Dialogue history and quality
vs
GuessWhat task success

Games mentioned in the papers

GuessWhat

- (de Vries et al., 2017)
- Asymetric game
- Q-bot asks Y/N questions to guess the target object of 20 candidates
- Sees image + dialogue history
- A-bot provides the answers
- Trained on human dialogues, crowdsourced
- Humans can stop asking Q at any time, bots have a fixed amount
- Visual grounding happens during question generation

GuessWhich

- (Das et al., 2017b)
- Asymetric game
- Q-bot cannot see image, has access to image captions
- Q-bot can ask any kind of Q
- Image is selected among 2K candidates
- A-bot sees caption + target image
- Human dialogues from VisDial dataset (chit-chat dialogues)
- Humans and bots ask 10 Q
- Visual grounding happens only during the guessing phase

MutualFriends

- (He et al., 2017)
- Symetric game
- Based only on text
- Two agents
- Both having a private list of friends described by a set of attributes
- Try to identify their mutual friend based on attributes
- Task only based on language

GuessWhat dataset

- Collected via Amazon Mechanical Turk, image from MS-COCO dataset
- One participant is assigned a target object in the image and the other has to guess it asking Y/N Q
- 155K Englihs dualogues
- 66K different images
- 52.2% No, 45% Yes, 2.2% N/A
- Training 128K datapoints, testing 23K
- 5.2 QA pairs on average
- Vocabulary 4900 words
- Between 3 and 20 candidates

Presenting papers:

- Testoni, A., & Bernardi, R. (2021). **The Interplay of Task Success and Dialogue Quality**: An in-depth Evaluation in Task-Oriented Visual Dialogues. In EACL 2021 (pp. 2071–2082)
- Greco, C., Testoni, A., & Bernardi, R. (2020). Which Turn do Neural Models Exploit the Most to Solve GuessWhat ? **Diving into the Dialogue History Encoding in Transformers and LSTMs**. In Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) (pp. 29–43).

Language/Dialogue quality

Testoni, A., & Bernardi, R. (2021)

Main points

- **Wheather and when language quality contributes to task success**
- Different complexity for guessing the target and asking questions
- Game can be won in a short time, generating human-like dialogues takes much longer
- Holds for all three tasks and models
- GuessWhat focus becase the dialogues play a major role in the guessing task
- Introducing the LD metric

Metrics (introducing LD)

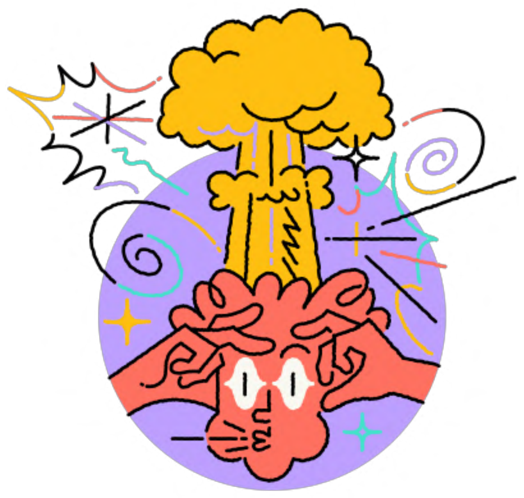
- **Task Success**: Accuracy for GuessWhat and MutualFriends, Mean Percentile Rank for GuessWhich
- **Linguistic metrics**:
  - *Unigram entropy*: unique unigrams/total number of tokens
  - *Mutual overlap*: average of a BLEU-4 by comparing each Q to with other Qs in same dialogue
  - *One question repeated verbatim* in a dialogue
  - *Global Recall*: % of learnable words that the models recall during generation
  - *Local Recall-d*: the normalised lexical overlap between a human and generated dialogue

LINGUISTIC DIVERGENCE

- all values normalised between 0 and 1
- "lower is better"
- overall vocabulary usage, diversity of questions/phrases, similarity of content with human dialogues

Models

- **GuessWhat**: A-Bot (Vries et al.,2017), Q-Bot: GDSE-SL and GDSE-CL (Shekhar at al., 2017) + reinforcement learning (Strub et al., 2017)
- **GuessWhich**: A-Bot from Das et al.(2017b), Q-Bot: Diverse (Muhari at al., 2019) and ReCap (Testoni et al., 2019)
- **MutualFriends**: DynoNet(He at al., 2017)



