NPFL099 Statistical Dialogue Systems
11. Multimodal Systems
(+some notes on voice-based systems & domain adaptation)

http://ufal.cz/npfl099

Ondřej Dušek, Zdeněk Kasner, Mateusz Lango, Ondřej Plátek

19. 12. 2024

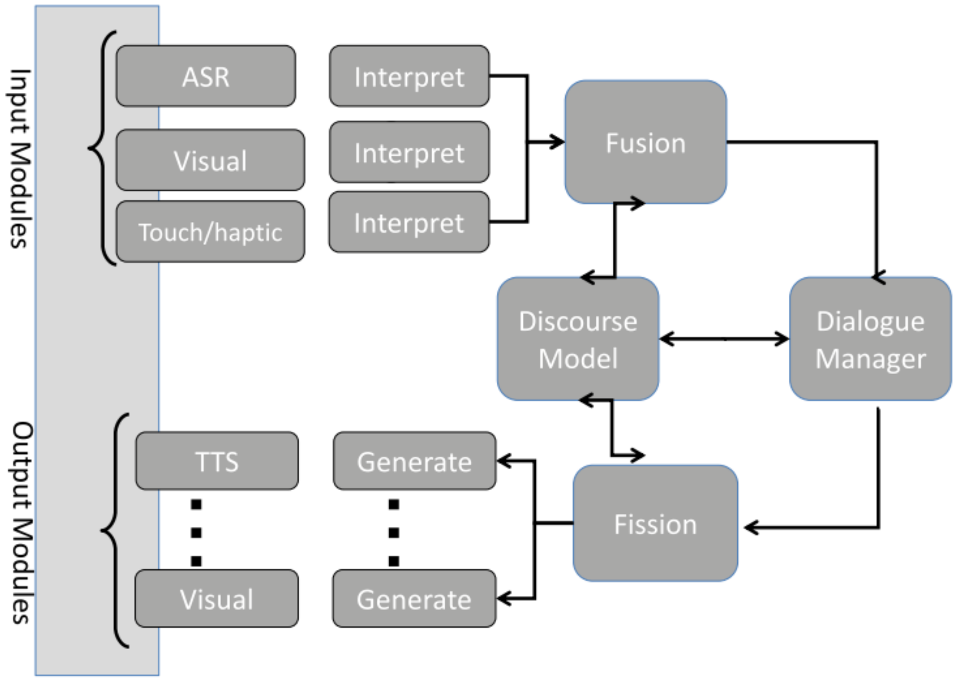
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Multimodal Dialogue Systems

- adding more modalities to voice/text
  - input:
    - touch
    - drawing
    - gaze, gestures, facial expressions
    - voice pitch/tone
    - image
  - output:
    - graphics
    - gaze, gestures, facial expressions, body movement
- either traditional/modular and mostly rule-based systems, or very experimental (not much use in practice)

Standard Multimodal DS Schema

- basically the same as voice/text DSs
- adding multiple input modules
  - for multiple modalities
  - each with its own NLU-like interpretation
  - interpretations are merged
- multiple output modules
  - each with its own generation
  - dialogue manager output is split
- typically ready-made off-the-shelf modules
  - it’s too complex/costly to build these custom



https://www.lsv.uni-saarland.de/multimodal-dialogue-systems-summer-2019/

Smart Devices

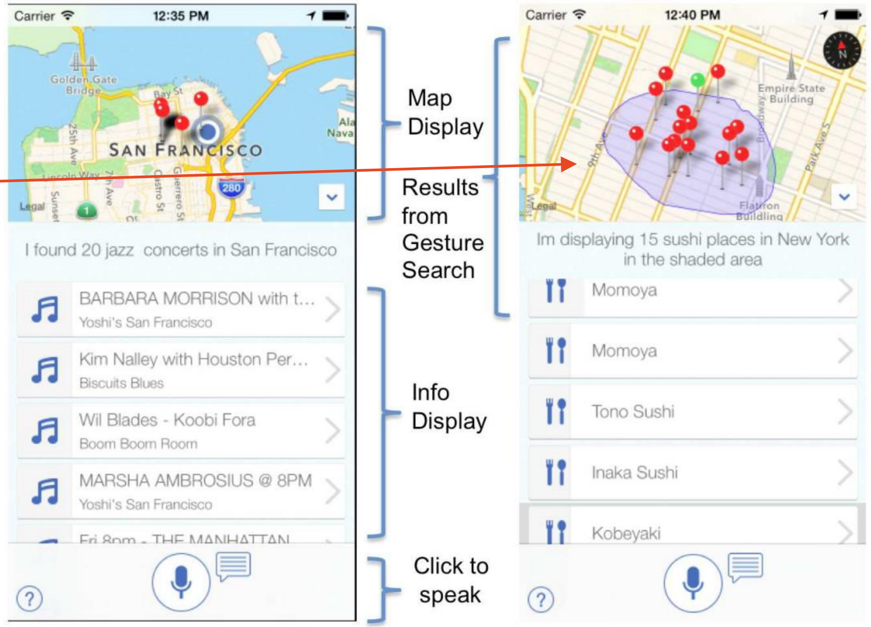
- Phones, wearables, smart speakers with a display
  - incl. Google Assistant, Alexa & Siri
  - admittedly not so much dialogue, more of commands
  - cloud-based operation for most
- Input
  - touch: active & passive gestures (touch/accelerometer)
    - "raise to speak"
    - rarely visually sensing gestures
    - doesn't support gaze
- Output
  - graphics: card interface
  - generation functions rule-based/low-level

https://www.wareable.com/android-wear/how-to-use-voice-commands-on-android-wear
https://www.cnet.com/reviews/amazon-echo-spot-review/
https://www.macrumors.com/how-to/use-siri-raise-to-speak-watchos-5/

“Classical” Multimodal Systems

- closed-domain task-oriented dialogue systems
- map-based: town information with map input & output
  - touch / pen – drawing, map display
    - reacting to zooming, area selection
    - handwriting recognition (as alternative input)
  - similar to Google Assistant, but more interactive
- in-car: voice & button control
- custom architectures
  - off-the-shelf modules
  - rule-based touch input processing

what about here?



I found 20 jazz concerts in San Francisco

BARBARA MORRISON with t…
Yoshi's San Francisco

Kim Nalley with Houston Per…
Biscuits Blues

Wil Blades - Koobi Fora
Boom Boom Room

MARSHA AMBROSIUS @ 8PM
Yoshi's San Francisco

Fri 8pm - THE MANHATTAN

Im displaying 15 sushi places in New York in the shaded area

Momoya

Momoya

Tono Sushi

Inaka Sushi

Kobeyaki

S: *I found 3 albums by The Beatles in your collection*
<shows listing on screen>
U: *Play the third one.*
U: *Which songs are on this one?*
<selects an album from listing on screen>

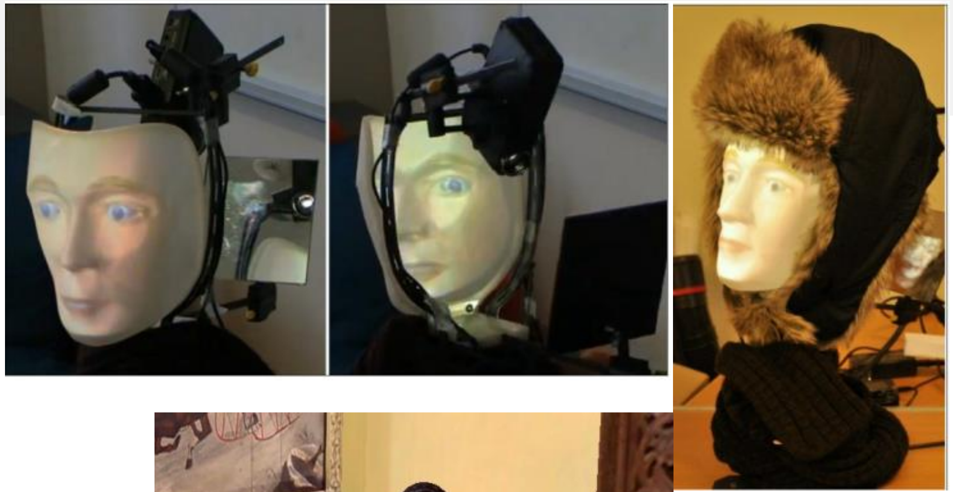
(Johnston et al., 2002) https://www.aclweb.org/anthology/P02-1048/
(Johnston et al., 2014) https://www.aclweb.org/anthology/W14-4335
(Becker et al., 2006) https://www.aclweb.org/anthology/P06-4015

Virtual Agents

https://youtu.be/ejczMs6b1Q4
https://vhtoolkit.ict.usc.edu/

- character face/full body
  - on screen or 3D projected (FurHat)
- a lot more outputs
  - full motion video – facial expressions, gaze, gestures, body movement
    - a lot of it "automatic", designed to look natural/match what's said
- additional inputs – gaze & facial expression
  - checking user engagement/sentiment
- dialogue management mostly rule-based
  - retrieval with non-linguistic inputs (Virtual Humans/SimSensei)
  - limited-domain custom rules (FurHat)
- tutoring/training, healthcare

FurHat

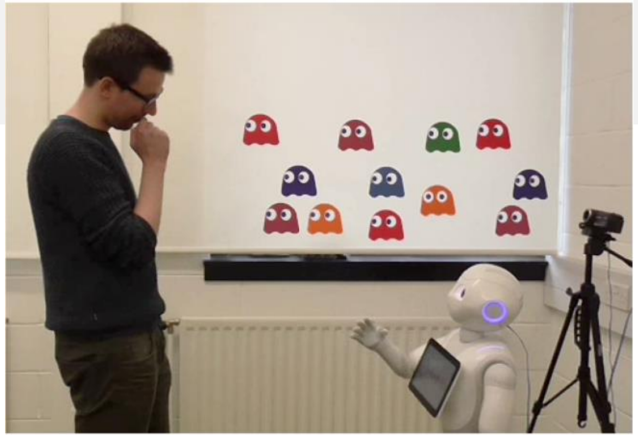


Virtual Humans

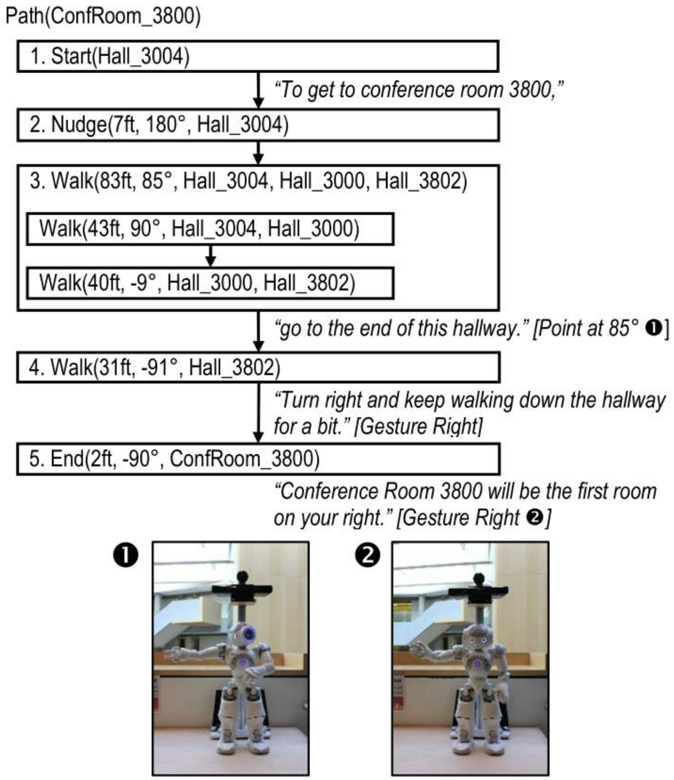
SimSensei

(Al Moubayed et al., 2012) https://doi.org/10.1007/978-3-642-34584-5_9
(Rushforth et al., 2009) https://doi.org/10.1007/978-3-642-04380-2_82
(DeVault et al., 2014) https://dl.acm.org/doi/10.5555/2615731.2617415

Robots



- similar to virtual agents, but with actual hardware
  - different user's perception
    - body gestures more prominent, touch is possible
  - situated deployment – need to track user engagement
    - is the user still talking to the robot?
  - hardware limitations
    - mostly no facial expr./gaze output, some sensors missing etc.
- off-the-shelf robots (Nao, Pepper, QTRobot)
  - built-in & additional sensors (e.g. Kinect)
  - custom rule-based gesture generation
  - remote-controlled (Nao, Pepper) / autonomous (QTRobot)
- "receptionist", education…



Path(ConfRoom_3800)

1. Start(Hall_3004)

"To get to conference room 3800,"

2. Nudge(7ft, 180°, Hall_3004)

3. Walk(83ft, 85°, Hall_3004, Hall_3000, Hall_3802)

Walk(43ft, 90°, Hall_3004, Hall_3000)

Walk(40ft, -9°, Hall_3000, Hall_3802)

"go to the end of this hallway." [Point at 85° ❶]

