W2_3: VISUAL GROUNDING AND MULTIMODAL NLP

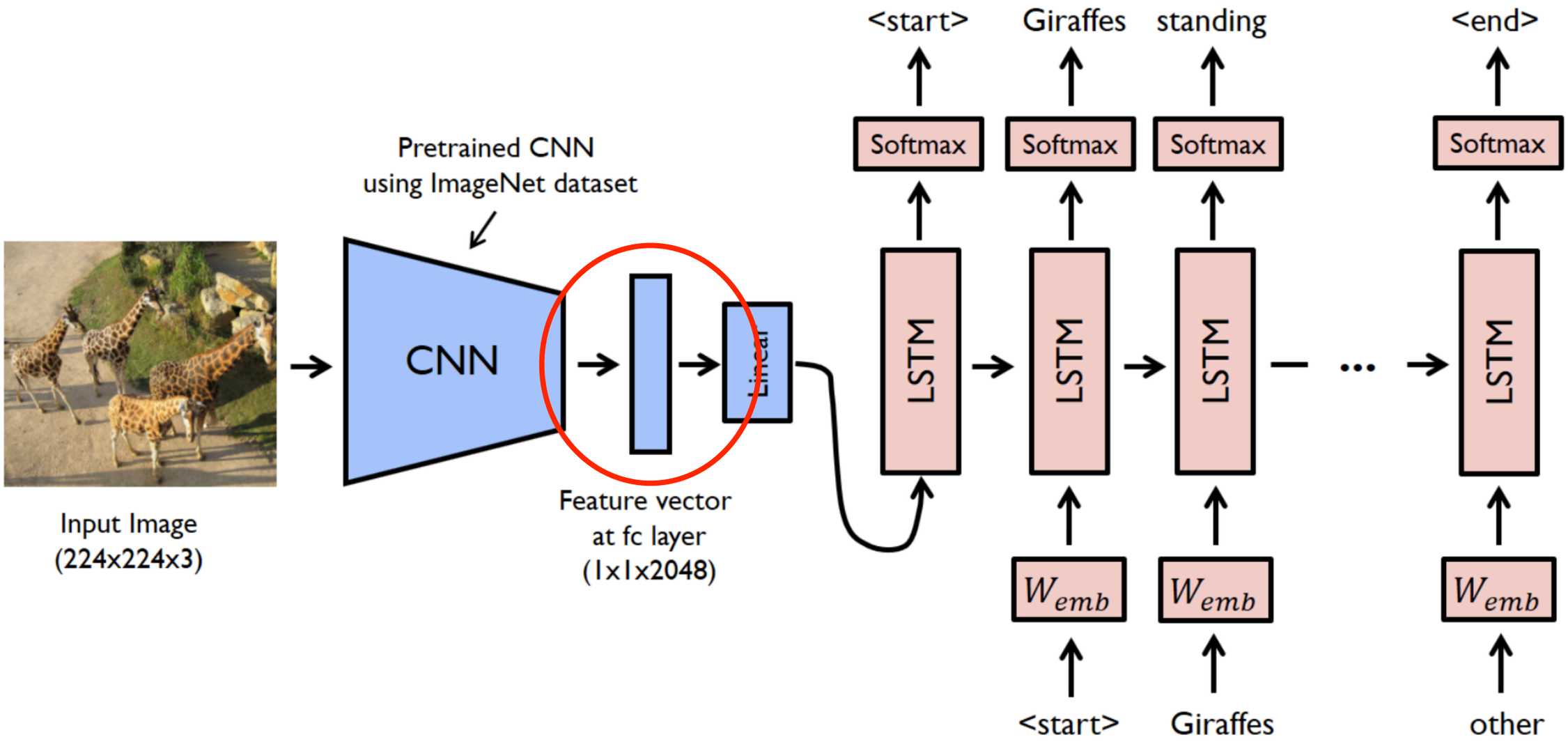
Cognitive Approaches to Multimodal Language Processing

Sandro Pezzelle
sandropezzelle.github.io

(SOME) LANGUAGE & VISION TASKS

1. Image Captioning (IC)
2. Visual Question Answering (VQA)
3. Visually-grounded dialogue

IMAGE CAPTIONING: OVERVIEW

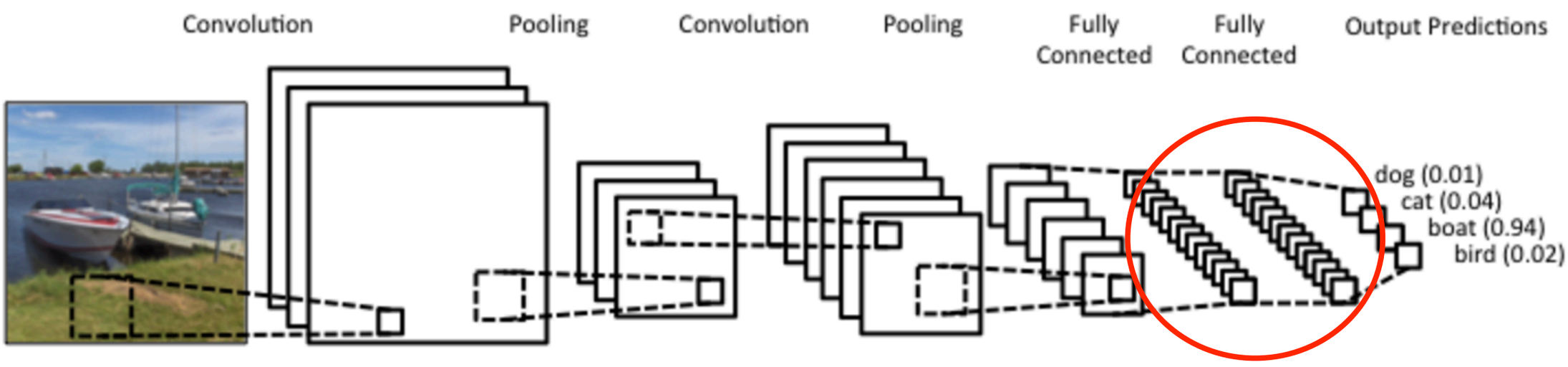


"Giraffes standing close to each other"