Quick results

with graphs

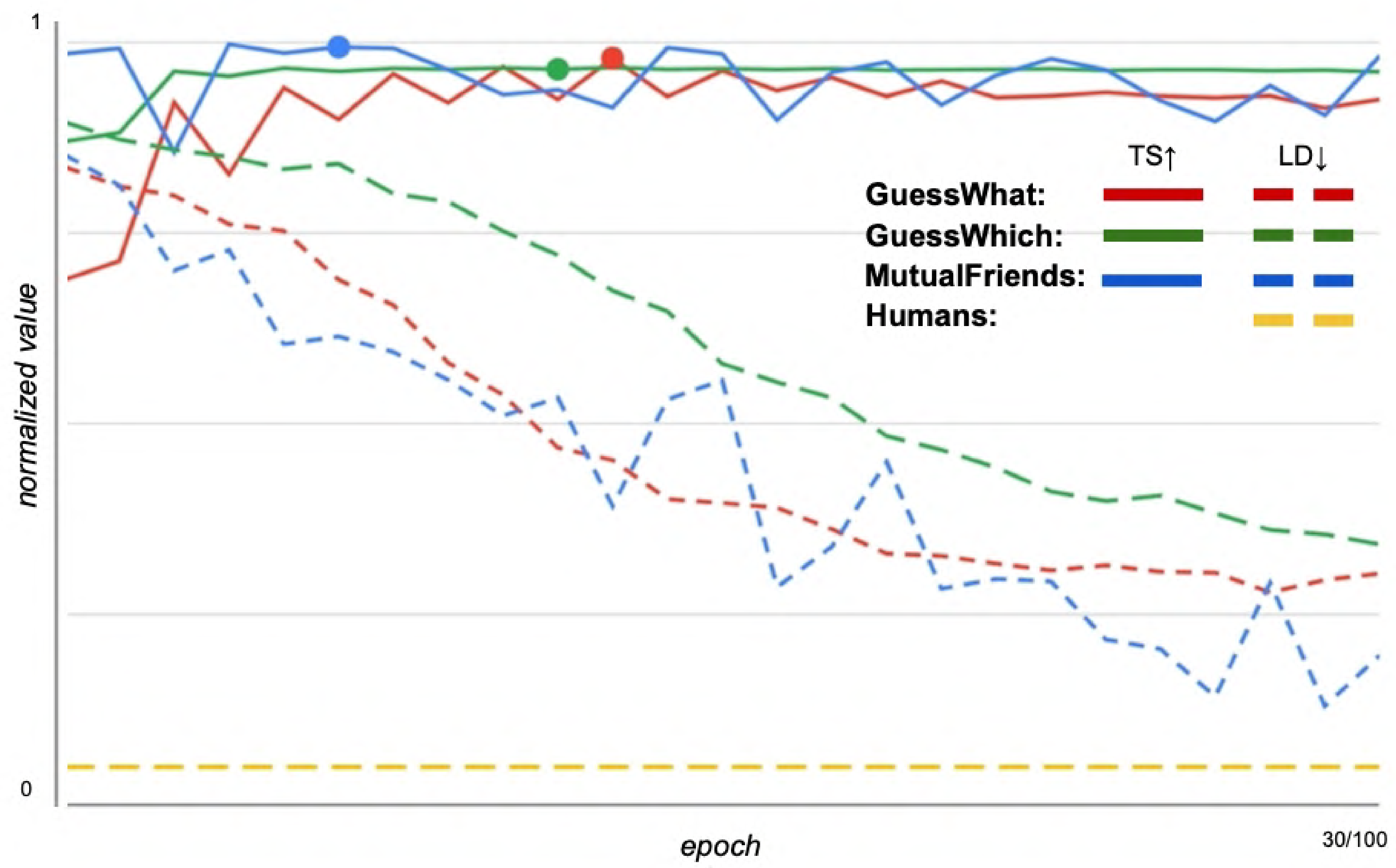
Models use very frequent words

	GuessWhich				GuessWhat				MutualFriends	
	D-SL	D-RL	ReCap-SL	Hum	GDSE-SL	GDSE-CL	RL	Hum	DynoNet-SL	H
TS ↑	95.2	94.89	96.76	-	48.21	59.14	56.3	84.62	0.98	0.82
GR ↑	6.46	9.04	14.4	27.69	34.73	36.35	12.67	72.98	51.15	65.2
LRd ↑	39.93	41.83	42.76	-	42.1	42.41	34.51	-	-	-
MO ↓	0.51	0.41	0.23	0.07	0.39	0.23	0.46	0.03	-	-
GRQ ↓	93.01	81.17	55.37	0.78	64.96	36.79	96.54	0.8	-	-
H ↑	4.03	3.92	4.19	4.55	3.52	3.66	2.42	4.21	3.91	4.57
LD ↓	0.58	0.52	0.38	-	0.46	0.36	0.67	-	0.18	-

