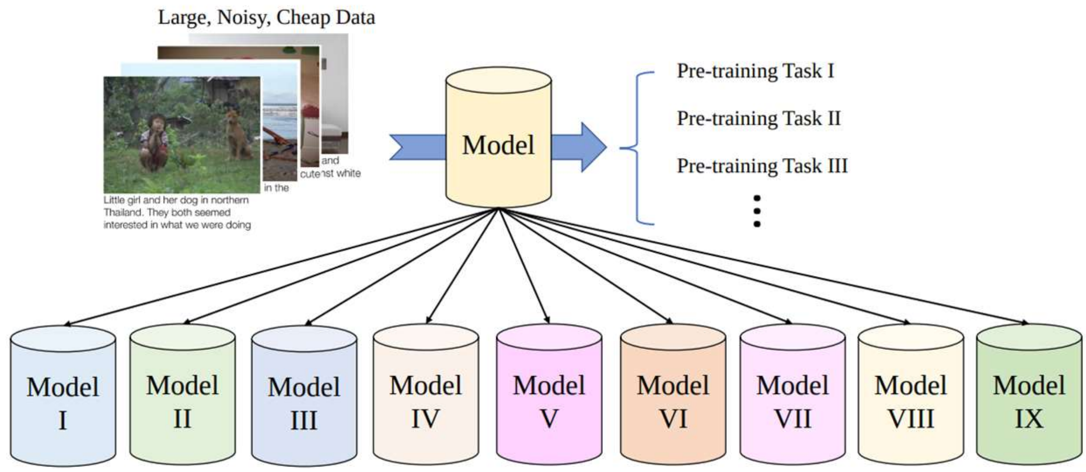
Two-Stage Training Pipeline



Slide credit: Licheng Yu , Linjie Li and Yen-Chun Chen CVPR tutorial

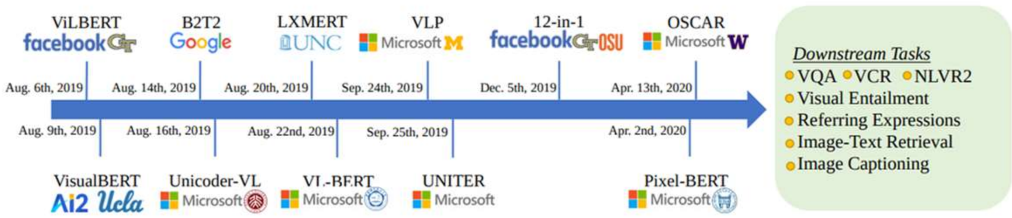
DEPARTMENT OF COMPUTER SCIENCE

Generalization



Slide credit: Licheng Yu , Linjie Li and Yen-Chun Chen CVPR tutorial

Recent VLP Methods



CLIP, ALIGN, Wenlan, VinVL

Slide credit: Licheng Yu , Linjie Li and Yen-Chun Chen CVPR tutorial

Common Pre-training Data for Vision + Language

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Conceptual Caption

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

SBU Caption

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

Slide credit: Licheng Yu , Linjie Li and Yen-Chun Chen CVPR tutorial

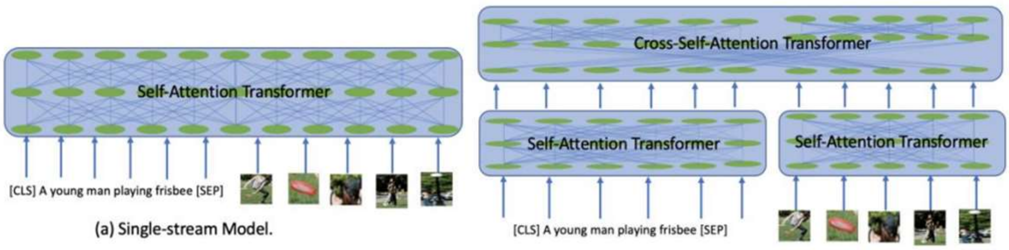
Recent Large Scale Pre-training

Clip	OpenAI	300M
ALIGN	Google	**1.8B**
Wenlan	Renmin University	500M
WIT	Google	37.6M

CLIP: 18 days to train on 592 V100 GPUs

Model Architecture

VLP: (1) architecture + (2) pre-training tasks



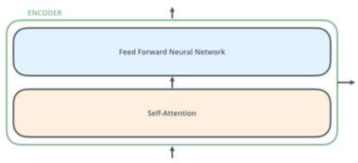
(a) Single-stream Model.