Image credits: https://github.com/anunay999/image_captioning_vgg16

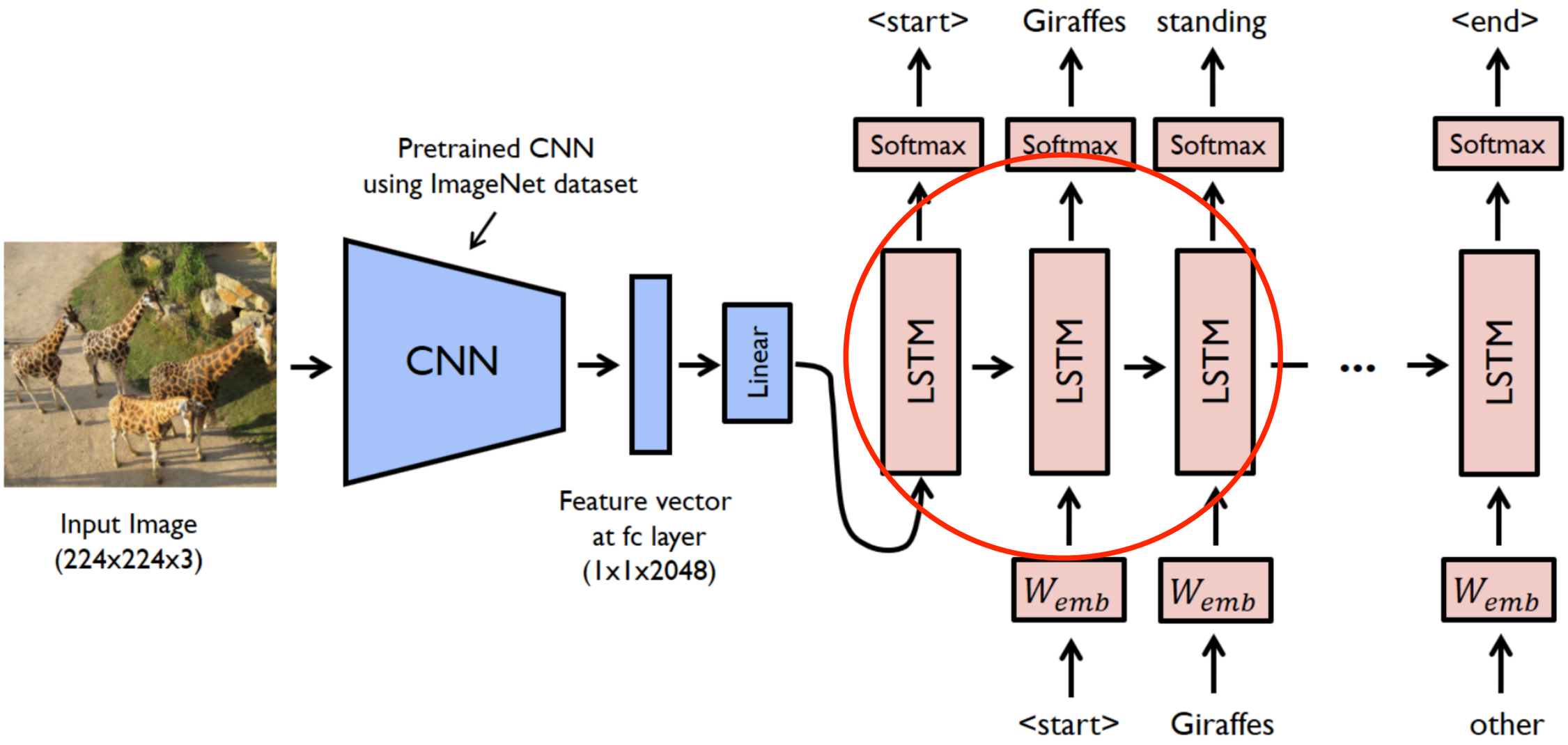
IMAGE CAPTIONING: VISUAL MODULE (CNN)

CNN pipeline: from *low-level* to ***abstract*** image features



Fully-connected layers: scene's abstract (***semantic***) representation

IMAGE CAPTIONING: OVERVIEW

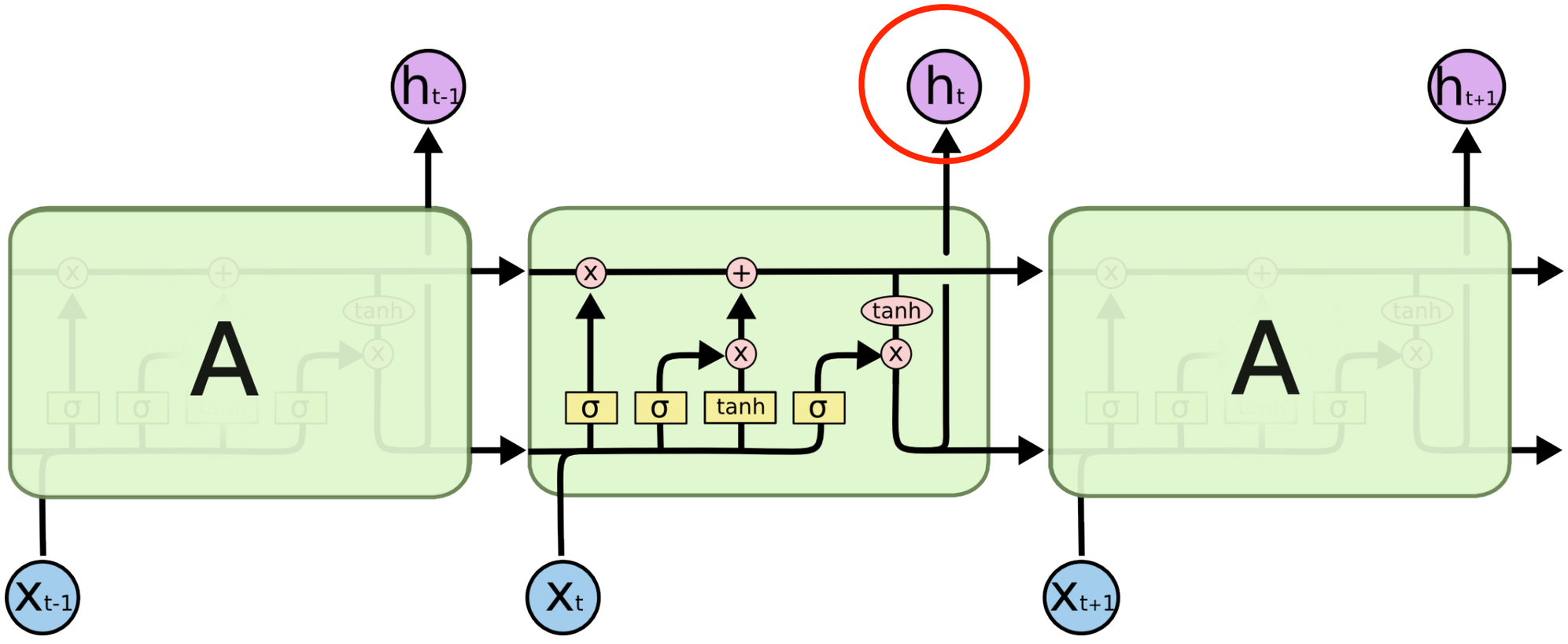


"Giraffes standing close to each other"

Image credits: https://github.com/anunay999/image_captioning_vgg16

IMAGE CAPTIONING: LANGUAGE MODULE (RNN/LSTM)

Hidden state: ***semantic*** representation of the sentence (so far)