Table 2: Comparative analysis of different models on several tasks and datasets.TS: task success. GR: global recall. LRd: local recall. MO: mutual overlap. GRQ: games with repeated questions. H: unigram entropy. LD: linguistic divergence. ↑: higher is better. ↓: lower is better.

LRd is maximum 42% indicating the models use very frequent words

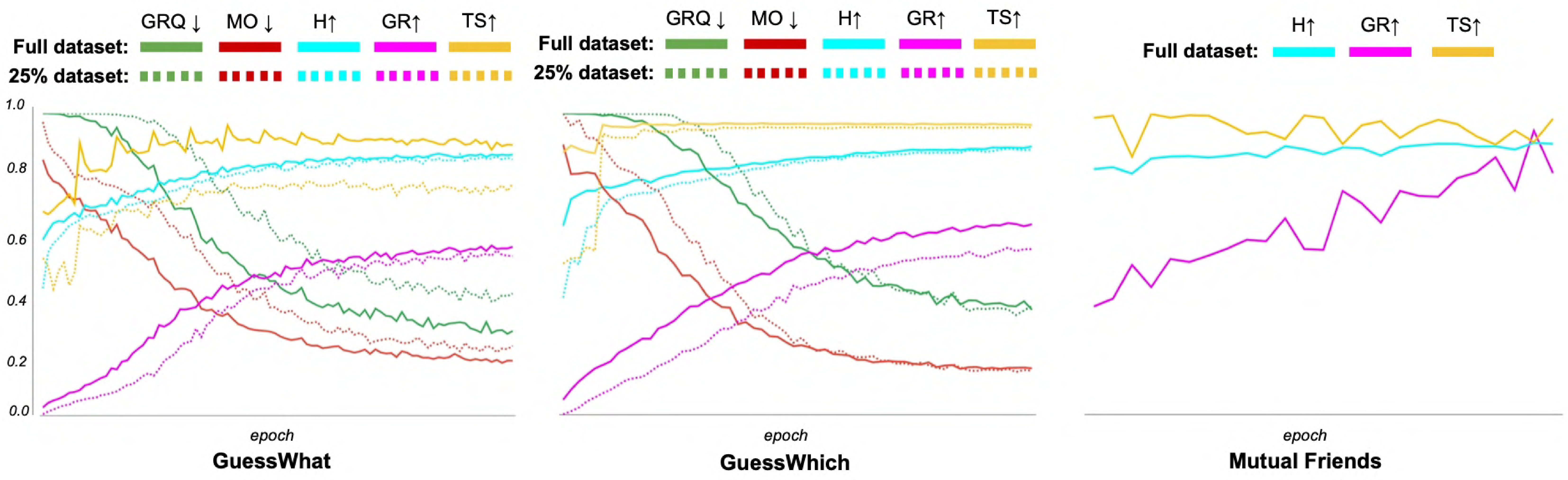
Choosing a model purely on TS prevents if from learning linguistic skills, but linguistic skills don't contribute to TS



The models learn to perform on the task quite quickly (see the distribution over epochs)