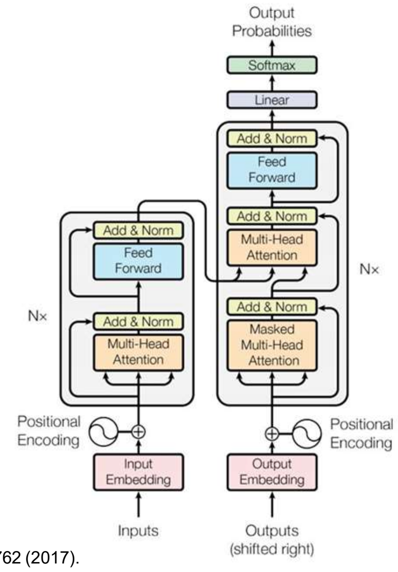
(b) Two-stream Model.

Transformer

Encoder

Decoder

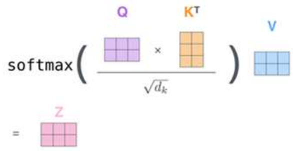




Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).

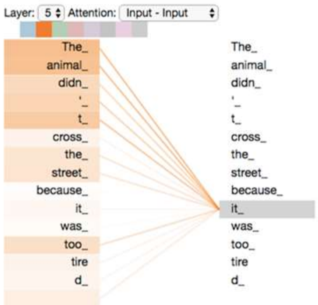
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

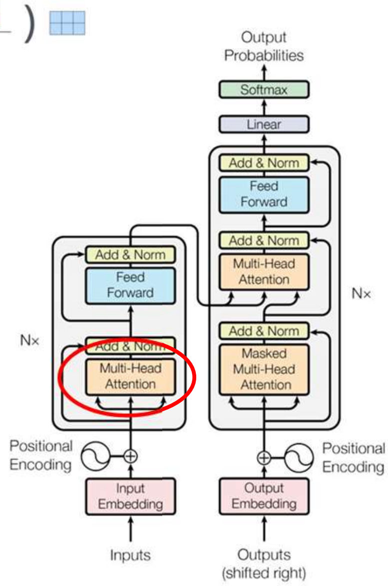
Transformer



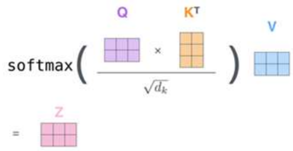
$$attention\_output = Attention(Q, K, V)$$

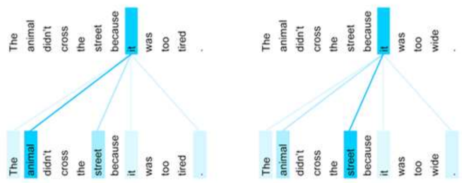
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

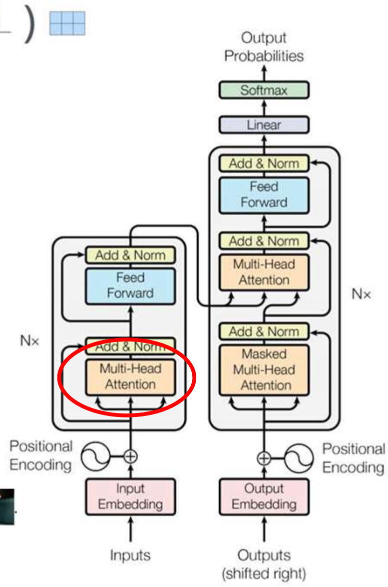




Transformer



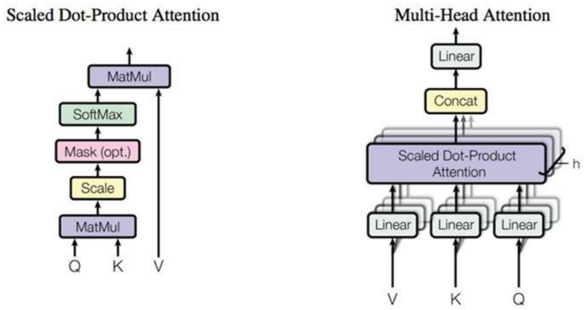




Transformer

$$attention\_output = Attention(Q, K, V)$$

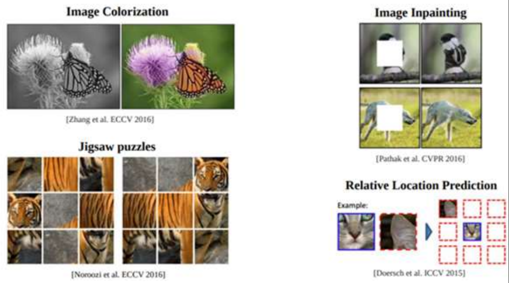
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