4. Walk(31ft, -91°, Hall_3802)

"Turn right and keep walking down the hallway for a bit." [Gesture Right]

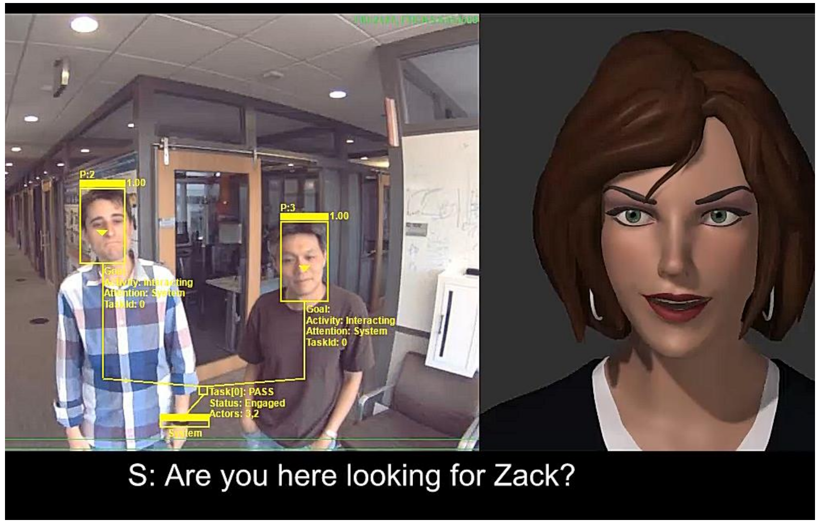
5. End(2ft, -90°, ConfRoom_3800)

"Conference Room 3800 will be the first room on your right." [Gesture Right ❷]

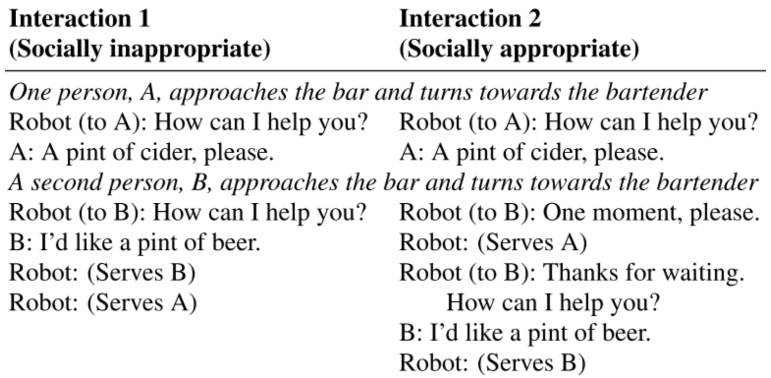
(Novikova et al., 2017) http://arxiv.org/abs/1706.02757
(Bohus et al., 2014) https://dl.acm.org/doi/10.5555/2615731.2615835
https://luxai.com/humanoid-social-robot-for-research-and-teaching/

Multi-party Dialogue

- Relevant for both virtual agents & robots
  - supported by most previously mentioned projects
- How to handle multiple counterparts?
  - users or other robots/virtual agents
- gaze/engagement/speech detection
  - who’s speaking/looking etc.
- rules for multiple counterparts
  - switching gaze to address them
    - here, 3D is better than 2D (otherwise gaze ambiguous)
  - telling one to wait for another
- customer service, information



S: Are you here looking for Zack?



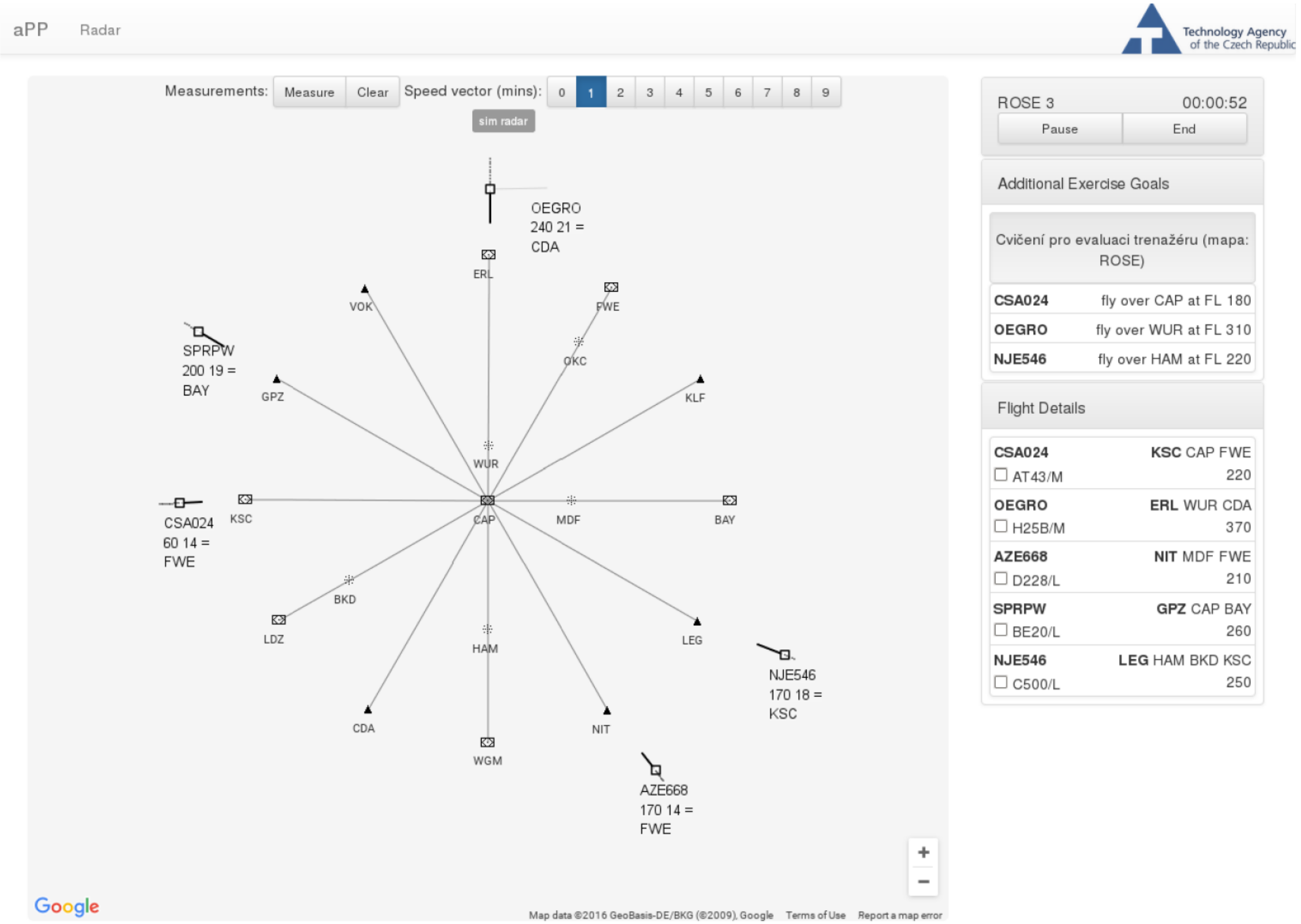
Interaction 1 (Socially inappropriate)	Interaction 2 (Socially appropriate)
One person, A, approaches the bar and turns towards the bartender	
Robot (to A): How can I help you?	Robot (to A): How can I help you?
A: A pint of cider, please.	A: A pint of cider, please.
A second person, B, approaches the bar and turns towards the bartender	
Robot (to B): How can I help you?	Robot (to B): One moment, please.
B: I’d like a pint of beer.	Robot: (Serves A)
Robot: (Serves B)	Robot (to B): Thanks for waiting. How can I help you?
Robot: (Serves A)	B: I’d like a pint of beer.
	Robot: (Serves B)

https://youtu.be/oOp4XP_ziMw
http://www.danbohus.com/

(Foster et al., 2012) http://dl.acm.org/citation.cfm?doid=2388676.2388680
(Bohus et al., 2014) https://dl.acm.org/doi/10.5555/2615731.2615835
(Skantze & Al Moubayed, 2012) https://doi.org/10.1145/2388676.2388698

Specific uses

- Air traffic controller training – radar as a modality
  - multiple agents/systems representing pilots
  - radar charting each agent's behavior
  - single ASR, many TTSs
    - varied accents
  - all rule-based
    - very limited domain
    - bearings, flight levels



ROSE 3 00:00:52

Additional Exercise Goals

Cvičení pro evaluaci trenažéru (mapa: ROSE)

CSA024	fly over CAP at FL 180
OEGRO	fly over WUR at FL 310
NJE546	fly over HAM at FL 220

Flight Details

CSA024 □ AT43/M	**KSC** CAP FWE 220
OEGRO □ H25B/M	**ERL** WUR CDA 370
AZE668 □ D228/L	**NIT** MDF FWE 210
SPRPW □ BE20/L	**GPZ** CAP BAY 260
NJE546 □ C500/L	**LEG** HAM BKD KSC 250

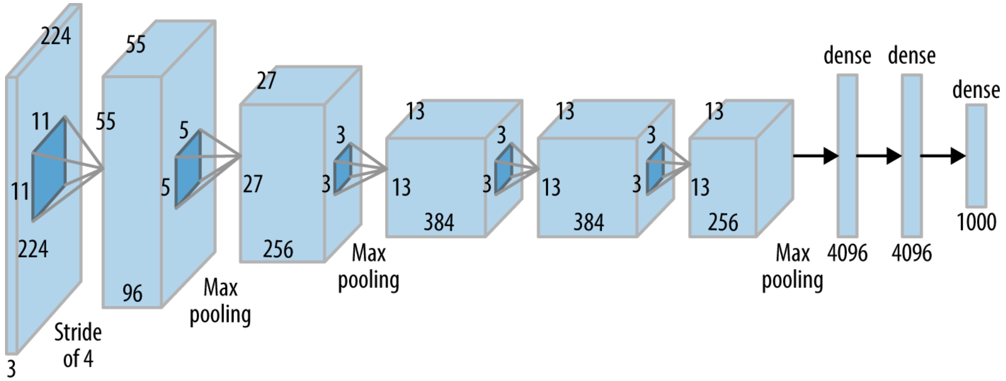
(Šmídl et al., 2016) https://www.isca-speech.org/archive/Interspeech_2016/abstracts/2002.html

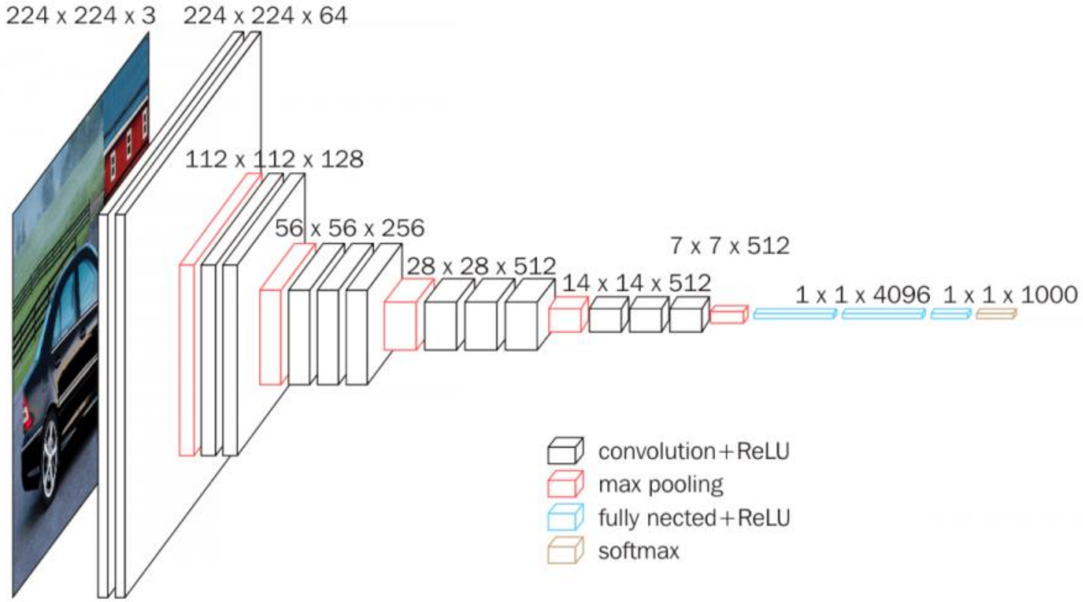
End-to-end Multimodal

- mostly experimental (save for latest LLMs)
- enhancing end-to-end DS architectures with image input
  - typically no video input
  - no avatars, facial expressions, gestures etc.
  - not much graphics output either
- also using off-the-shelf components
  - image input: ready-made convolutional/Transformer architectures
  - textual: known architectures (HRED, MemNN…) or pretrained Transformer models
- mostly just end-to-end prediction
  - pretrained image recognition parts are kept fixed, no end-to-end training