Pitch

How well the model learns to ground language plays an important role in TS

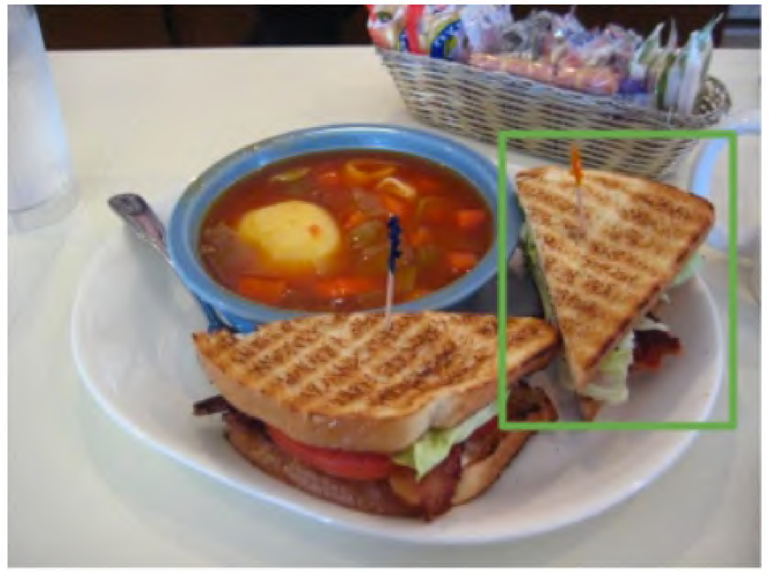


The impact of downsizing training data over epochs

Pitch

The GuessWhat model "struggles" with rare words

Human dialogues		Generated dialogues	
Questioner	**Answerer**	**Questioner**	**Answerer**
1. Is it a man?	No	1. Is it a person?	No
2. Is it food?	Yes	2. Is it food?	Yes
3. Is *pancake*?	No	3. Is it pizza?	Yes
4. Is egg?	Yes	4. Is it the pizza in front?	Yes
		5. . . .	
⇝ **model suceeds guessing**		⇝ **model fails guessing**	



Human dialogues		Generated dialogues	
Questioner	**Answerer**	**Questioner**	**Answerer**
1.Is it edible?	Yes	1. Is it food?	Yes
2.Is it a sandwich?	Yes	2. Is it a sandwich?	Yes
3. Does it have an orange *toothpick* in it?	Yes	3. Is it on the right?	Yes
		. . .	
⇝ **model fails guessing**		⇝ **model succeeds guessing**	

This could be due to the inability to generate or encode/ground rare words

Dialogue history

Greco, C., Testoni, A., & Bernardi, R. (2020)

Main points

- Focus on visually grounded dialogue history encoding
- GuessWhat as a "diagnostic" dataset
- Comparing SOTA models accross: architecture, input modalities and model background knowledge
- Transformers are less sensitive than LSTMs to the order in which QA pairs are provided
- Pre-trained versions are stronger at detecting salient information, independently of the poisiton
- ROBERTA provides the Guesser with the most informative representation
- The *blind* version of both the LSTM and Transformer models obtains higher/comparable results with the multimodal counterpart

Dataset

- Reasons why the GuessWhat dataset is suitable as **"diagnostic" dataset**: simplicity, the dialogue length mirrors the level of difficulty of the game and the most quetions in the last turs are answered positively and are longer than earlier ones
- Using only human dialogues, at most 10 turns (90K train, 18K eval and test)
- **Some findings of the analysis**: !

  - the shorter the dialogue the higher the % of Yes answers (average is balanced)