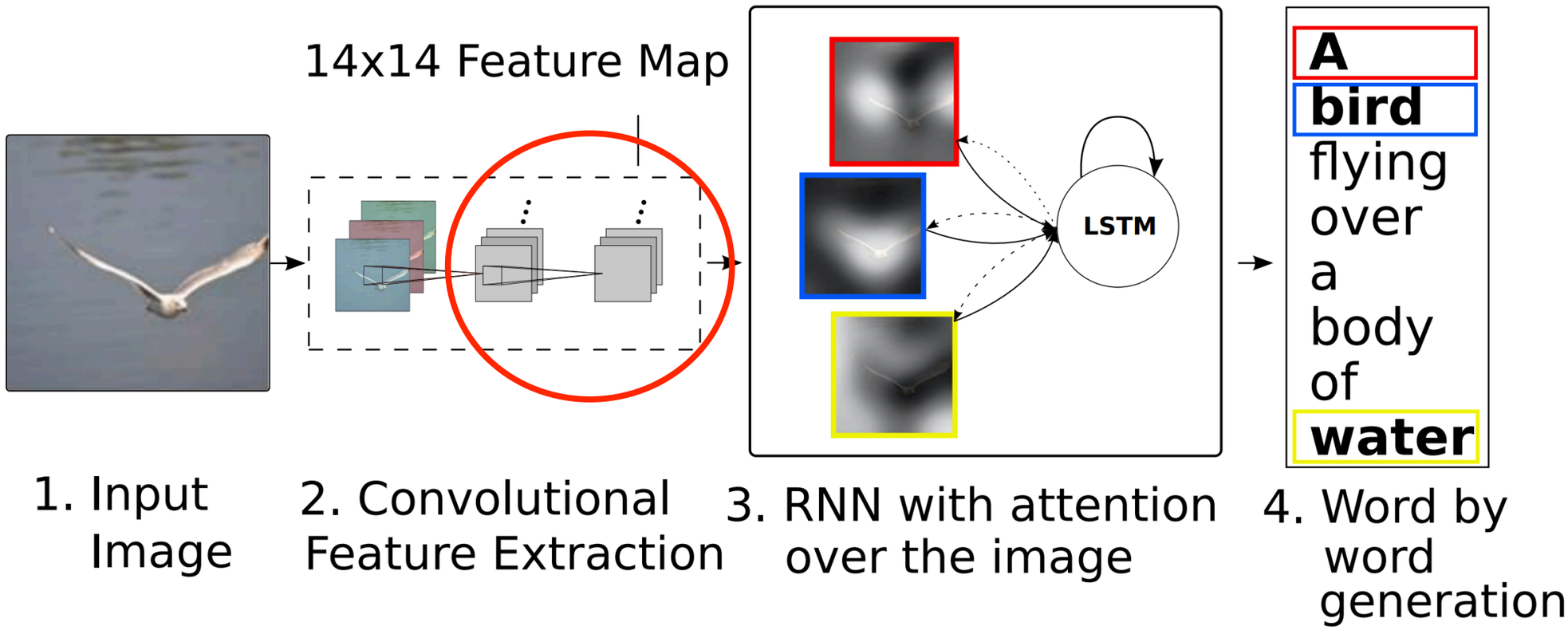


<start> *Giraffes* *standing*

Image credits: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

IMAGE CAPTIONING: ATTENTION

- Instead of ***static*** visual representation (i.e., FC layers), ***dynamic*** **focus (~ attention) on salient image features!**