Image Classification

- Data: ImageNet Challenge
  - >1M images, 1000 classes
  - just classify the object in the image
  - CNNs are way better than anything that came before them
- **AlexNet** – 1st deep CNN
  - 5 conv layers, ReLU activations, max pooling & 3 dense layers
- **VGGNet** – improvement
  - more layers, smaller CNN kernels (3x3, 2x2 pooling with stride 2)
    - reduces # of parameters, same function

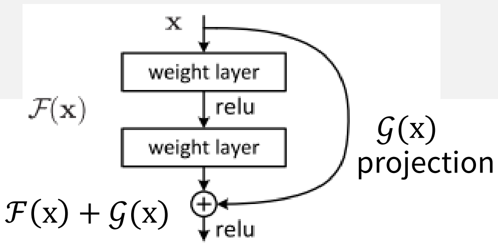


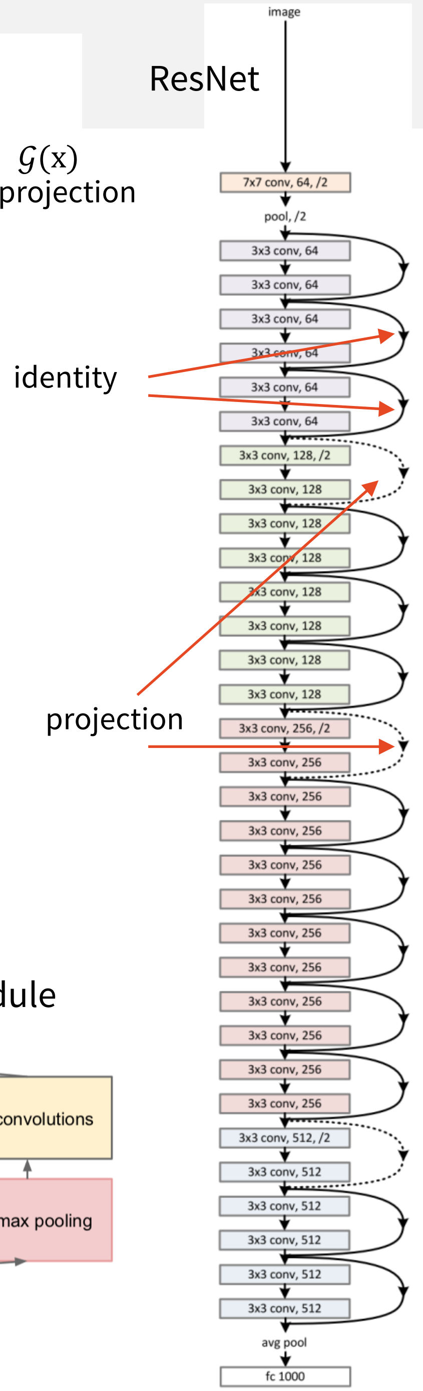


https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

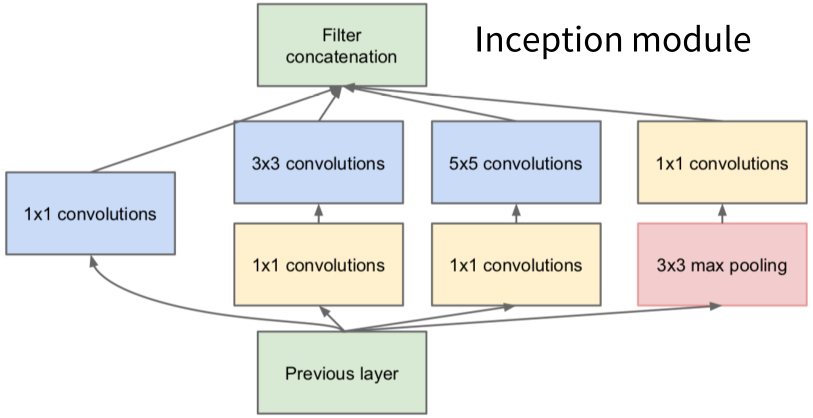
(Krizhevsky et al., 2012) https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
(Simonyan & Zisserman, 2015) http://arxiv.org/abs/1409.1556

Pretrained CNNs


x
weight layer
relu
weight layer
$\mathcal{F}(\mathrm{x})$
$\mathcal{G}(\mathrm{x})$
projection
$\mathcal{F}(\mathrm{x}) + \mathcal{G}(\mathrm{x})$
relu



- **ResNet** – residual networks
  - trying to simplify the mappings found by CNNs
    - with regular CNNs, deeper might not be better (vanishing gradient problem)
  - “shortcuts”: adding identity / linear projection to convolutions
    - learning a residual CNN mapping (“what projection can’t handle”)
  - allows much deeper networks – alleviates vanishing gradients
- **Inception** – more CNN kernels in parallel
  - for detecting different-sized object features
  - 1x1 depth reductions, depth-wise concatenations
  - better results with shallower networks



https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

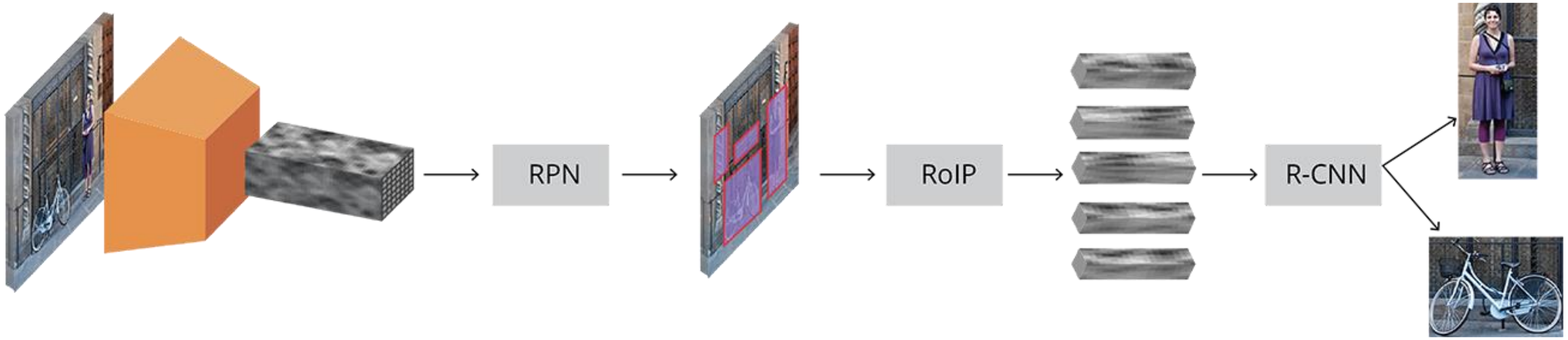
(He et al., 2016) https://arxiv.org/abs/1512.03385
(Szegedy et al., 2015) http://arxiv.org/abs/1409.4842

Pretrained CNNs

(Girschick, 2015) http://arxiv.org/abs/1504.08083
(Ren et al., 2015) https://arxiv.org/abs/1506.01497

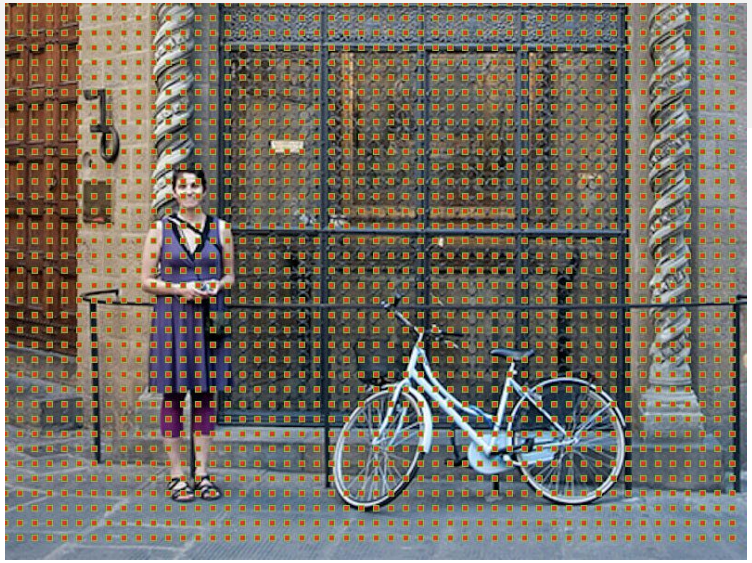
https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-rabbit-hole-of-modern-object-detection/

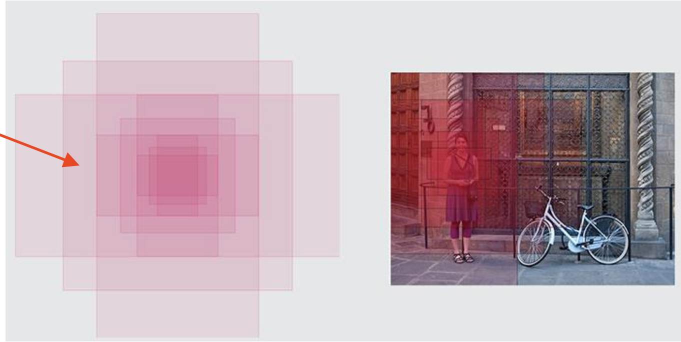
- **Faster R-CNN**
  - object detection – harder task
  - detecting boxes (regions) for multiple objects in image
- Pipeline:
  - Region prediction network (detect salient boxes)
  - Region-of-interest pooling (consolidate features)
  - Region-based CNN (classify)

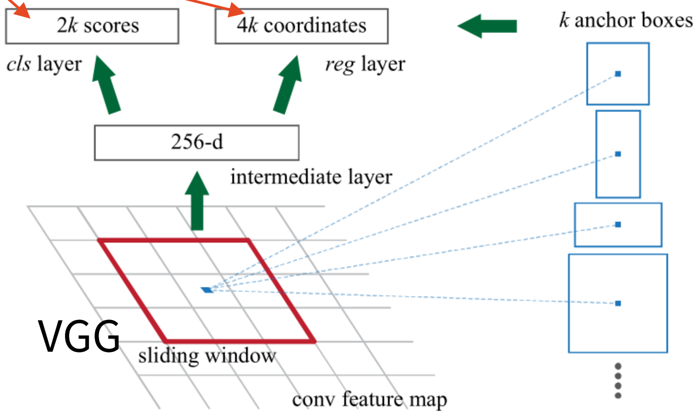


Region prediction

- pretrained VGG as feature extraction
  - features for each of the anchor points (regularly spaced in the image)
- for each anchor point, predict:
  - anchor base size & h/w ratio (e.g. 64-128-256px, 0.5/1/1.5)
  - p(this is object) & p(this is background)
  - anchor $\Delta x, \Delta y, \Delta h, \Delta w$
  - all of this via convolutions ☺
- trained using object/non-object anchors
- overlapping predictions unified







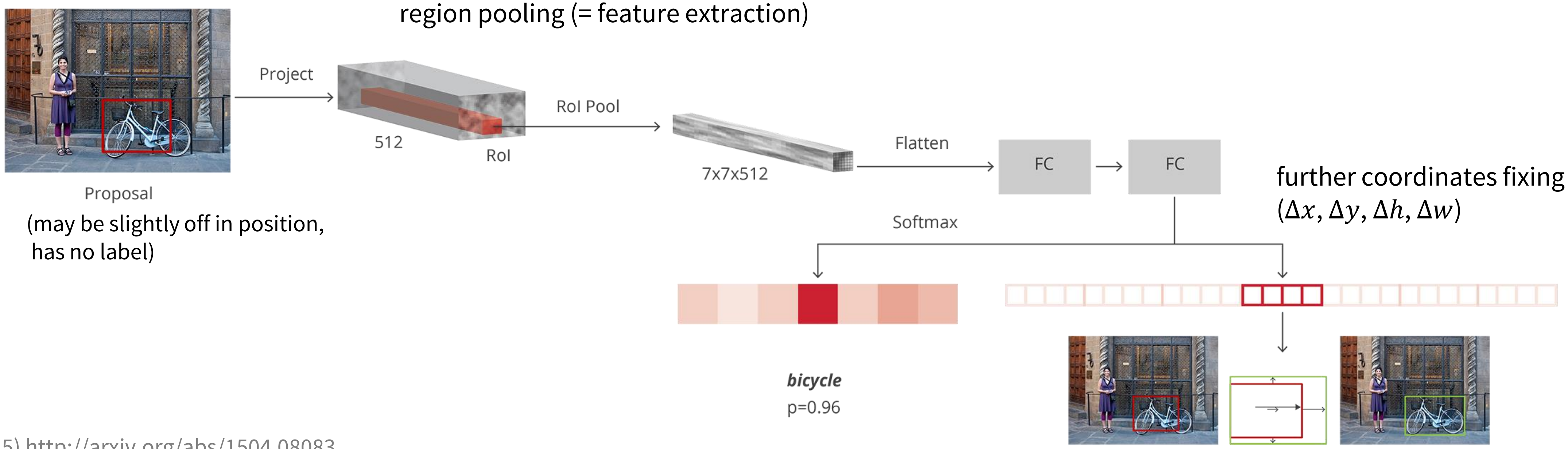
(Ren et al., 2015) https://arxiv.org/abs/1506.01497

https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-rabbit-hole-of-modern-object-detection/