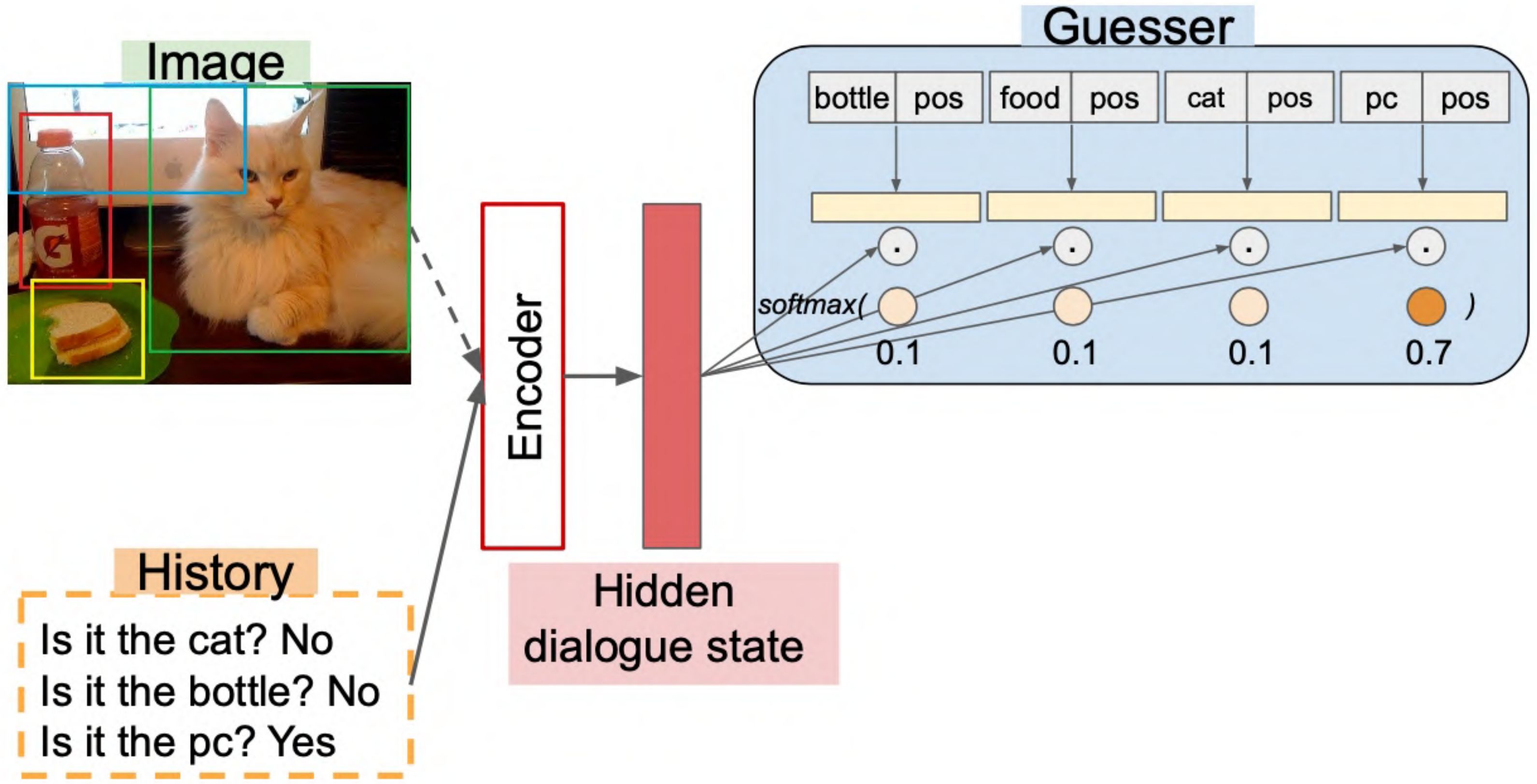
  - most of the Q in the last turns obtain a positive answer and these Q are longer than the previous ones

  - more difficult games have smaller target area, more distractors, the object is most likely a person