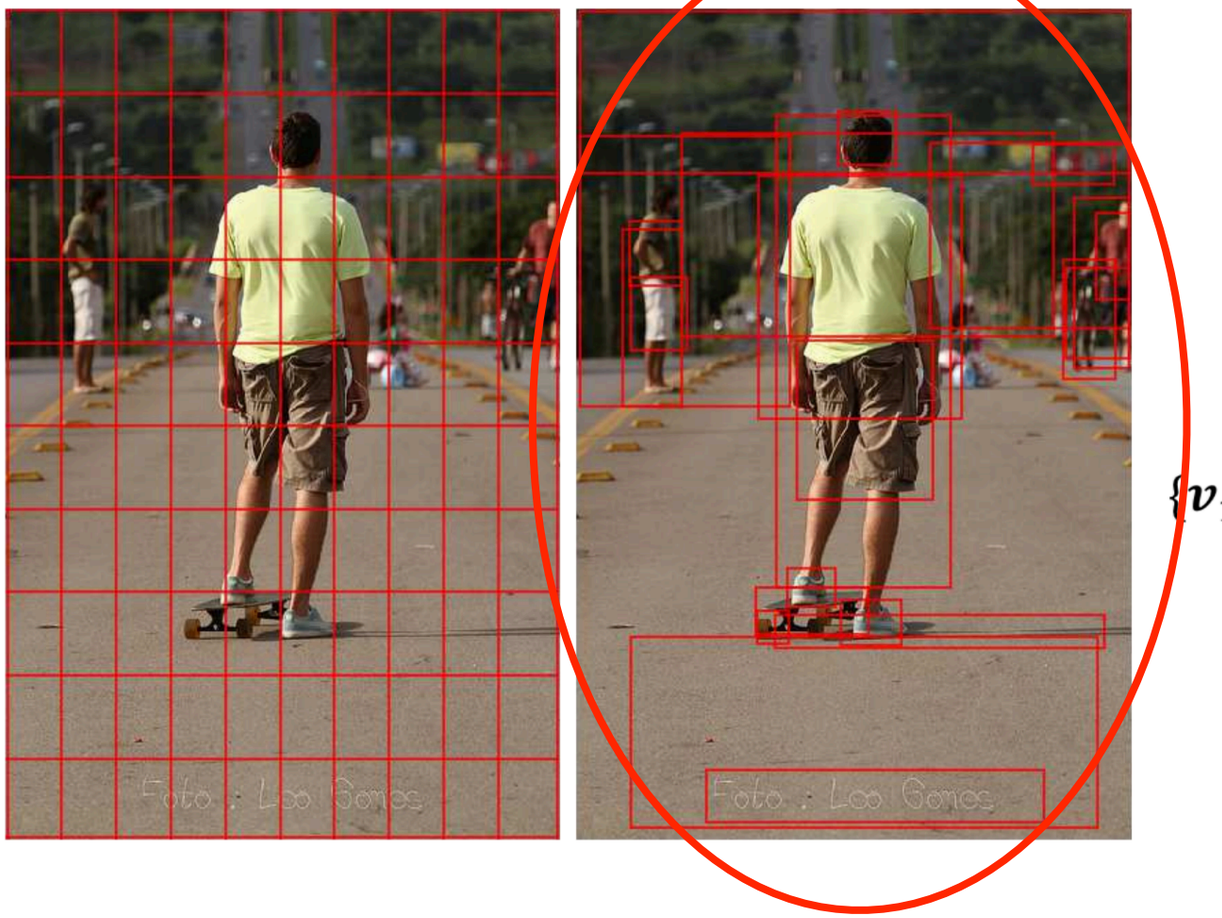


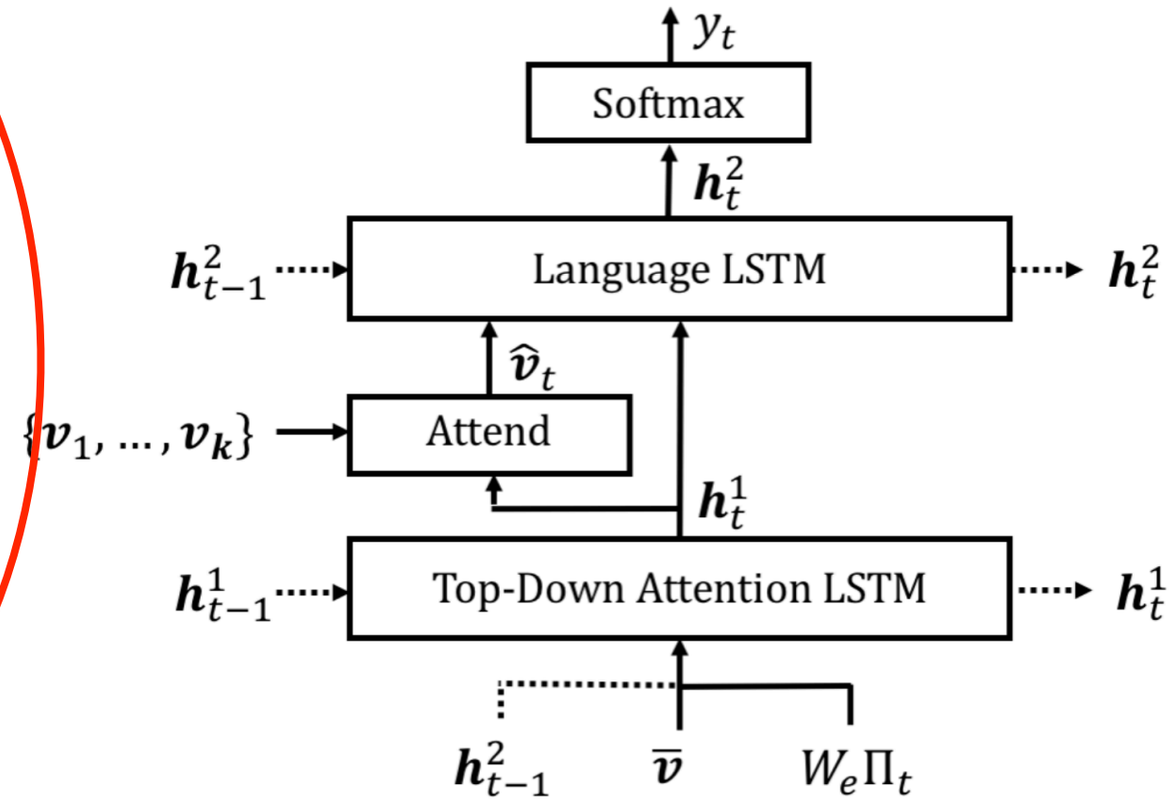
Feature maps: a step ***before*** FC layers — spatial properties preserved!

[Xu et al. 2015]

IMAGE CAPTIONING: SOTA MODEL

- **Bottom-up** + **Top-down attention**: attention calculated at the level of (**proposed**) **objects** and other salient regions

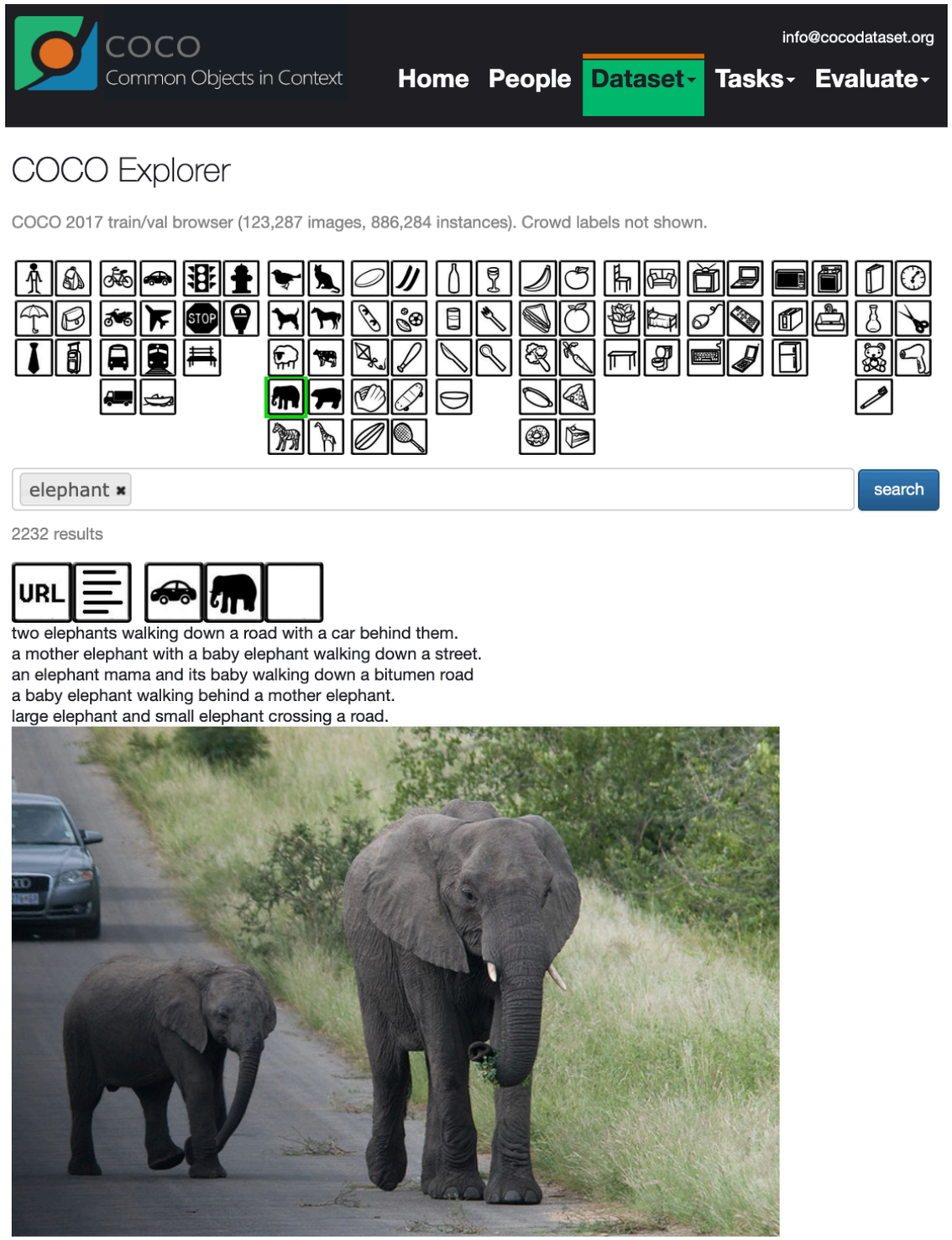




- Best-performing model on **MS-COCO dataset**!

[Anderson et al. 2018]

IMAGE CAPTIONING: MS-COCO DATASET



Dataset statistics

- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories (e.g., elephant, car, etc.)
- **5 captions per image**
- SOTA models leaderboard

webpage: http://cocodataset.org/#home

IMAGE CAPTIONING: METRICS

- Metrics from **machine translation!**
- Degree of ***word overlap*** between **original** (source language) and **generated** (target language) description
  - BLEU, SPICE, METEOR, CIDEr, ROUGE, WMD, etc.

IMAGE CAPTIONING: METRICS

- Metrics from **machine translation!**
- Degree of ***word overlap*** between **original** (source language) and **generated** (target language) description
- BLEU, SPICE, METEOR, CIDEr, ROUGE, WMD, etc.

More on W3!