R-CNN classification

- basically the same as image classification (given region)
  - with one more box coordinates fix
- sharing VGG features from RPN
  - this makes it much faster (only the pooling & prediction layers are new)

region pooling (= feature extraction)



(may be slightly off in position, has no label)

further coordinates fixing $(\Delta x, \Delta y, \Delta h, \Delta w)$

classification

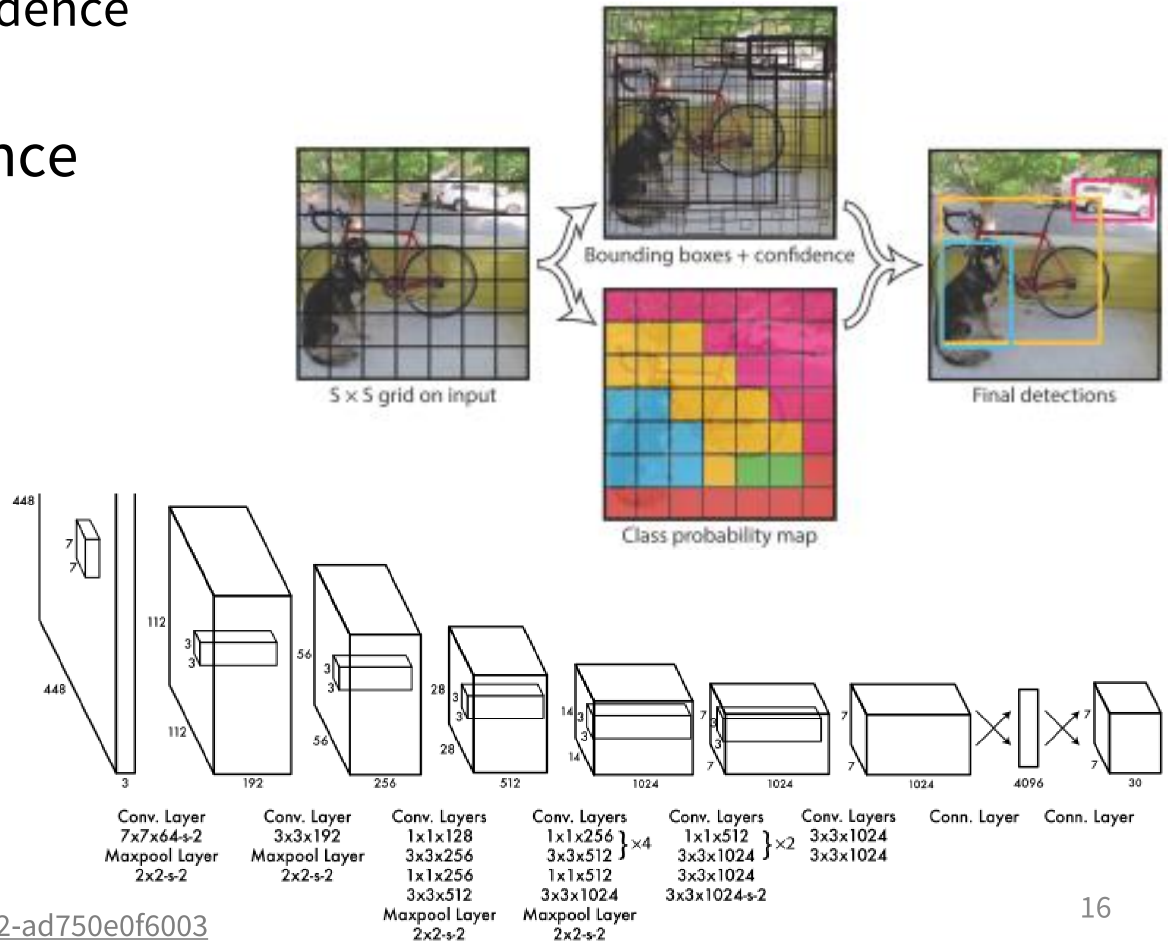
(Girschick, 2015) http://arxiv.org/abs/1504.08083

https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-rabbit-hole-of-modern-object-detection/

YOLO: Single-shot object detection

You Only Look Once

- Use CNN + fully connected grid to detect all objects in an image at once
  - Divide input to a grid
  - For each grid point (in parallel, via fully connected layer):
    - predict bounding boxes & object detection confidence
    - predict class probability
  - Threshold output objects on model confidence
- CNN processes whole image
  - unlike R-CNN which is local
- Works on real-time video
- Iteratively improved
  - multiple versions with better CNNs & other tricks



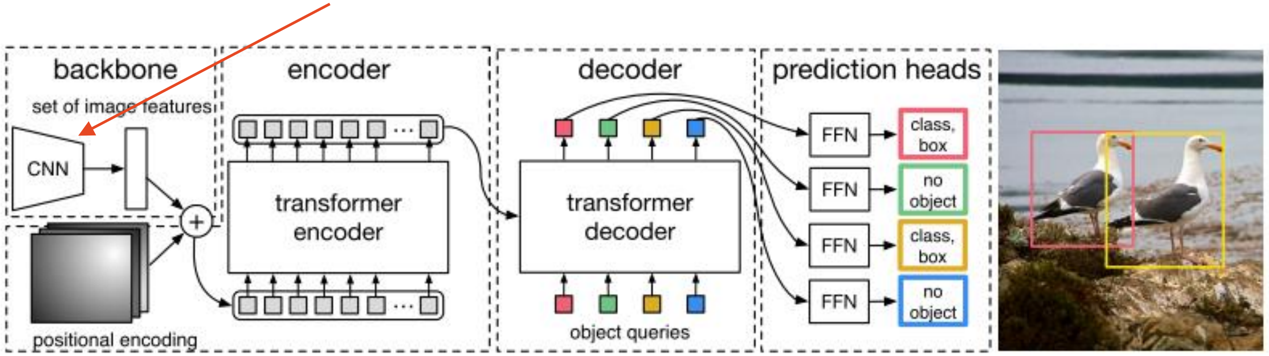
(Redmon et al., 2016) http://ieeexplore.ieee.org/document/7780460/
(Wang et al., 2022) http://arxiv.org/abs/2207.02696
https://www.v7labs.com/blog/yolo-object-detection
https://medium.com/@pedroazevedo6/object-detection-state-of-the-art-2022-ad750e0f6003

Using Transformers

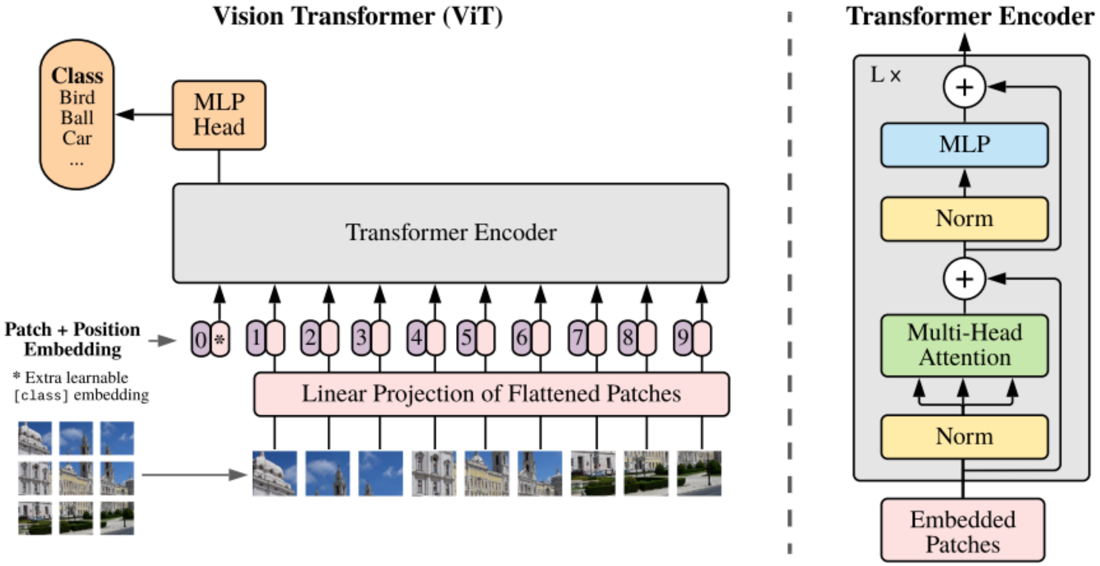
(Carion et al., 2020)
https://arxiv.org/abs/2005.12872

- Object detection: CNN + Transformer
  - trained end-to-end
  - predicts all objects at once
    - max. *k* objects: class + box (or “null”)
  - set-based loss (any order allowed)

ResNet

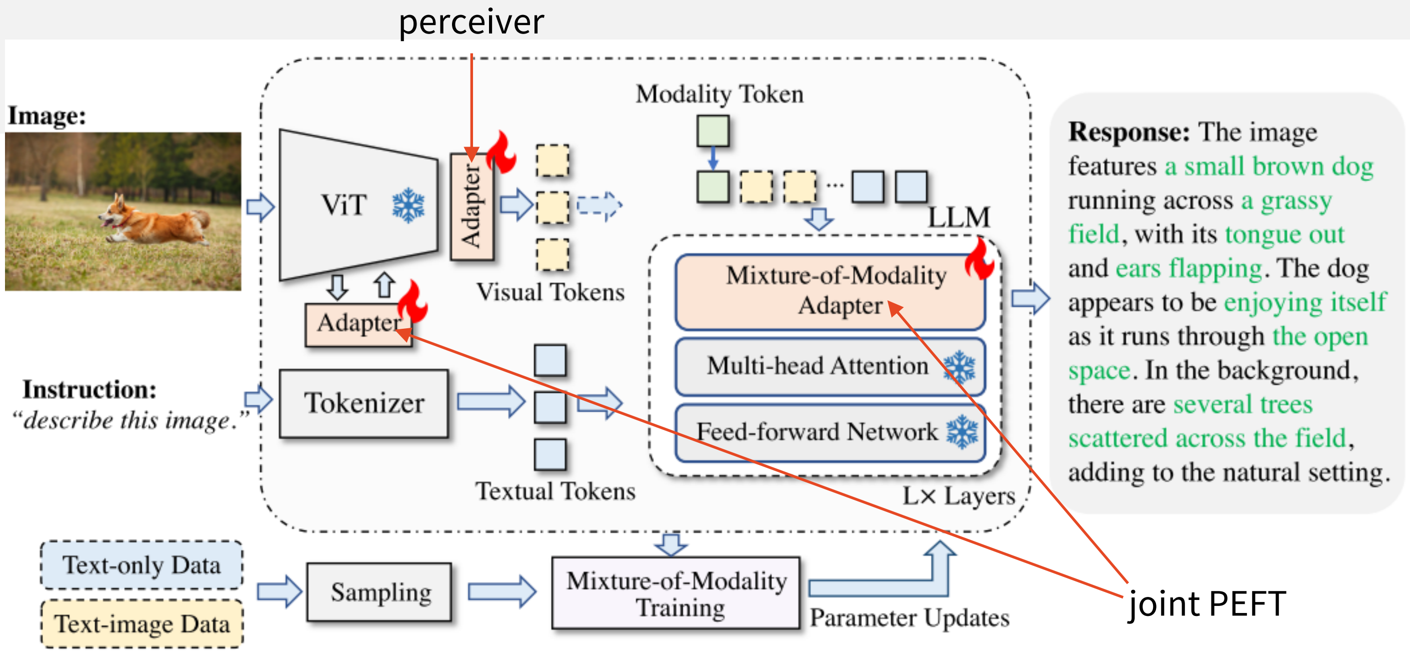


- Classification: Transformer-only (**Vision Transformer**)
  - split image into 16x16 patches
    - flatten (256xRGB=768-dim)
    - pass through a single “embedding” matrix
    - feed into standard transformer
  - adding positional embeddings
  - special “start” token
    - attention here used in a feed-forward classifier



(Dosovitskiy et al., 2020)
http://arxiv.org/abs/2010.11929 https://youtu.be/TrdevFK_am4

Multimodal LLMs

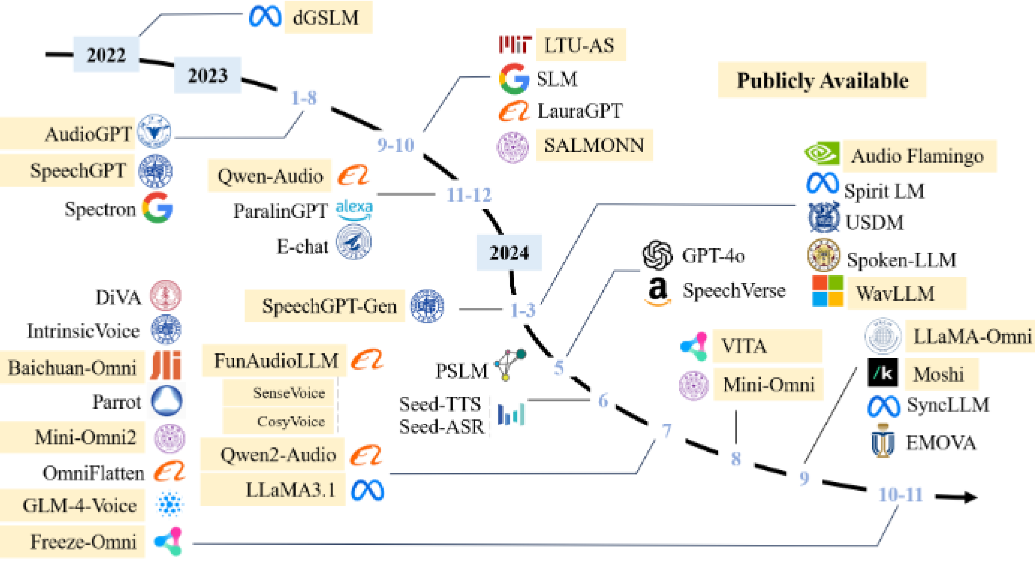


Response: The image features a small brown dog running across a grassy field, with its tongue out and ears flapping. The dog appears to be enjoying itself as it runs through the open space. In the background, there are several trees scattered across the field, adding to the natural setting.