  - the more distractors from the same category the more difficult the game

Pitch

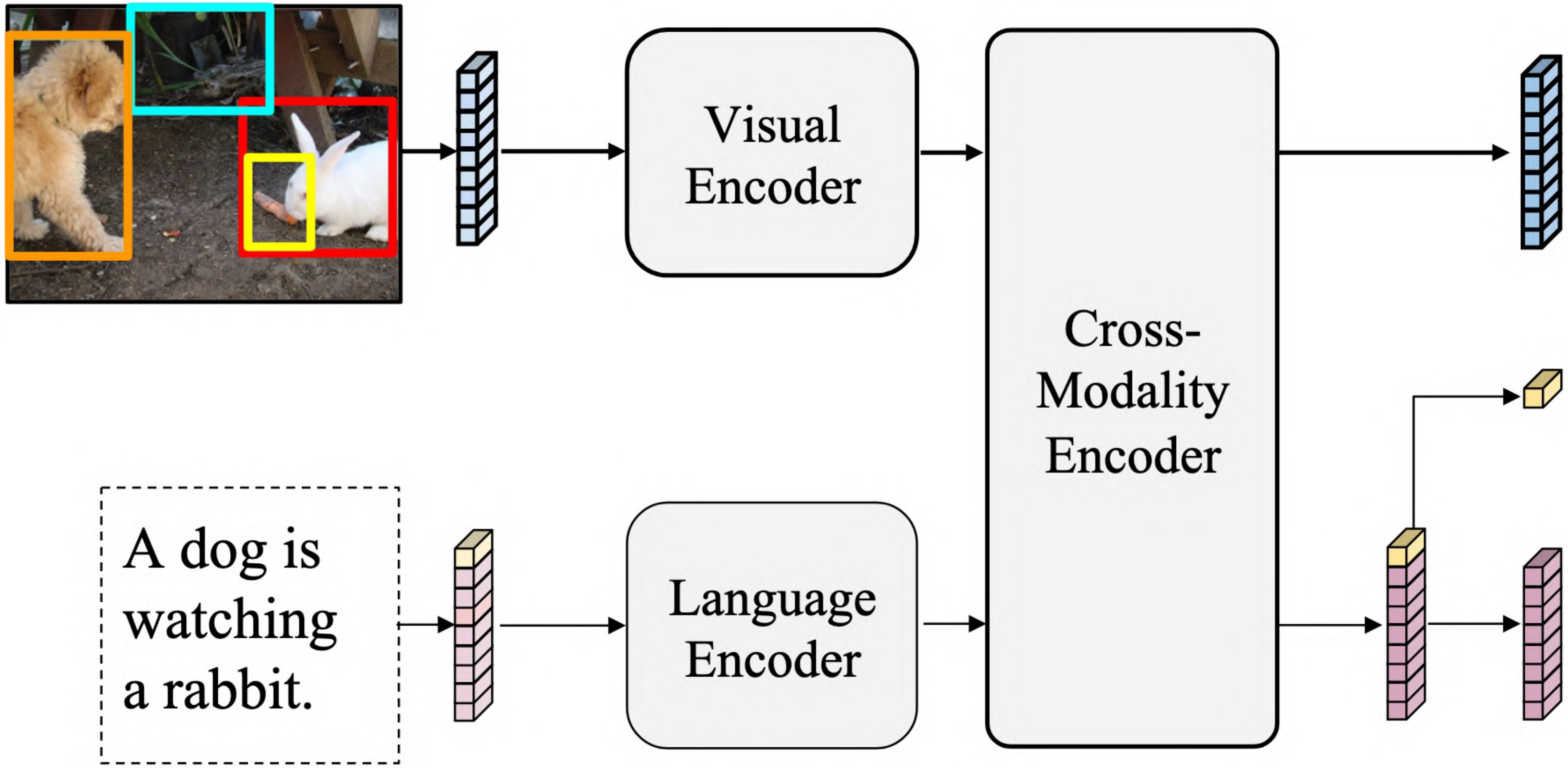
Models



- **Language Encoders:** representations of the candidates + hidden state obtained by an LSTM = only processes the dialogue history
- **RoBERTa:** RoBERTaBase, special tokens (CLS, SEP, EOS). The CLS token output is given to a linear layer with a tanh activation to obtain the hidden state then given to the Guesser.
- Both pretrained (RoBERTa) and trained from scratch(RoBERTa-S).

- **V-LSTM**: linguistic + visual representation (scaled), passed through a linar layer with tahn activation to obtain the hidden state. Frozen ResNet-152 pre-trained on ImageNet for the visual vectors.
- **LXMERT:** image = the set of position-aware object embeddings for the 36 most salient regions detected by a Faster R-CNN, text = position-aware randomly initialised word embeddings. Both representations are processed by a transformer encoder based on self-attention layers and their outputs are then processed by a cross-modality encoder that generates representations of the single modality enhanced with the other modality and their joing represenation. CLS and SEP. LXMERT (pre-trained) and LXMERT-S (from scratch).

LXMERT



Visual representation of the LXMERT architecture: source

Quick results

with graphs

Task sucess

		GT	Reversed
Blind	LSTM	64.7	56.0
	RoBERTa-S	64.2	57.8
	RoBERTa	**67.9**	66.5
MM	V-LSTM	64.5	51.3
	LXMERT-S	64.7	58.3
	LXMERT	64.7	60.3

Table 1: We compare the accuracy of models on the test set containing dialogues in the Ground Truth (GT) order of turns vs. the reversed order (reversed).