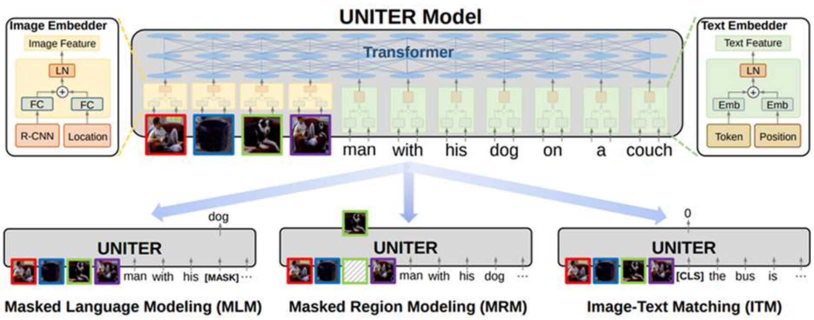


HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER

DEPARTMENT OF COMPUTER SCIENCE

VLP Pretraining Tasks



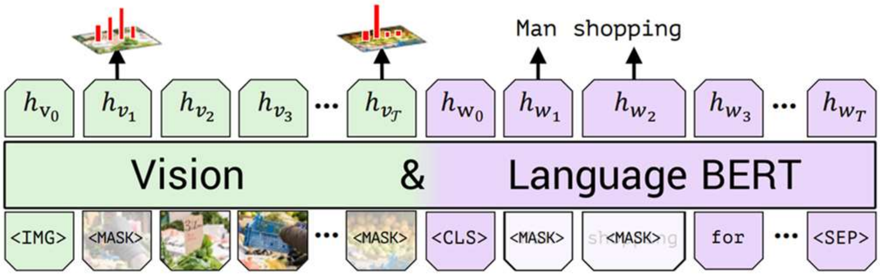


HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER

DEPARTMENT OF COMPUTER SCIENCE

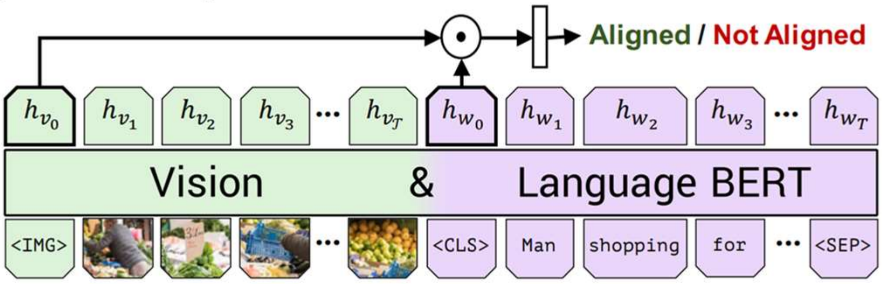
Pretraining Tasks

Masked region/language modeling



Pretraining Tasks

Image-text matching



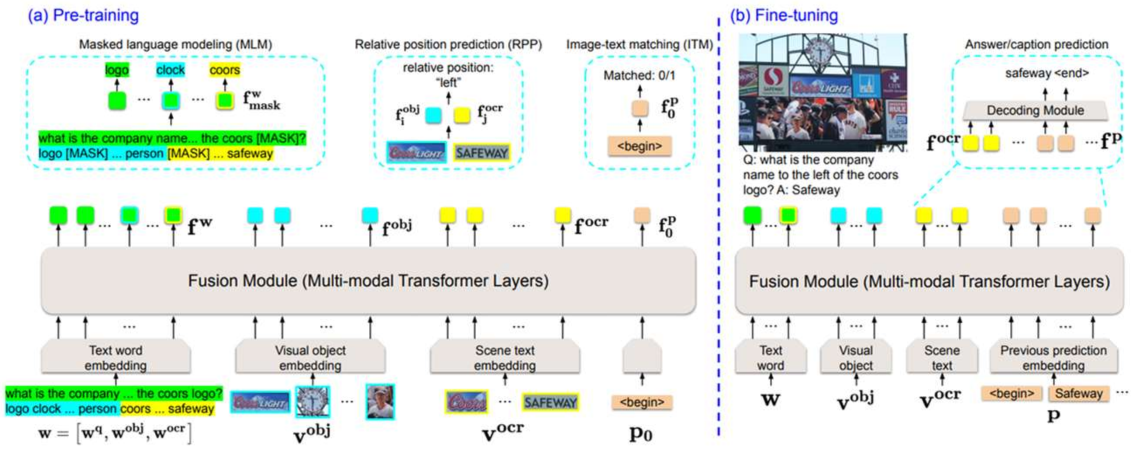
[1] * Tianlang Chen, Jiajun Deng and Jiebo Luo. "Adaptive Offline Quintuplet Loss for Image-Text Matching." ECCV 2020.
[2] * Tianlang Chen and Jiebo Luo. "Expressing Objects just like Words: Recurrent Visual Embedding for Image-Text Matching." AAAI 2020.
[3] * Quanzeng You, Zhengyou Zhang, Jiebo Luo. “End-to-end Convolutional Semantic Embeddings.” CVPR 2018.

VL Pretraining with Reading Comprehension

a	b	c
Model: a macdonald ' s sign that is on a brick wall	**Model:** a sign that has the time of 12 : 37 on it	**Model:** a ruler that has the number 2003 on it
Human: A tile wall with a red circle on it reading Mornington Crescent	**Human:** A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks	**Human:** An old artifact being measured by a ruler that shows it is around 40 millimeters wide

* Yang, Lu, Yin, Florencio, Wang, Zhang, Zhang, Luo. "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption." In CVPR 2021.

VL Pretraining with Reading Comprehension



2

Method

- **VL alignment tasks**

Masked language modeling (MLM)

[CLS] what [MASK] bar is … [SEP] honey [MASK] yellow tail ... [SEP] juice wine ... [SEP]

Question OCR Token Object Token

Contrastive loss

[CLS] what candy bar is … [SEP] [Other text seq. in batch] [SEP] juice wine ... [SEP]