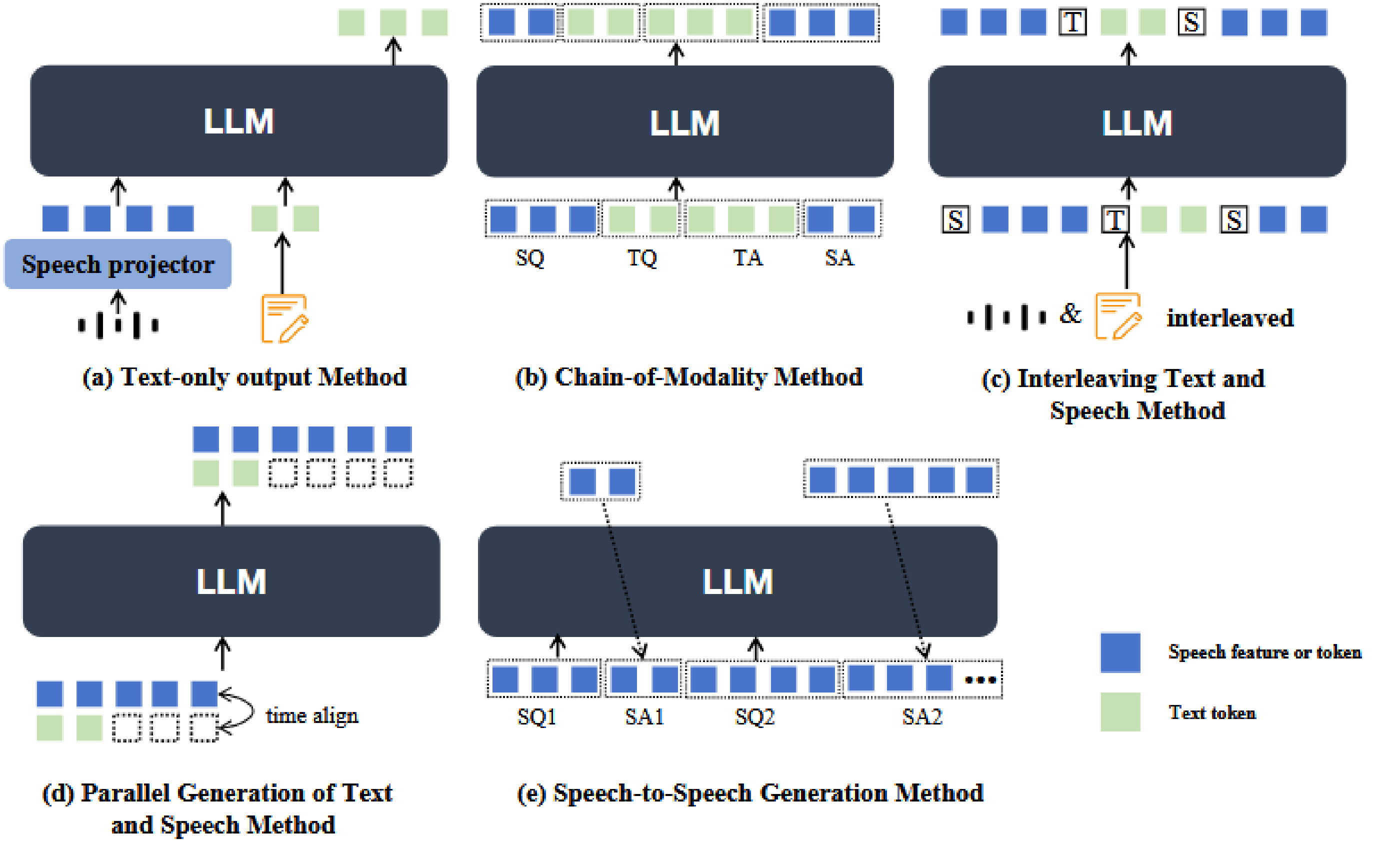
- Text-based LLM
  + image/video/audio encoders
  - mostly based on existing architectures
    (e.g. Vision Transformer + Llama)
- Aligning representations
  - interfacing – **perceiver modules**: produce sequences of trainable tokens
    - i.e. expanding LLM vocabulary for image/audio tokens
  - alternative: ("expert model") translate to text & use standard LLM
- Instruction tuning on multimodal tasks
  - e.g. image description, image QA
  - 2-step (training adapter + instruction tuning) vs. joint parameter-efficient

(Yin et al., 2023) http://arxiv.org/abs/2306.13549
(Luo et al., 2023) http://arxiv.org/abs/2305.15023
(Zhao et al., 2023) http://arxiv.org/abs/2305.16103
(Li et al., 2023) http://arxiv.org/abs/2305.03726

Speech End-to-End LLMs



- Very similar, work with audio tokens
- Multiple paradigms, different capabilities
  - voice input only (+external TTS) / input & output
  - internal use of text ?
- Aim: full duplex
  - = listen & speak at the same time
  - no explicit turns & waiting
- Aim: minimal latency
- Some are video + speech + text
- Work but worse than text LLMs



(a) Text-only output Method

(b) Chain-of-Modality Method

(c) Interleaving Text and Speech Method

(d) Parallel Generation of Text and Speech Method

(e) Speech-to-Speech Generation Method

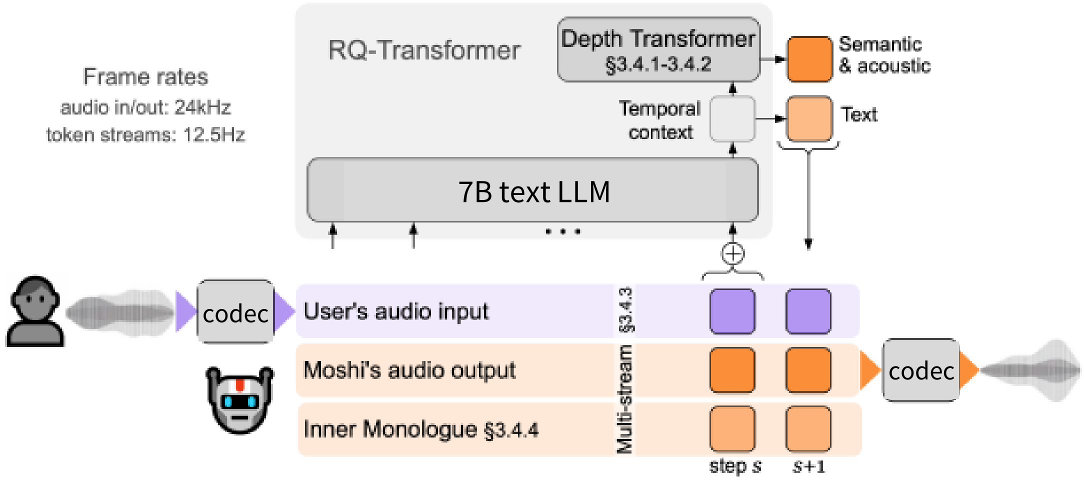
SQ = speech question
SA = speech answer
TQ = text question
TA = text answer

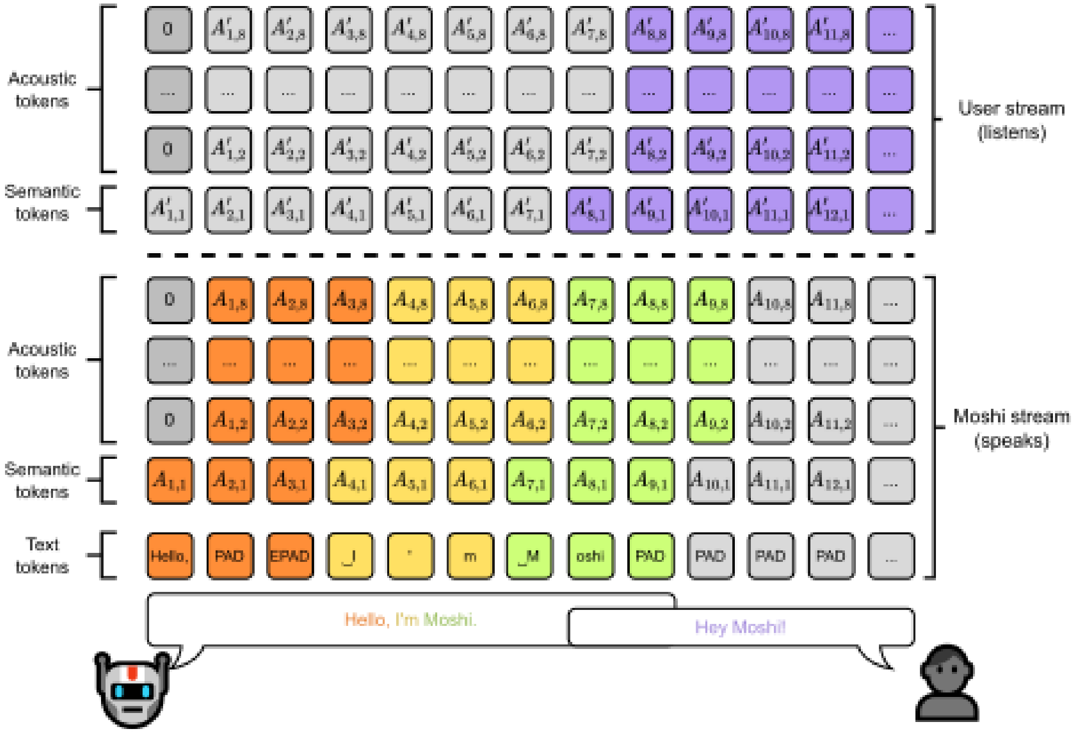
(Ji et al., 2024)
https://arxiv.org/abs/2411.13577

Speech LLM example: Moshi

(Defossez et al., 2024)
https://arxiv.org/abs/2410.00037

- Parallel speech & text, full duplex (200ms latency)
- Text LLM + audio codec network distilled from pretrained ASR Model
    - 2.1T tokens of text
    - 7M hrs of audio transcribed with Whisper
    - 20k synthetic speech instruction data (TTS, consistent voice)





Acoustic tokens	0	$A'_{1,8}$	$A'_{2,8}$	$A'_{3,8}$	$A'_{4,8}$	$A'_{5,8}$	$A'_{6,8}$	$A'_{7,8}$	$A'_{8,8}$	$A'_{9,8}$	$A'_{10,8}$	$A'_{11,8}$	...	User stream (listens)
	...	...	...	...	...	...	...	...	...	...	...	...	...	
	0	$A'_{1,2}$	$A'_{2,2}$	$A'_{3,2}$	$A'_{4,2}$	$A'_{5,2}$	$A'_{6,2}$	$A'_{7,2}$	$A'_{8,2}$	$A'_{9,2}$	$A'_{10,2}$	$A'_{11,2}$	...	
Semantic tokens	$A'_{1,1}$	$A'_{2,1}$	$A'_{3,1}$	$A'_{4,1}$	$A'_{5,1}$	$A'_{6,1}$	$A'_{7,1}$	$A'_{8,1}$	$A'_{9,1}$	$A'_{10,1}$	$A'_{11,1}$	$A'_{12,1}$	...	
Acoustic tokens	0	$A_{1,8}$	$A_{2,8}$	$A_{3,8}$	$A_{4,8}$	$A_{5,8}$	$A_{6,8}$	$A_{7,8}$	$A_{8,8}$	$A_{9,8}$	$A_{10,8}$	$A_{11,8}$	...	Moshi stream (speaks)
	...	...	...	...	...	...	...	...	...	...	...	...	...	
	0	$A_{1,2}$	$A_{2,2}$	$A_{3,2}$	$A_{4,2}$	$A_{5,2}$	$A_{6,2}$	$A_{7,2}$	$A_{8,2}$	$A_{9,2}$	$A_{10,2}$	$A_{11,2}$	...	
Semantic tokens	$A_{1,1}$	$A_{2,1}$	$A_{3,1}$	$A_{4,1}$	$A_{5,1}$	$A_{6,1}$	$A_{7,1}$	$A_{8,1}$	$A_{9,1}$	$A_{10,1}$	$A_{11,1}$	$A_{12,1}$	...	
Text tokens	Hello,	PAD	EPAD	_I	'	m	_M	oshi	PAD	PAD	PAD	PAD	...	

https://github.com/kyutai-labs/moshi

Multimodal Tasks: Visual Dialogue

https://visualdialog.org/

(Das et al., 2017) http://arxiv.org/abs/1611.08669

- **Task**: have a meaningful dialogue about an image
  - close to visual QA: **human asks**, **system responds**
  - but VD is multi-turn & human doesn't see the image (just a caption)
    - follow-up questions possible – coreference
    - people are not primed by the image when asking questions
- not much realistic purpose other than to test the models
  - dataset of 10-turn dialogues on 120k images
    - collected via crowdsourcing
      - connecting 2 people
        live to talk about an image
    - not very deep dialogue:
      history only needed in ~11% cases

(Agarwal et al., 2020)
https://www.aclweb.org/anthology/2020.acl-main.728

Caption: A man and woman on bicycles are looking at a map.
Person A (1): where are they located
Person B (1): in city
Person A (2): are they on road
Person B (2): sidewalk next to 1
Person A (3): any vehicles
Person B (3): 1 in background
Person A (4): any other people
Person B (4): no
Person A (5): what color bikes
Person B (5): 1 silver and 1 yellow
Person A (6): do they look old or new
Person B (6): new bikes
Person A (7): any buildings
Person B (7): yes
Person A (8): what color
Person B (8): brick
Person A (9): are they tall or short
Person B (9): i can't see enough of them to tell
Person A (10): do they look like couple
Person B (10): they are

Caption:
A sink and toilet in a small room.