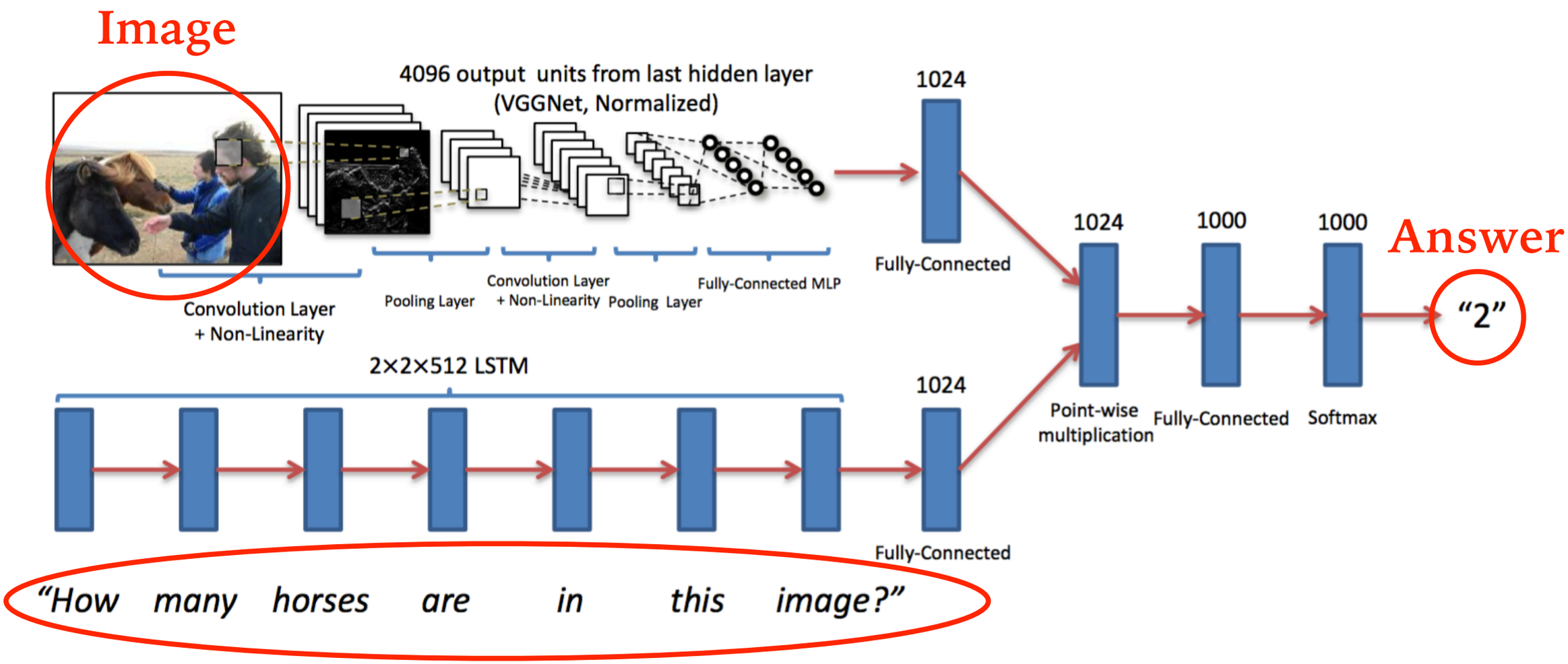
PRAGMATIC IMAGE CAPTIONING: RSA-BASED APPROACHES

RSA-based Image Captioning: Pragmatically-informative captions

*"A man wearing a white shirt and **glasses**"*

[Cohn-Gordon et al. 2018]

VISUAL QUESTION ANSWERING: OVERVIEW



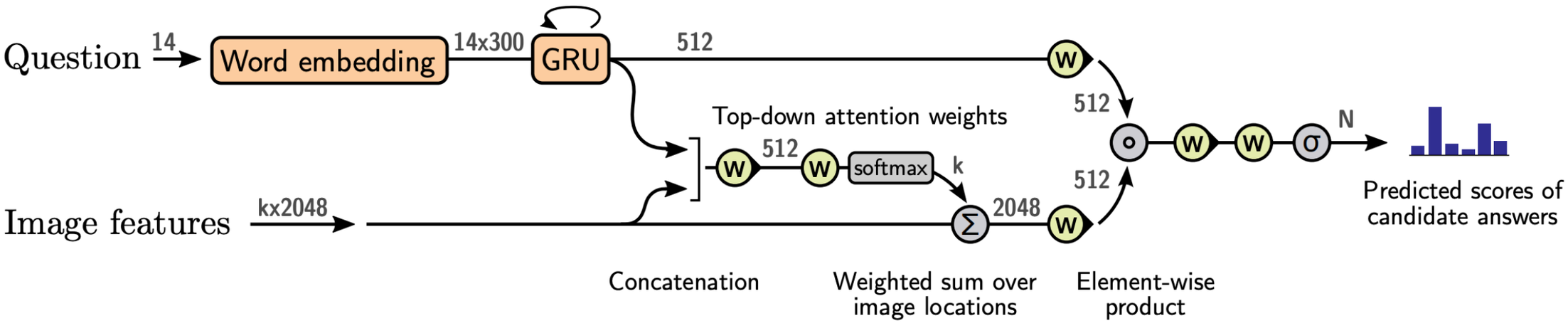
[Antol et al. 2015]