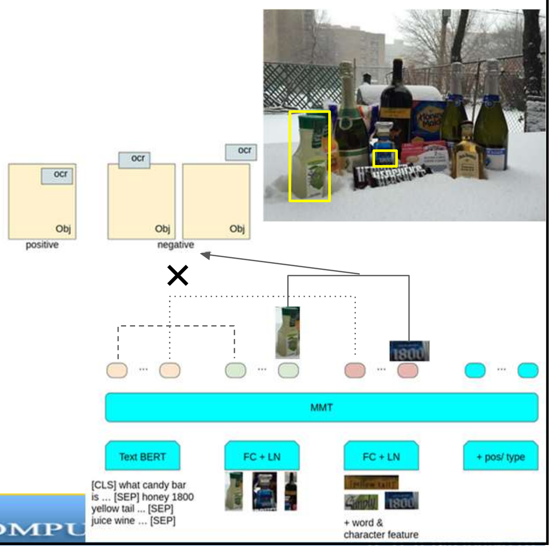
Question OCR Token Object Token

Region Alignment Tasks

- Obj ⇔ OCR region relationship
- Relative position prediction

Q: what is written ***on*** the man's shirt? A: Life cycle

Q: what number is ***on*** the bike on the right? A: 317



Experiments

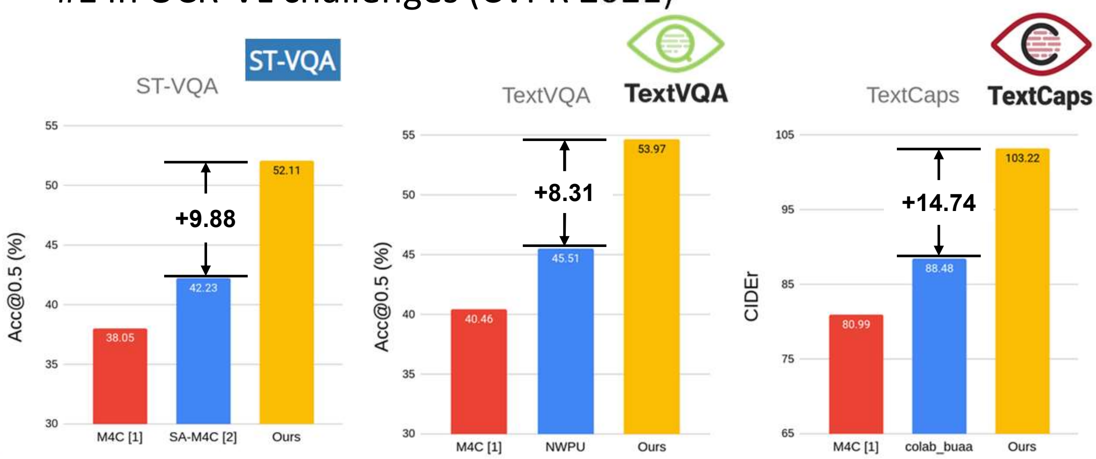
- **Datasets and Metrics**
- TextVQA
  - 28,408 images (22K training)
  - Soft-voting of 10 answers
- Text($Acc = \{0.3\#matched, 1\}$
  - Same images, 110K captions
  - Captioning metrics

Question: what is this sign warning drivers of?
Prediction Answer: road work.
GT Answer: [**'road work'**, 'road work ahead', ... , 'road work ahead', **'road work'**]. (#matched = 2)
Acc: 0.6.

[1] Hu, Ronghang, et al. "Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA." In CVPR 2020. (M4C)
[2] Singh, Amanpreet, et al. "Towards vqa models that can read." In CVPR 2019.
[3] Sidorov, Oleksii, et al. "TextCaps: a Dataset for Image Captioning with Reading Comprehension." In ECCV 2020.

TextVQA, TextCaps

- #1 in OCR-VL challenges (CVPR 2021)

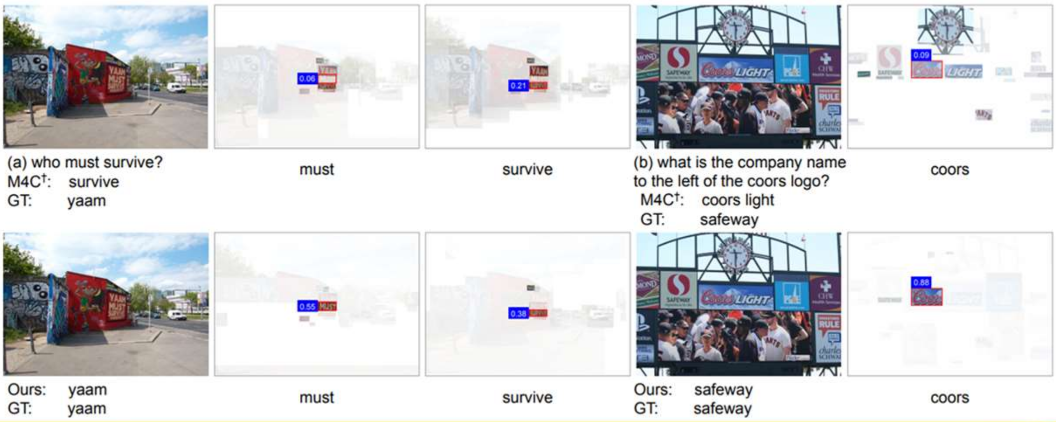


[1] Hu, Ronghang, et al. "Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA." In CVPR 2020. (M4C)
[2] Kant, Yash, et al. "Spatially Aware Multimodal Transformers for TextVQA." In ECCV 2020. (SA-M4C)