Q3: can you see anything else ?
A3: there is a shelf with items on it
Q4: is anyone in the room ?
A4: nobody is in the room
Q5: can you see on the outside ?
A5: no, it is only inside
Q6: what color is the sink ?
A6: the sink is white
Q7: is the room clean ?
A7: it is very clean
Q8: is the toilet facing the sink ?
A8: yes the toilet is facing the sink
Q9: can you see a door ?
A9: yes, I can see the door
Q10 what color is the door ?
A10 the door is tan colored

Base Visual Dialogue Models

(Das et al., 2017) http://arxiv.org/abs/1611.08669

Late fusion

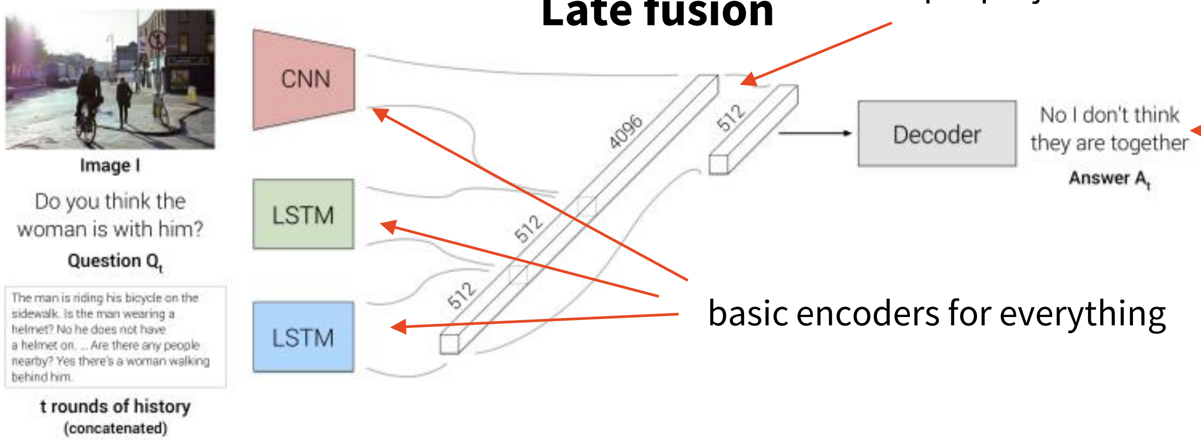


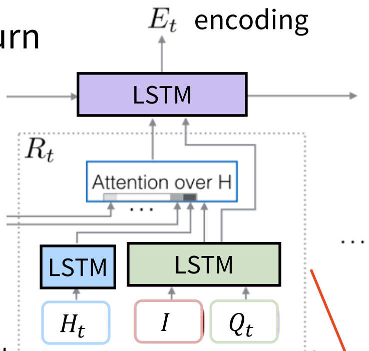
Image I
Do you think the woman is with him?
Question Q$_t$
The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on. ... Are there any people nearby? Yes there's a woman walking behind him.
t rounds of history
(concatenated)
CNN
LSTM
LSTM
4096
512
512
512
Decoder
No I don't think they are together
Answer A$_t$

simple projection to initialize decoder

LSTM decoder (same for all)

basic encoders for everything

each turn



encoding

all previous turns

preceding turn

CNN output

current user input

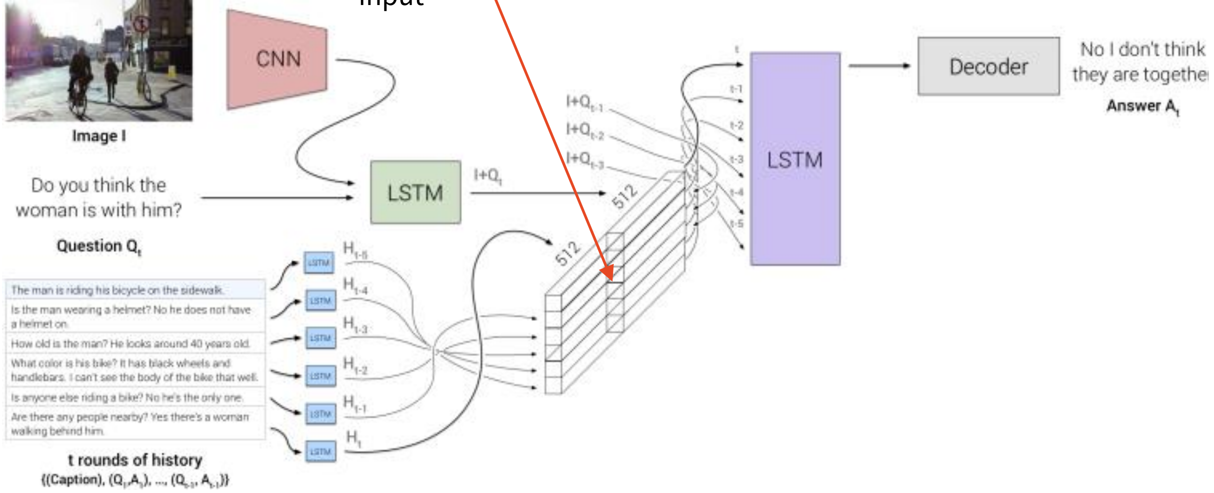


Image I
CNN
Do you think the woman is with him?
Question Q$_t$
LSTM
I+Q$_t$
I+Q$_{t-1}$
I+Q$_{t-2}$
I+Q$_{t-3}$
512
512
t
t-1
t-2
t-3
t-4
t-5
LSTM
Decoder
No I don't think they are together
Answer A$_t$
The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.
LSTM
H$_{t-5}$
H$_{t-4}$
H$_{t-3}$
H$_{t-2}$
H$_{t-1}$
H$_t$
t rounds of history
{(Caption), (Q$_1$,A$_1$), ..., (Q$_{t-1}$, A$_{t-1}$)}

Hierarchical Recurrent Encoder

VGG (same for all)

Memory Network

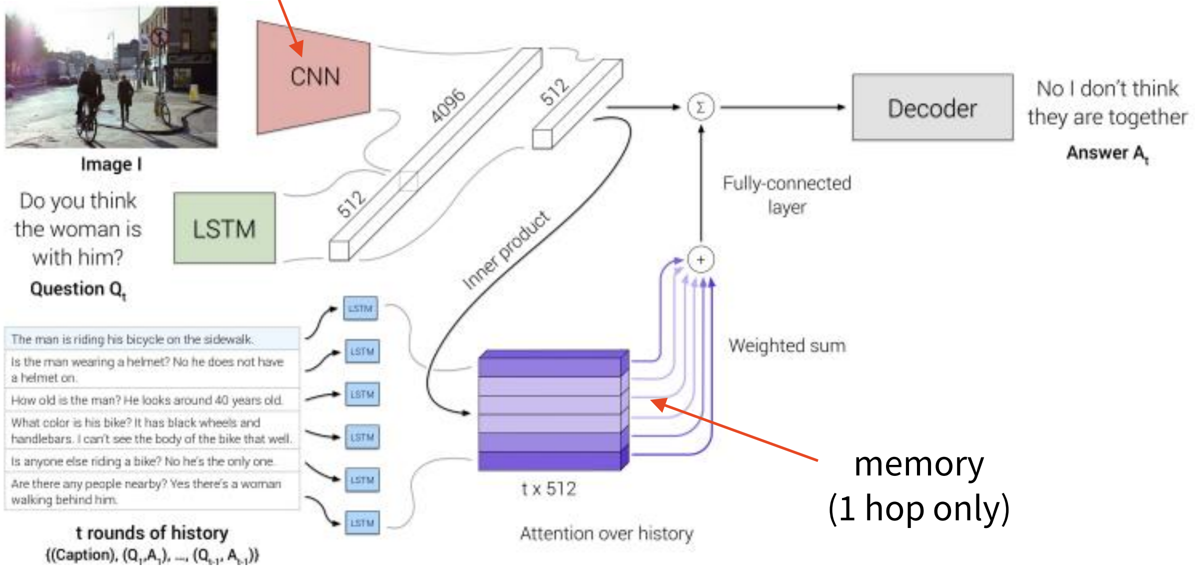


Image I
CNN
4096
512
512
Do you think the woman is with him?
Question Q$_t$
LSTM
Σ
Decoder
No I don't think they are together
Answer A$_t$
Fully-connected layer
Inner product
+
Weighted sum
The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.
LSTM
t rounds of history
{(Caption), (Q$_1$,A$_1$), ..., (Q$_{t-1}$, A$_{t-1}$)}
t x 512
Attention over history

memory
(1 hop only)

Visual Dialogue Evaluation

- BLEU etc. possible but not used here
- IR setup used instead
  - system given ground-truth dialogue history + user input & 100 candidate answers to score/rank
- IR metrics:
  - **ground-truth response rank** (average)
  - **recall@k** (% cases where ground-truth is included in top k)
  - **mean reciprocal rank**: $\emptyset \frac{1}{\text{ground truth rank}}$ (1 if ground truth is first, 0.5 if second etc.)
  - **normalized discounted cumulative gain**
    - for multiple acceptable answers out of the 100 candidates
    - DCG: $\sum_{i=1}^{100} \frac{c_i \text{ relevant?}}{\log_2(i+1)}$, normalize by highest possible DCG (all good answers on top)
- problem: images only give modest gain over text-only models

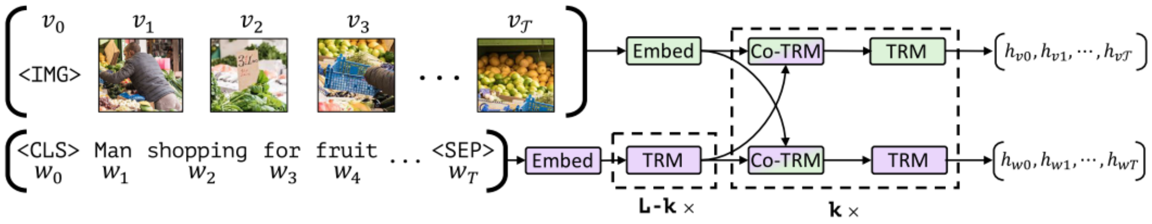
https://visualdialog.org/challenge/2019#evaluation

https://en.wikipedia.org/wiki/Discounted_cumulative_gain
https://medium.com/@_init_/notes-on-the-ndcg-metric-used-in-the-visual-dialog-challenge-2019-90cf443b93dc