	LSTM	RoBERTa-S	RoBERTa	V-LSTM	LXMERT-S	LXMERT
All	64.7	64.2	67.9	64.5	64.7	64.7
3	72.5	72.7	75.3	71.9	73	73.8
5	59.3	58.3	60.1	59.3	59.2	58.7
8	47.3	45.1	51.0	47.2	46.8	43.3

Table 2: Accuracy with GT dialogues: results for all games, and for those of 3/5/8 dialogue length.

Dialogue history alone is quite informative to accomplish the task.

Does the order of the questions matter?

- Following a strategy: shorter questions in the beginning, longer in the end
- Reversing the dialogue shows that transformers are less senstive than LSTMs to the order
- A performace drop with models trained from scratch
- **Transformers seem to be able to identify salient information independently of the position in which it is provided within the dialogue history (Table1 previous slide)**

The role of the last question

Model	3-Q		5-Q		8-Q	
	All turns	W/o last turn	All turns	W/o last turn	All turns	W/o last turn
LSTM	72.5	53.4	59.3	46.8	47.3	38.4
RoBERTa-S	72.7	55.4	58.3	44.9	45	38.9
RoBERTa	75.3	58.2	60.1	49.3	51	42
V-LSTM	71.9	53.8	59.3	43.7	47.2	36.5
LXMERT-S	73	55.8	59.2	45	46.8	38.8
LXMERT	73.8	55.3	58.7	45.6	43.3	34.1

Table 3: Accuracy of the models when receiving all turns of the dialogue history and when removing the last turn for dialogues with 3, 5, and 8 turns.

- Results without the last turn
- All models have a similar drop in accuracy thus **last turn is the most informative**
- RoBERTa superitority - better encodes a full dialogue history, holds for all lenghts

How the attention is distributed across turns

- How much each turn contributes to the overall self-attention withing a dialogue by summing the attention of each toke within a turn
- **All models put more attention on the last turn**
- The attention heads of RoBERTa and LXMERT both in pre-trained and from scratch versions focus more on the last turn even in the reverse order

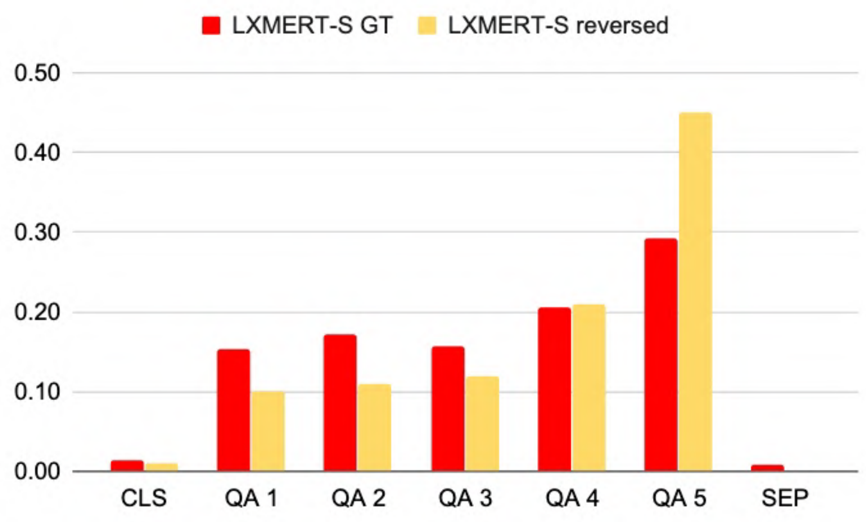


Fig. 5: Attention assigned by LXMERT-S to each turn in a dialogue when the dialogue history is given in the GT order (from QA1 to QA5) or in the reversed order (from QA5 to QA1).

Reproduction idea

What from the paper we will implement?

- The paper has excellent reproducibility details, including hyperparamter values
- Set up V-LSTM and Transformer models mentioned:

  - https://github.com/GuessWhatGame/guesswhat

  - https://huggingface.co/unc-nlp/lxmert-base-uncased
- Experiments: baseline, reverse dialogue history, remove last questions, blind
- Finding inspiration on how to build upon the project for the final

Reimplementation/Reproduction plan

Prep

Manipulate dataset

Run code LSTM

Set up Transformer

Set up docs

Experiment

Baseline: blind, multimodal

Reverse dialogue history

Remove last term

Document

Use W&B for running

Document results

Report

Package findings in a report

Thank you!

Ask away, about today's paper or the one I didn't present