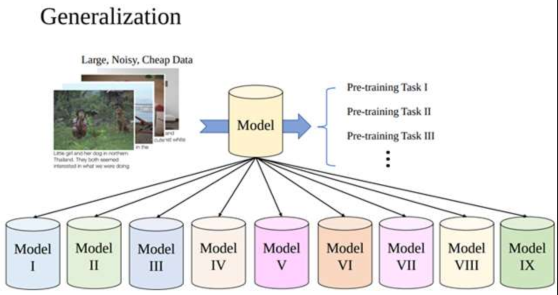
TextVQA, TextCaps

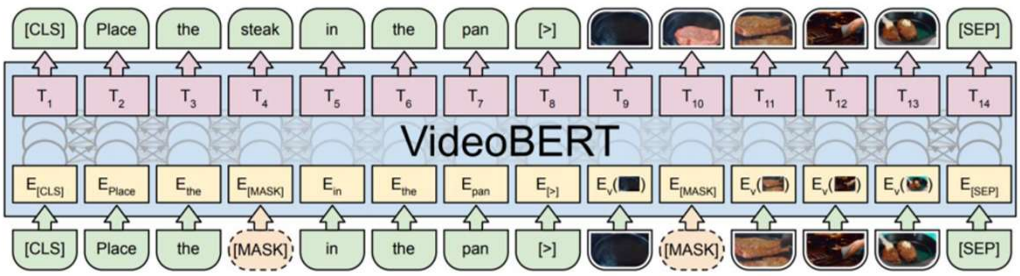
- **Latent Grounding**

Coref Type	W/O TAP	With TAP
Text Word → Scene Text	0.0477	**0.3514**
Scene Text → Text Word	0.0473	**0.5206**
Visual Object → Scene Text	0.0045	**0.0130**
Scene Text → Visual Object	0.0337	**0.0680**



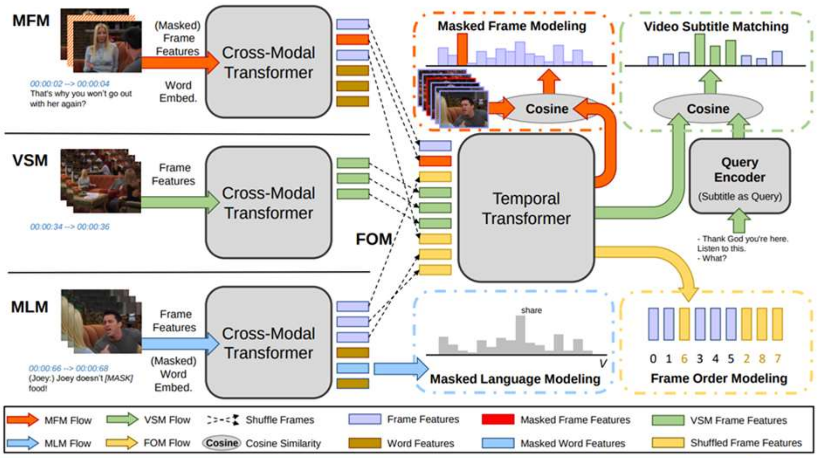
Video-Language Pretraining





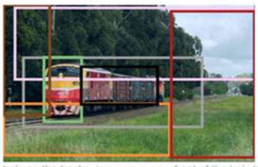
Video-Language Pretraining

- Token representation
- Pretraining tasks



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


Is there something to cut the vegetables with?

VQA

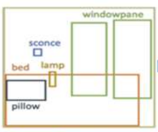

Guy in yellow dribbling ball

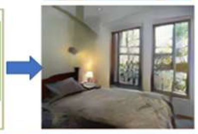
Referring Expressions

- ***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

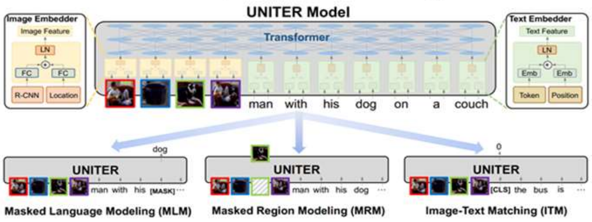




Popular Tasks:
- Text-to-image
- Layout-to-image
- Scene-graph-to-image
- Text-based image editing
- Story visualization

SOTA Models:
- StackGAN
- AttnGAN
- ObjGAN
- ...

Self-supervised Learning

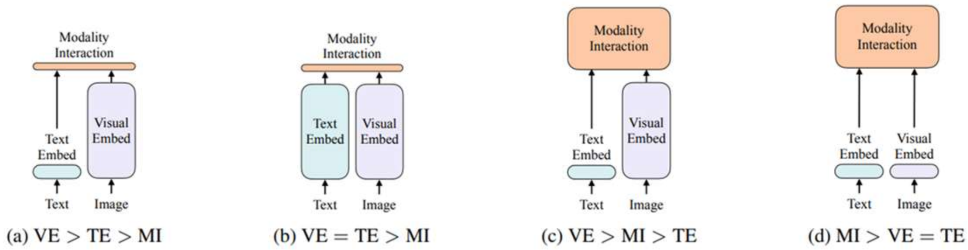


SOTA Models:
- ***Image+Text:*** ViLBERT, LXMERT, Unicoder-VL, UNITER, etc.
- ***Video+Text***: Video-BERT, CBT, UniViLM, etc.