Visual Dialogue

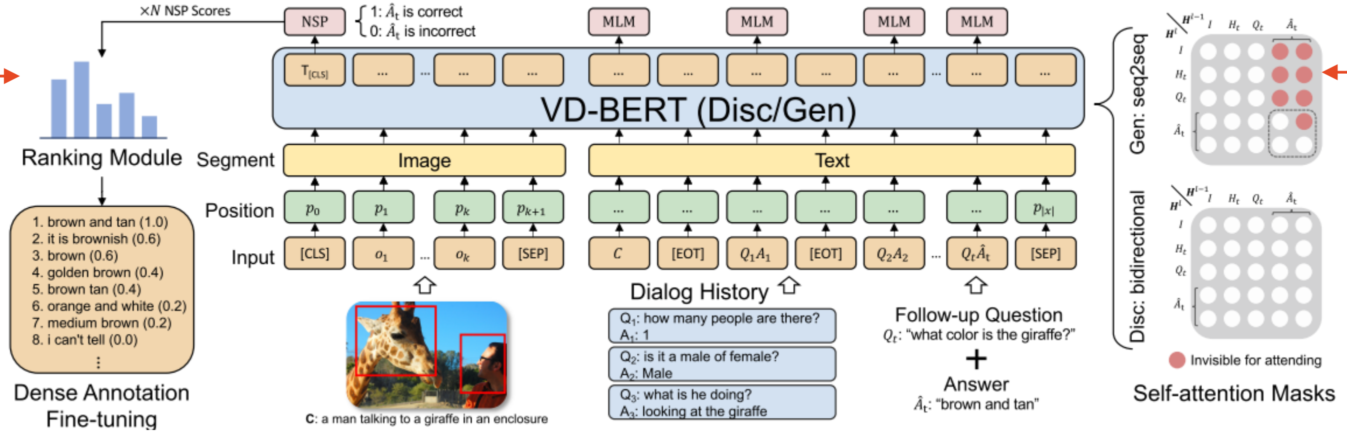
- Pretrained vision & language models
  - BERT as text features
  - R-CNN labels & positions as image features
  - cross-attention
  - additional Transformer decoder

(Lu et al., 2019) http://arxiv.org/abs/1908.02265



v_0 v_1 v_2 v_3 v_T
<IMG>
<CLS> Man shopping for fruit ... <SEP>
w_0 w_1 w_2 w_3 w_4 w_T
Embed
Embed
Co-TRM
TRM
TRM
Co-TRM
TRM
$(h_{v0}, h_{v1}, \cdots, h_{vT})$
$(h_{w0}, h_{w1}, \cdots, h_{wT})$
L-k ×
k ×

tasks: masked LM
generation
matching sentences



×N NSP Scores
NSP
1: $\hat{A}_t$ is correct
0: $\hat{A}_t$ is incorrect
MLM
MLM
MLM
MLM
$T_{[CLS]}$
VD-BERT (Disc/Gen)
Ranking Module
Segment
Image
Text
Position
p_0 p_1 p_k p_{k+1} $p_{|x|}$
Input
[CLS] o_1 o_k [SEP] C [EOT] Q_1A_1 [EOT] Q_2A_2 $Q_t\hat{A}_t$ [SEP]
1. brown and tan (1.0)
2. it is brownish (0.6)
3. brown (0.6)
4. golden brown (0.4)
5. brown tan (0.4)
6. orange and white (0.2)
7. medium brown (0.2)
8. i can't tell (0.0)
Dense Annotation Fine-tuning
C: a man talking to a giraffe in an enclosure
Dialog History
Q1: how many people are there?
A1: 1
Q2: is it a male of female?
A2: Male
Q3: what is he doing?
A3: looking at the giraffe
Follow-up Question
Q_t: "what color is the giraffe?"
+
Answer
$\hat{A}_t$: "brown and tan"
Gen: seq2seq
Disc: bidirectional
H^{l-1} I H_t Q_t $\hat{A}_t$
H^l
Invisible for attending
Self-attention Masks

generation (with decoder):
image visible always, text left-to-right

(Wang et al., 2020)
https://aclanthology.org/2020.emnlp-main.269/

- Option: synthetic data generation & retraining
  - retrieve additional images, generate dialogues, add to training data

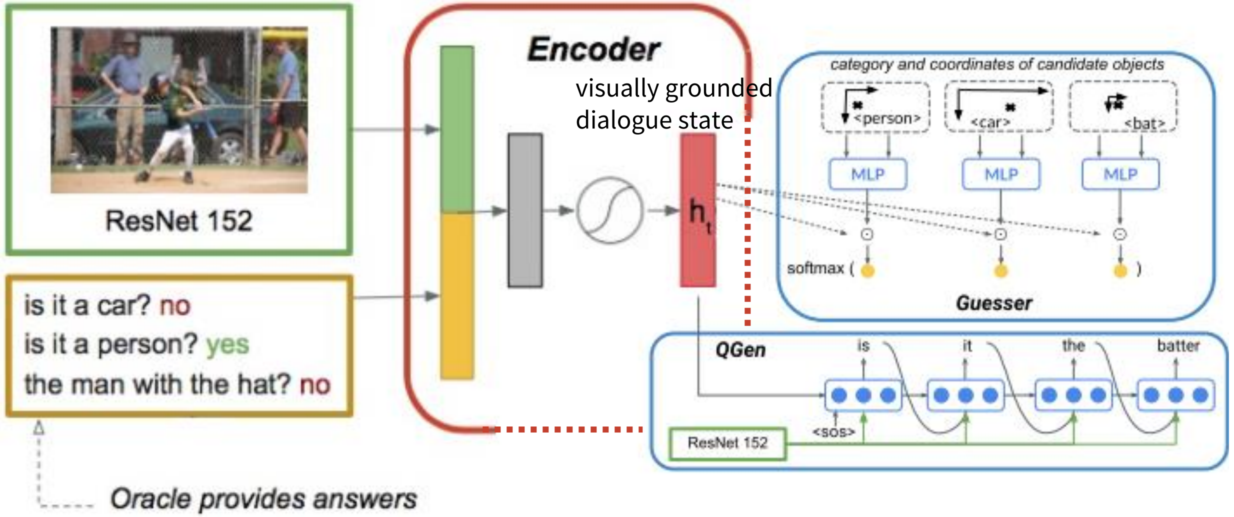
(Kang et al., 2023) https://ieeexplore.ieee.org/document/10203535

Guess What

(Strub et al., 2017) https://www.ijcai.org/proceedings/2017/385
(Shekhar et al., 2019) https://aclanthology.org/N19-1265

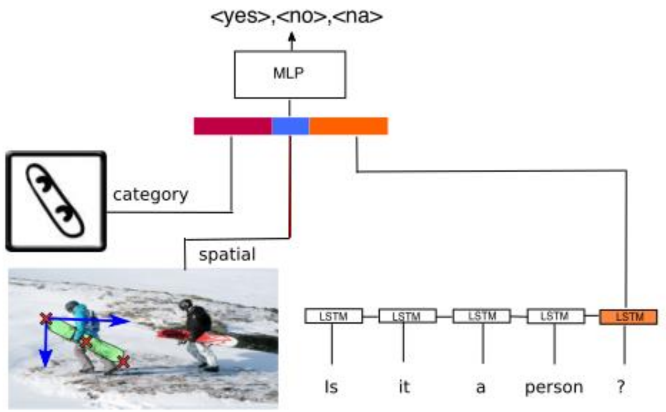
Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

- guessing one of the objects in an image
  - GuessWhat data (150k guessing dialogues)
- 3 models:
  - question generator – seq2seq decoder
    - encoder: fusion of image & text encodings
  - guesser – classifier:
    softmax from candidates
    - uses the same encoder as Q-gen
  - oracle – reply classifier (Y/N/NA)
    - seq2seq question encoding
      + object category + object size & position in image
- (potentially) trained jointly with RL



is it a car? no
is it a person? yes
the man with the hat? no

visually grounded dialogue state

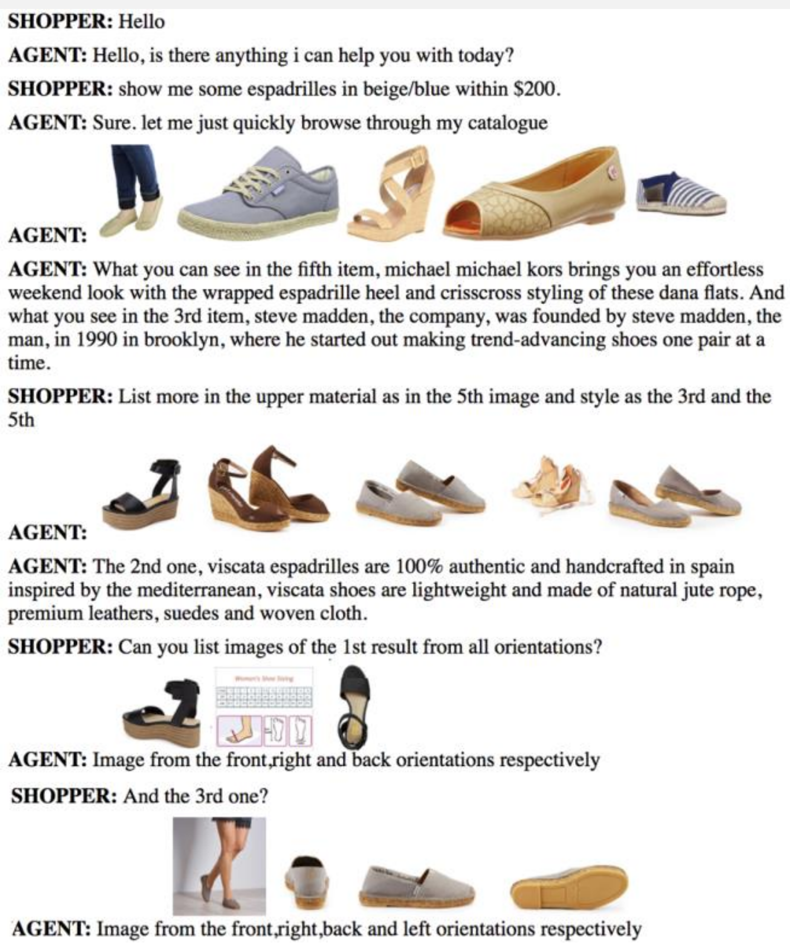


extensions: (Suglia et al., 2020) https://www.aclweb.org/anthology/2020.coling-main.95
(Chaudhary et al., 2022) https://aclanthology.org/2022.inlg-main.24/

Shopping Dialogues

(Saha et al., 2018)
http://arxiv.org/abs/1704.00200

- a more practical task: fashion e-shop
  - both user & system can show images
  - user can reference/talk about images
- semi-automatic data (150k dialogues)
  - 84 predefined intents
  - generated dialogue flows
    - based on predefined state automata over intents
  - utterances for each intent written by annotators
  - backed by a large manually curated DB
    of fashion items with images (1M items)
    - taxonomy (man > apparel > lower body > trousers)
    - (generated) celebrity profiles/endorsements



SHOPPER: Hello

AGENT: Hello, is there anything i can help you with today?

SHOPPER: show me some espadrilles in beige/blue within $200.

AGENT: Sure. let me just quickly browse through my catalogue

AGENT:

AGENT: What you can see in the fifth item, michael michael kors brings you an effortless weekend look with the wrapped espadrille heel and crisscross styling of these dana flats. And what you see in the 3rd item, steve madden, the company, was founded by steve madden, the man, in 1990 in brooklyn, where he started out making trend-advancing shoes one pair at a time.

SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and the 5th

AGENT:

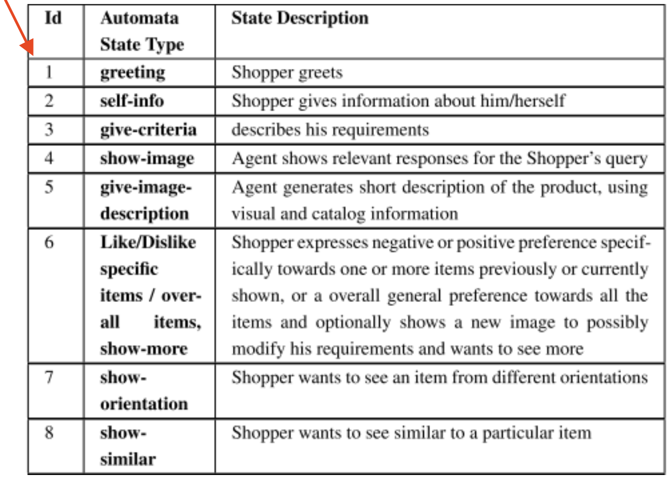
AGENT: The 2nd one, viscata espadrilles are 100% authentic and handcrafted in spain inspired by the mediterranean, viscata shoes are lightweight and made of natural jute rope, premium leathers, suedes and woven cloth.

SHOPPER: Can you list images of the 1st result from all orientations?

AGENT: Image from the front,right and back orientations respectively

SHOPPER: And the 3rd one?