VISUAL QUESTION ANSWERING: UP-DOWN ATTENTION

70.3% accuracy on VQA 2.0 test set *(ensemble)*

Question

Answer

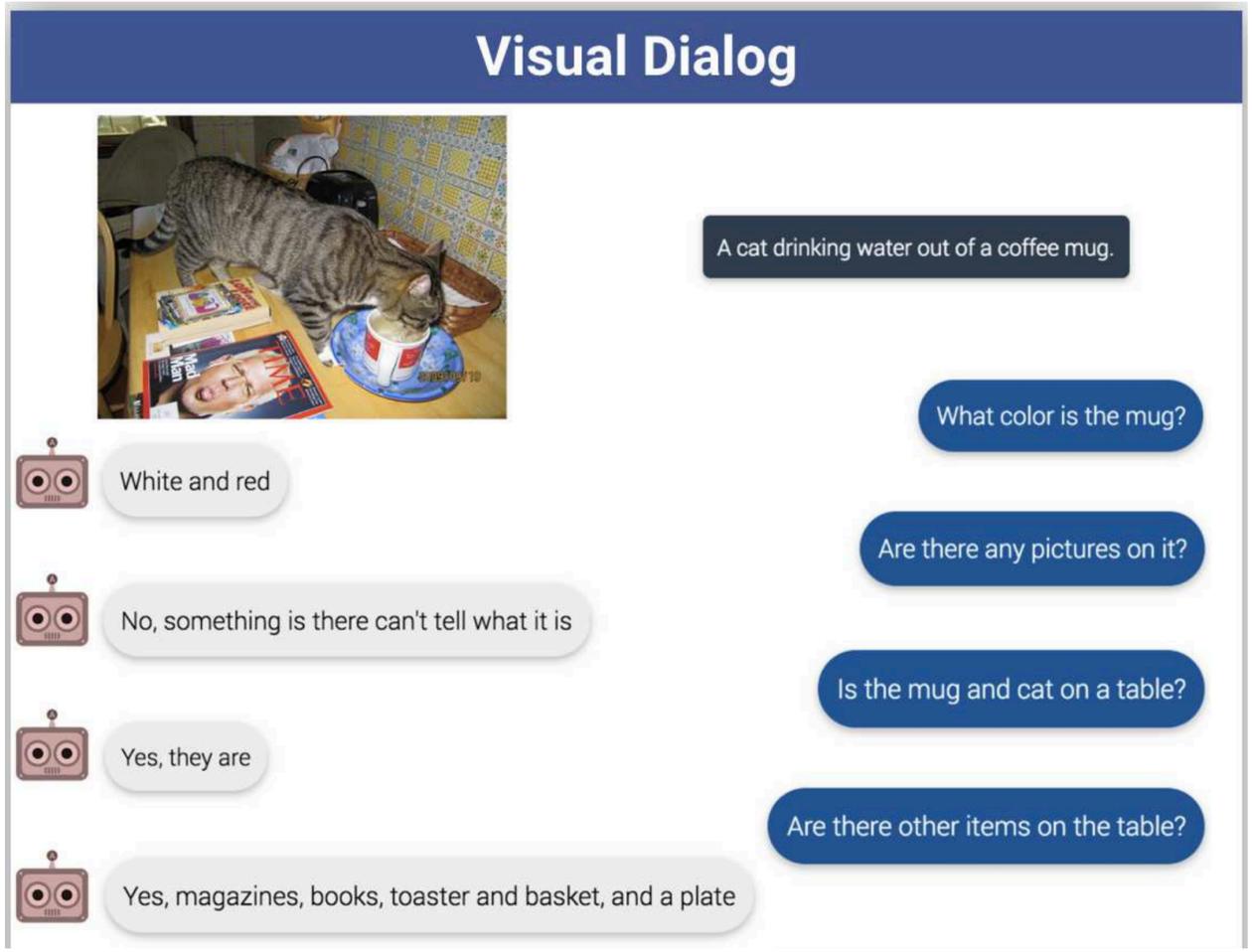


Image

[Anderson et al. 2018]

(best model **72.5%** acc.: In a few slides…)

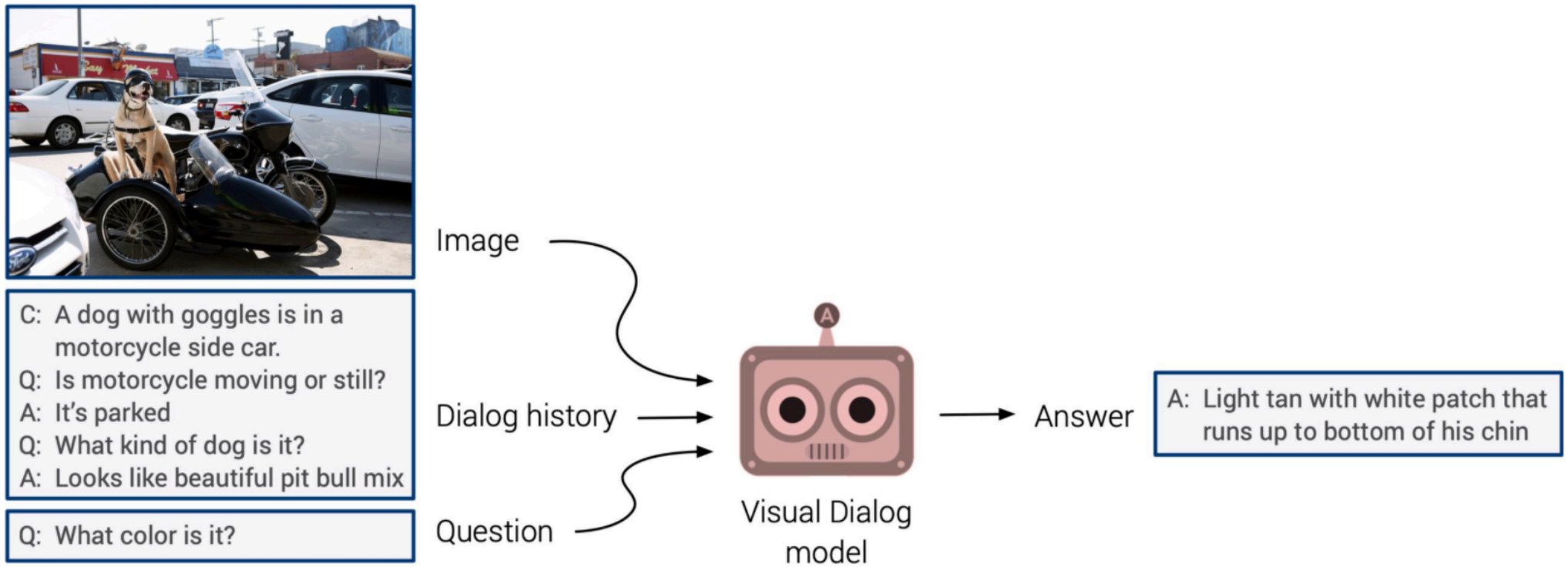
VISUALLY-GROUNDED DIALOGUE: OVERVIEW



- **Dialogue** about an **image**: ask and answer questions
- Dialogue ~ sequence of ***(question, answer)*** **pairs**
- Questioner's goal: get **info on image** *(not seen)*
- Answerer's goal: **answer** Questioners' questions

[Das et al. 2017]

VISUALLY-GROUNDED DIALOGUE: TASK



Model/answerer **task**:

answer questions correctly based on **Image + History**

[Das et al. 2017]

VISUALLY-GROUNDED DIALOGUE: GUESSWHAT?! APPROACH

Questioner	**Oracle**
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

- Model tasks: Ask questions (Q) / Guess the object (O)

[De Vries et al. 2017]

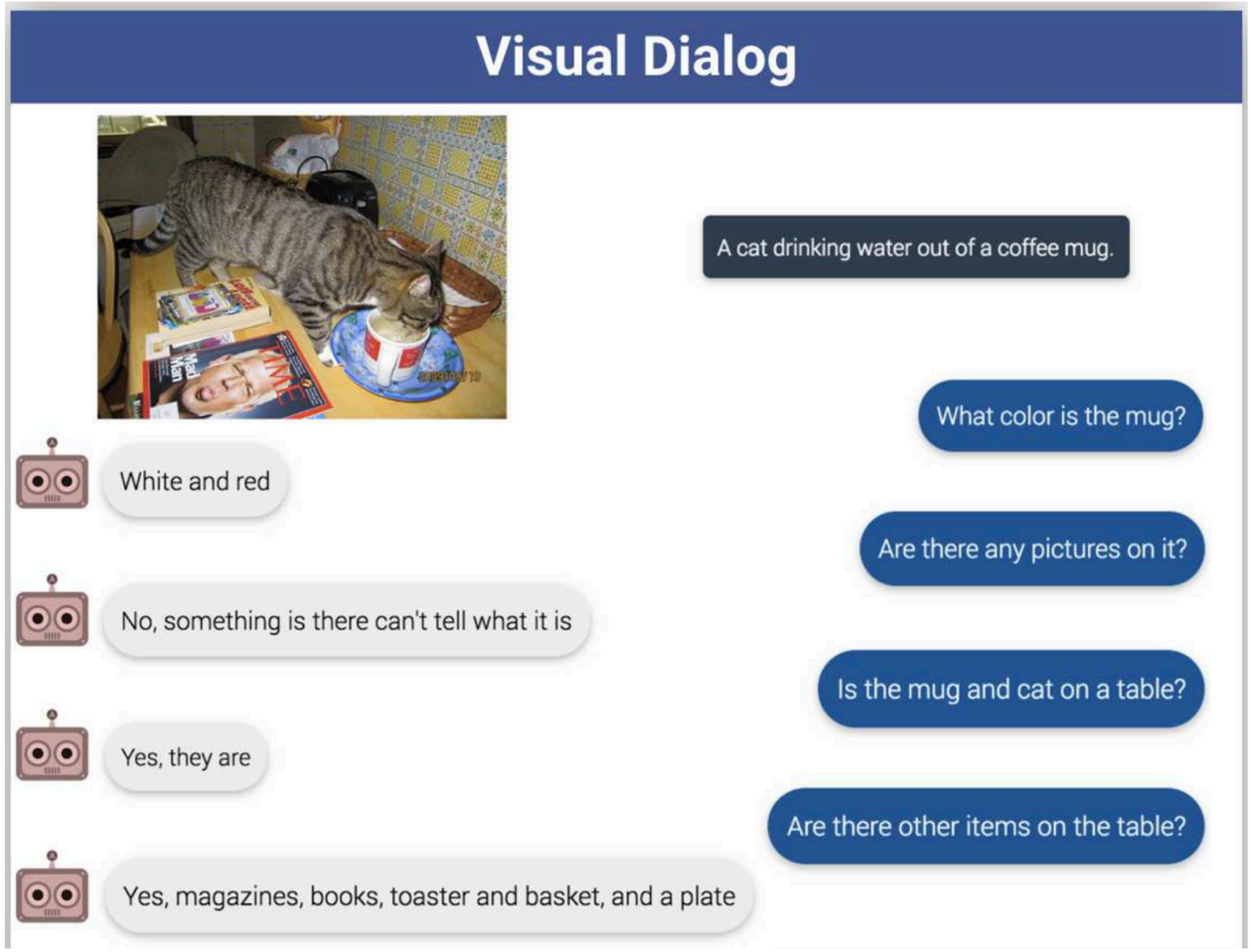
VISUALLY-GROUNDED DIALOGUE: GUESSWHAT?! APPROACH