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

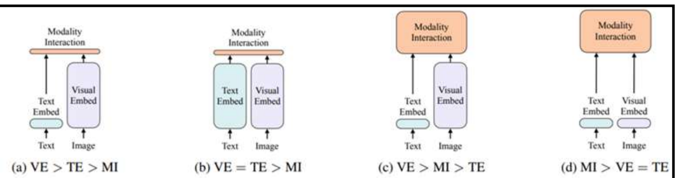
Future Directions

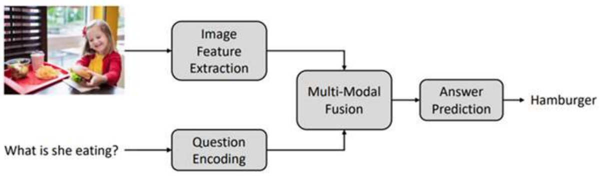
- General vision-language models
  - Unified pipeline
    - Architecture

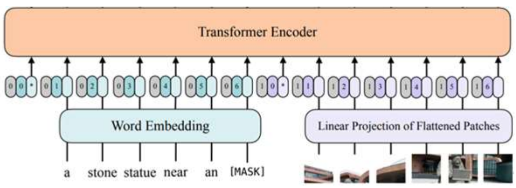


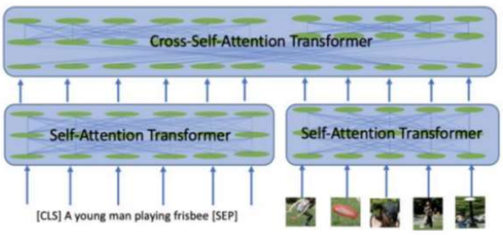
Wonjae Kim, et al. "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision." In Arxiv 2102.03334

Future Directions





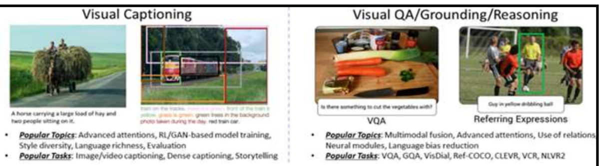




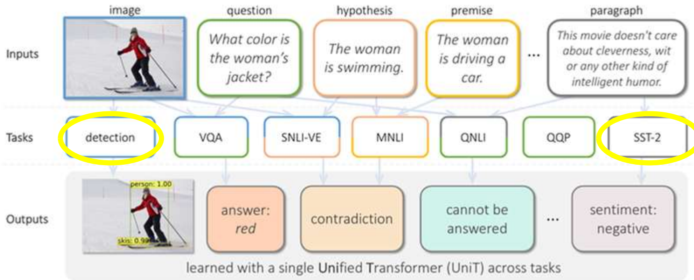
- Unified visual-text representation

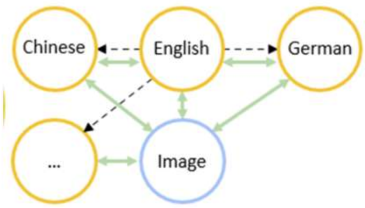
HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES UNIVERSITY of ROCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Future Directions



- General vision-language models
  - Unified pipeline
    - Tasks: V-L => V-V, V-L, L-L

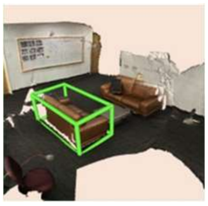


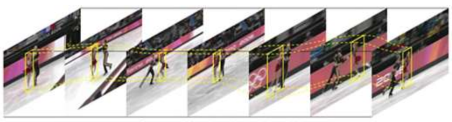


[1] Hu, Ronghang, et al. "Transformer is all you need: Multimodal multitask learning with a unified transformer." arXiv:2102.10772.
[2] Mingyang Zhou, et al. "UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training." arXiv:2104.00332.