AGENT: Image from the front,right,back and left orientations respectively

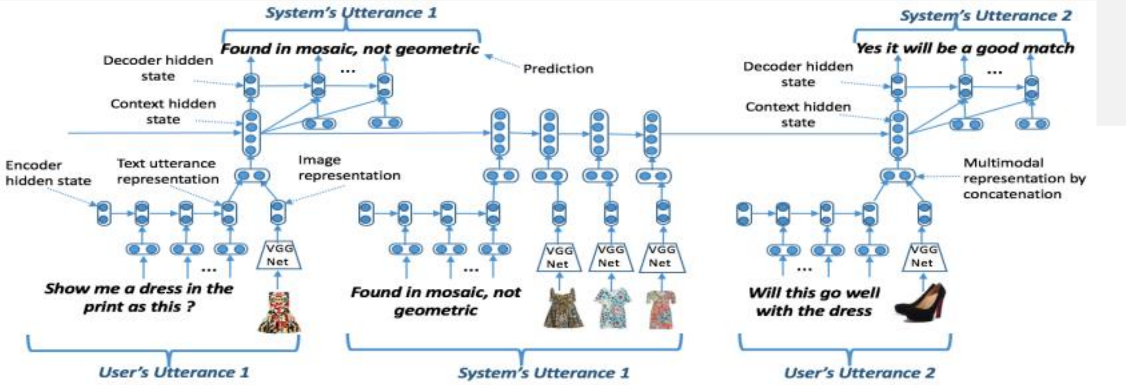


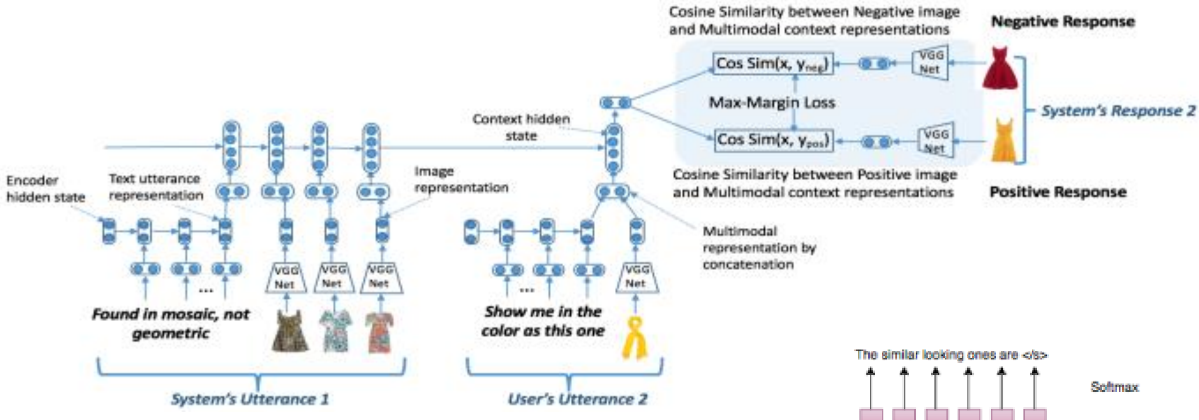
Id	Automata State Type	State Description
1	**greeting**	Shopper greets
2	**self-info**	Shopper gives information about him/herself
3	**give-criteria**	describes his requirements
4	**show-image**	Agent shows relevant responses for the Shopper's query
5	**give-image-description**	Agent generates short description of the product, using visual and catalog information
6	**Like/Dislike specific items / overall items, show-more**	Shopper expresses negative or positive preference specifically towards one or more items previously or currently shown, or a overall general preference towards all the items and optionally shows a new image to possibly modify his requirements and wants to see more
7	**show-orientation**	Shopper wants to see an item from different orientations
8	**show-similar**	Shopper wants to see similar to a particular item

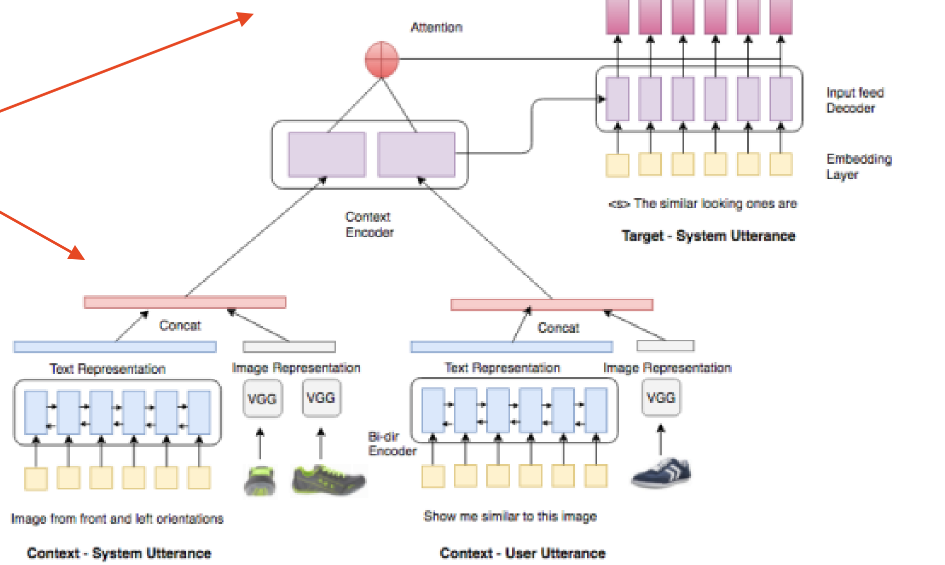
Shopping Dialogues

- Models similar to visual dialogue
  - variants of multimodal HRED
  - VGG image input
- image input
  - turn-level
  - concatenated with utterance
    - seems to work better (fewer turns)
- text/image responses
  - shared encoder
  - text generation (word-by-word)
  - image ranking (needs rough retrieval)
    - so far just "select 1 out of 5"

(Saha et al., 2018) http://arxiv.org/abs/1704.00200
(Agarwal et al., 2018) http://aclweb.org/anthology/W18-6514

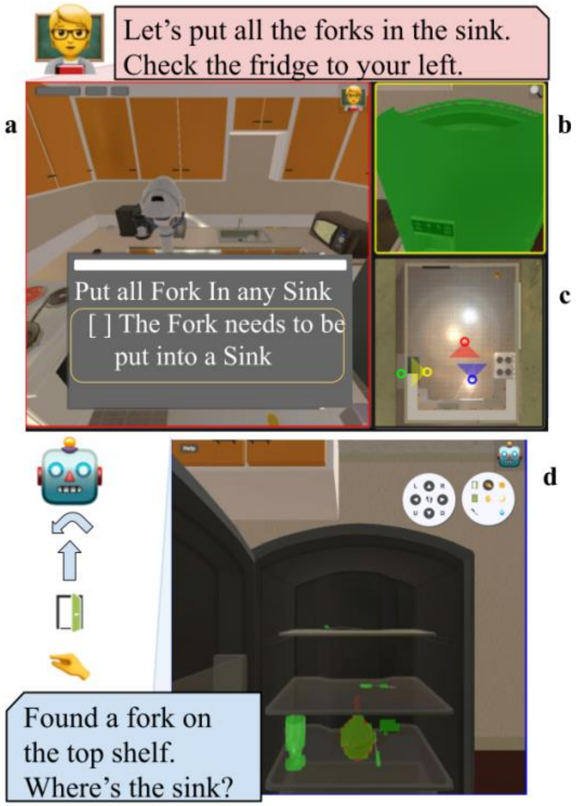


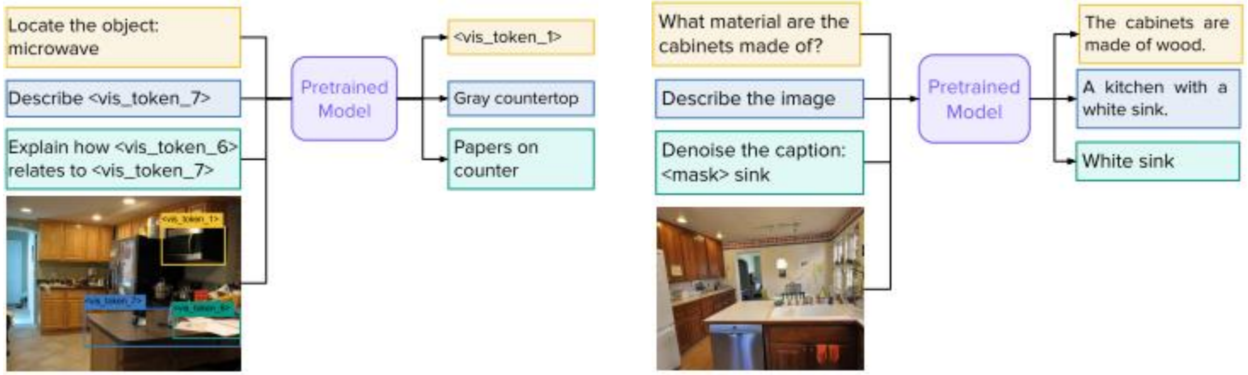

Cosine Similarity between Negative image and Multimodal context representations
Negative Response
Cos Sim(x, y_{neg})
VGG Net
Max-Margin Loss
Cos Sim(x, y_{pos})
System's Response 2
Cosine Similarity between Positive image and Multimodal context representations
Positive Response
Context hidden state
Encoder hidden state
Text utterance representation
Image representation
Multimodal representation by concatenation
Found in mosaic, not geometric
Show me in the color as this one
System's Utterance 1
User's Utterance 2

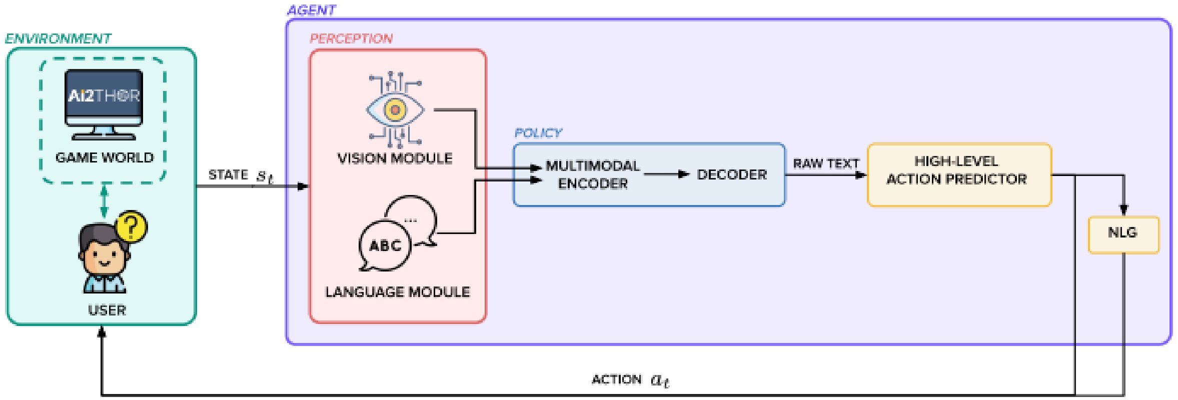


Situated Tasks (Amazon SimBot Challenge)

- Home tasks – commander & follower (Padmakumar et al., 2022) https://arxiv.org/abs/2110.00534
  - find best trajectory of actions to complete task
  - include dialogue, ask if unsure (Suglia et al., 2022) https://aclanthology.org/2022.sigdial-1.62/
- EMMA – integrated architecture for this
  - core: Transformer LM (+sparse attention)
  - "visual tokens" to represent video input
  - explicit actions / words on the output
  - pretraining on many vision & language tasks
    - captioning, image QA, location, relations


Let's put all the forks in the sink. Check the fridge to your left.
a
b
c
d
Put all Fork In any Sink
[ ] The Fork needs to be put into a Sink
Found a fork on the top shelf. Where's the sink?


Locate the object: microwave
Describe <vis_token_7>
Explain how <vis_token_6> relates to <vis_token_7>
Pretrained Model
<vis_token_1>
Gray countertop
Papers on counter
What material are the cabinets made of?
Describe the image
Denoise the caption: <mask> sink
Pretrained Model
The cabinets are made of wood.
A kitchen with a white sink.
White sink



Domain Adaptation

- pretraining
  - BERT, but also any other model
  - weight sharing: copy weights for similar slots in target domain
- delexicalization
  - assuming your domains are similar (e.g. TVs → PCs)
- pseudo in-domain data selection
  - find data similar to your domain in the source domain
- forcing shared latent space (see few-shot end-to-end models)
- multi-task training
  - your task in source domain & different task in target domain
- partial handcrafting (see Hybrid Code Networks)

Summary

- “traditional” multimodal systems, with components
    - combination of off-the-shelf components
        - parallels for ASR/NLU & NLG/TTS in I/O modalities
    - dialogue typically quite simple
    - modalities: static graphics / touch / gaze / facial expr. / avatars / robots
    - often support multi-party dialogue
- end-to-end multimodal systems
    - mostly experimental, based on HRED/Transformer with pretrained CNNs
        - VGG, ResNet, Inception (just image classification), Faster R-CNN, YOLO (+object detection)
    - MLLMs: multimodal instruction tuning, adapter modules, visual & audio tokens
    - visual dialogue: questions & answers about an image, finding an object in image
    - task-oriented: shopping dialogue with product images
    - situated tasks: discussing & executing household actions

Thanks

9 January
Last lecture

Contact us:

https://ufaldsg.slack.com/
{odusek,hudecek,kasner}@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

Get the slides here:

http://ufal.cz/npfl099

References/Inspiration/Further:

- Volha Pethukova's course (Uni Saarland): https://www.lsv.uni-saarland.de/multimodal-dialogue-systems-summer-2019/
- McTear et al. (2016): The Conversational Interface – Talking to Smart Devices
- Delgado & Araki (2005): Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment
- papers referenced on slides