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

➤ Model tasks: Ask questions (Q) / Guess the object (O)

[De Vries et al. 2017]

VISUALLY-GROUNDED DIALOGUE: DATASET



- 120k MS-COCO images
- 1 dialogue / image
- **10 rounds** of question-answers / dialogue
- Total 1.2M dialogue question-answers

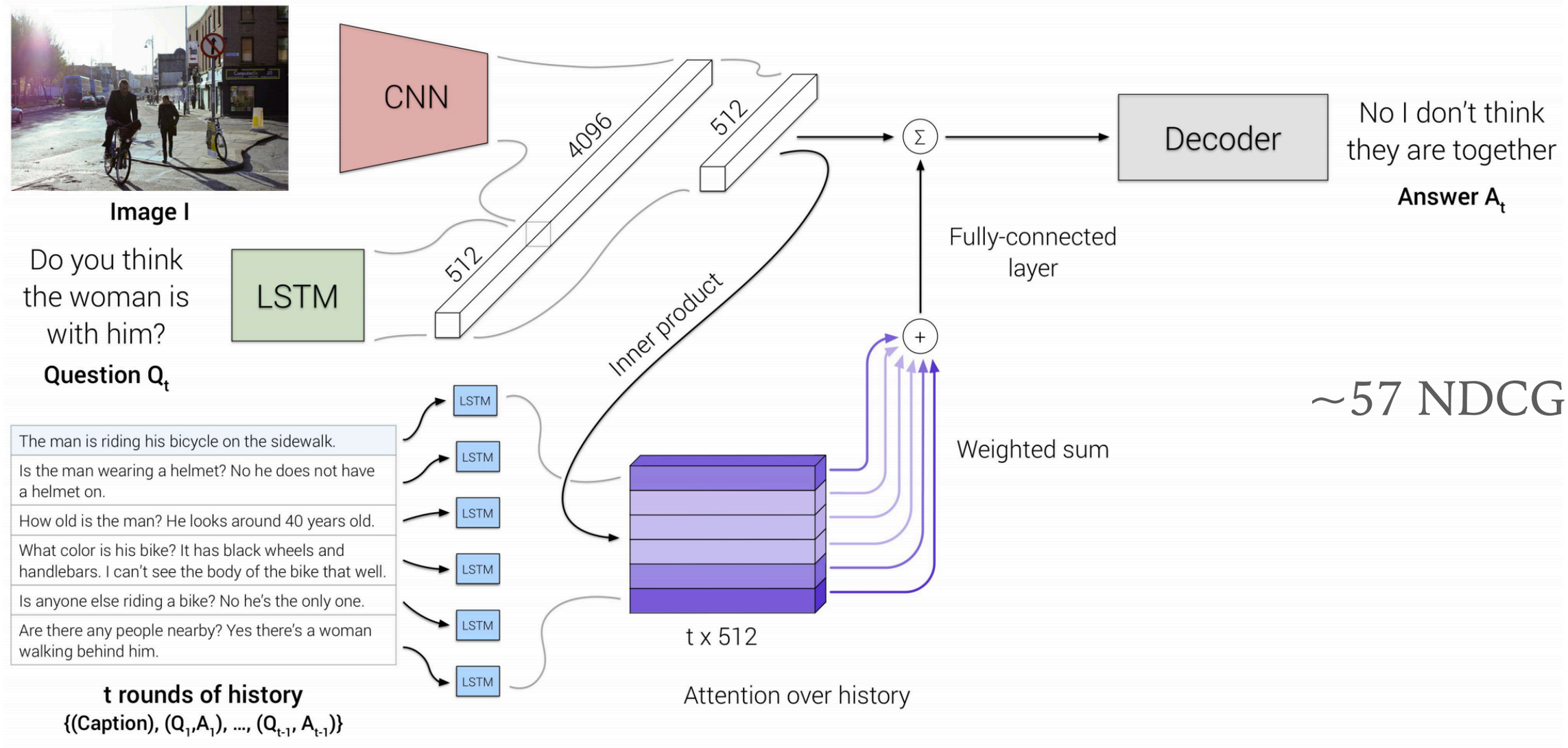
[Das et al. 2017]

VISUALLY-GROUNDED DIALOGUE: EVALUATION

- Treated as either **discriminative** or **generative** task
- **Retrieval metrics:** Mean Reciprocal Rank (**MRR**)
  - Normalized Discounted Cumulative Gain (**NDCG**): *alignment* between model-generated answers ranking and ideal ranking (based answer ***relevance*** by humans)

Nice blogpost here: shorturl.at/lqZ12

VISUALLY-GROUNDED DIALOGUE: MEMORY NETWORK



~57 NDCG

(c) Memory Network Encoder

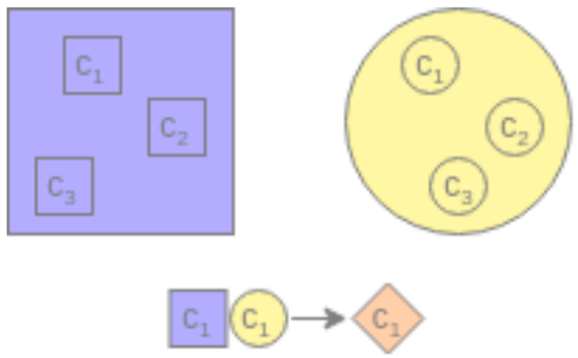
(best model **74.5 NDCG**: In a few slides…)

[Das et al. 2017]

VISUAL GROUNDING: 2 (MAIN) COMPUTATIONAL APPROACHES

Visually-grounded semantics

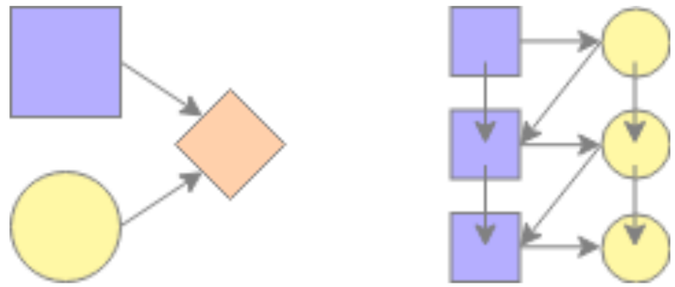
- **combination** of *known* textual and visual features to obtain **human-like representations** (task-agnostic)



(a) Multimodal fusion. Concatenate known representations from modality A and B and apply dimensionality reduction.