Future Directions

- General vision-language models
  - Extra modalities, multi-lingual

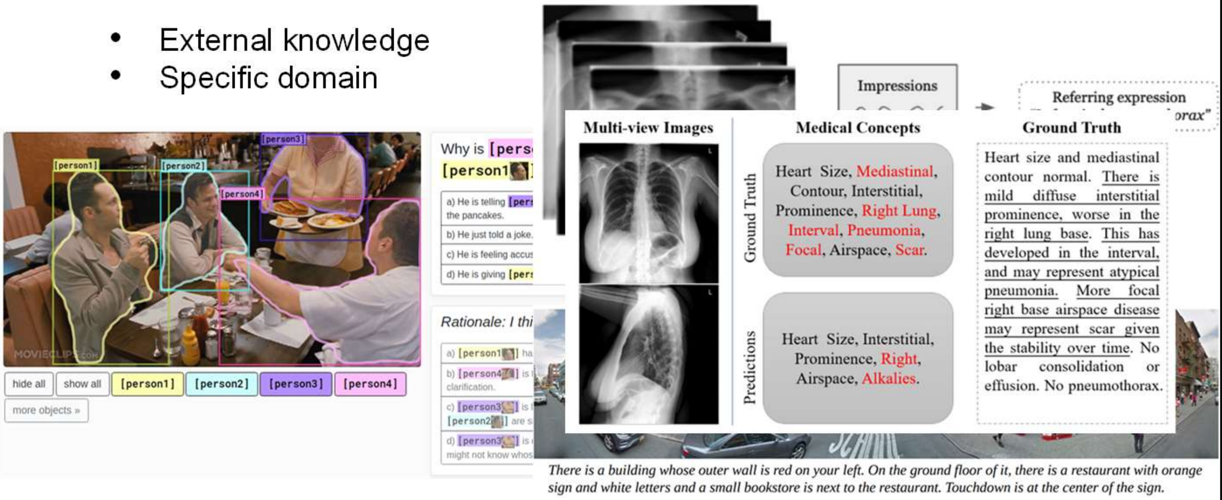




Query: "female skater in red."

Future Directions

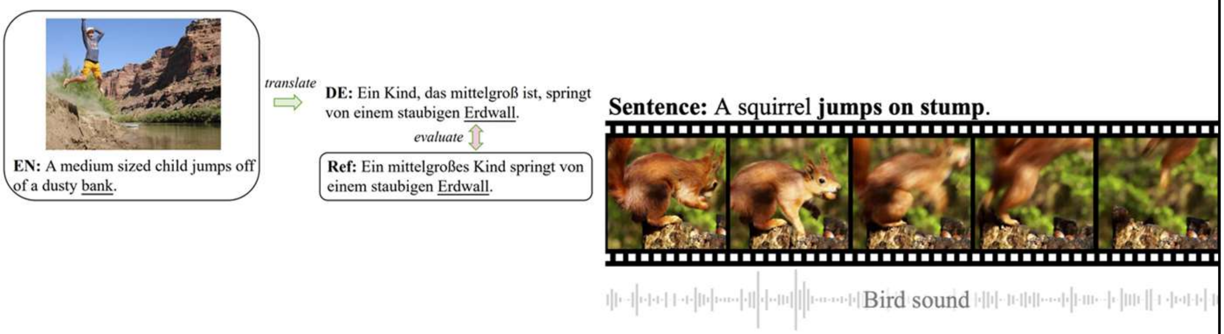
- External knowledge
- Specific domain



[1] Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." In CVPR 2019.
[2] * Yuan, Jianbo, et al. "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment." In MICCAI 2019.
[3] Tam, Leo K., et al. "Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies." In MICCAI 2020.
[4] Luo, Weixin, et al. "SIRI: Spatial Relation Induced Network For Spatial Description Resolution." In Neurips 2020.

Future Directions

- New NLP tasks
  - Multimodal machine translation
  - Video-aided grammar induction



[1] * Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, Jiebo Luo. “Dynamic Context-guided Capsule Network for Multimodal Machine Translation.” ACM MM 2020. (oral presentation)
[2] * Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, Jiebo Luo. “Video-aided Unsupervised Grammar Induction.” NAACL 2021. (**Best Long Paper**)

Keep in mind our original aspiration. Keep marching forward.

Computer vision is an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos. (Wikipedia)

“Vision is the process of discovering (and describing) from images what is present in the world, and where it is.”

-- David Marr, *Vision (1982)*

Thank you!