Multimodal machine learning (aka Language & Vision tasks)

- **joint multimodal processing** to understand/align/integrate **information from language and vision** (task-oriented)



(c) Joint multimodal processing. Left: Modality A and B both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

[Beinborn et al. 2018]

TASK-AGNOSTIC "VISIOLINGUISTIC" REPRESENTATIONS

- Recently, several approaches based on **BERT-like techniques** to obtain **task-agnostic multimodal representations**
- "the dominant strategy is to start with separate language and vision models [...] and then learn grounding as part of task training [...] **we seek to pretrain for visual grounding**"

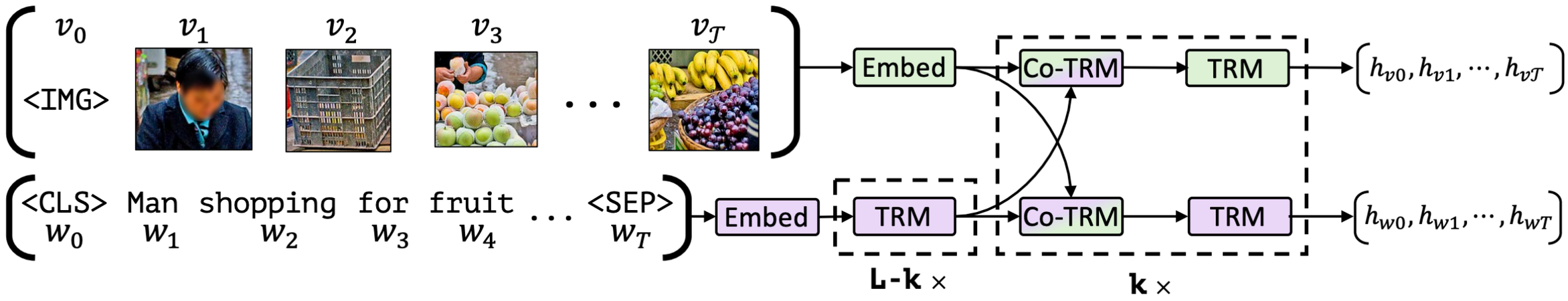
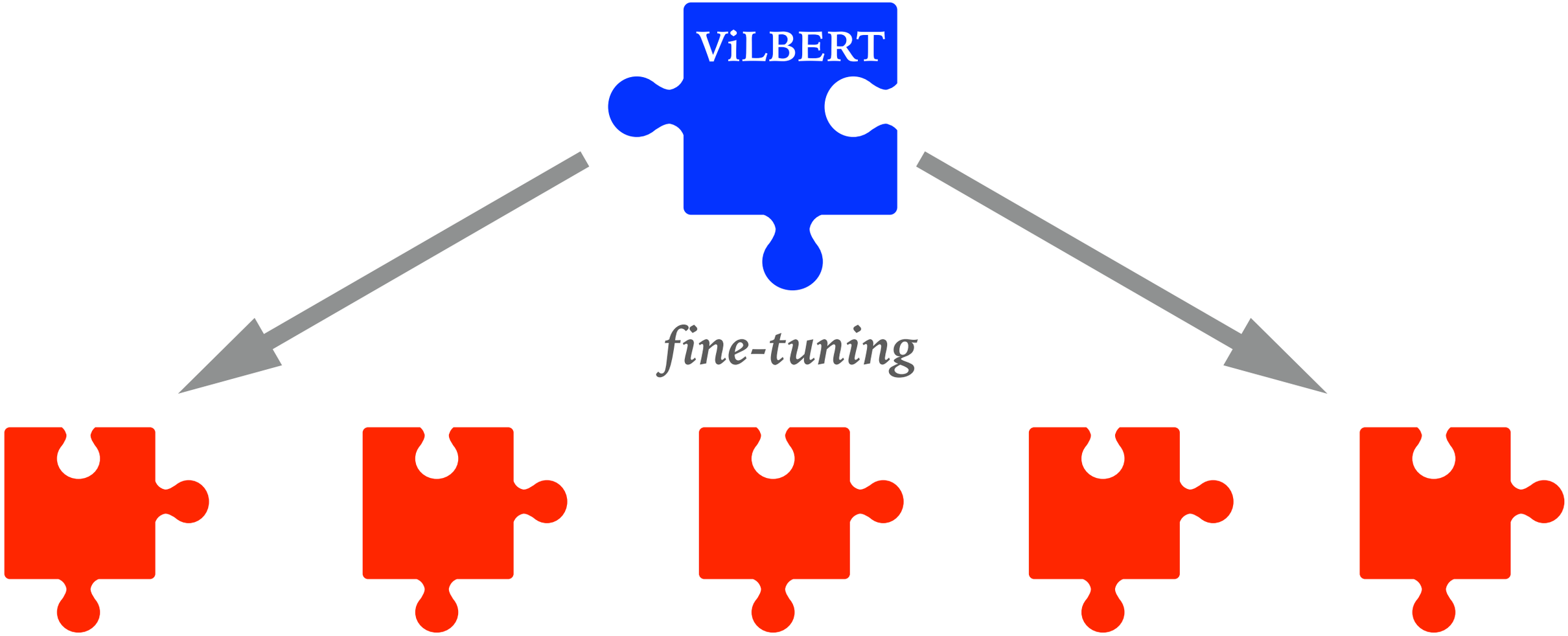


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

[Lu et al. 2019]

VILBERT-LIKE MODELS + FINE-TUNING: SOTA RESULTS



state-of-the-art performance!

[Lu et al. 2019; Tan & Bansal 2019; Murahari et al. 2019]

VILBERT-LIKE MODELS: PROS & CONS

PROs

- Very **powerful** models: state-of-art in many tasks
- **Task-agnostic** representations, similar to traditional approaches in visually-*grounded* semantics

CONs

- Highly-parametrized architectures + extensive pre-training + lot of data …
- Limited understanding of what type of representations are learned —> ***are they in line with human judgments?***

W2 REFERENCES

- Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., ... & Fernández, R. (2016, August). The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of ACL (Volume 1: Long Papers) (pp. 1525-1534).

- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In Proceedings of the ICML Deep Learning Workshop, Lille, France.

- Baroni, M. (2016). Grounding distributional semantics in the visual world. Language and Linguistics Compass, 10(1), 3-13.

- Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012, July). Distributional semantics in technicolor. In Proceedings of ACL: Long Papers-Volume 1 (pp. 136-145). Association for Computational Linguistics.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119).

- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of EMNLP (pp. 1532-1543).

W2 REFERENCES

- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep Contextualized Word Representations. In Proceedings of NAACL-HLT, Volume 1 (Long Papers) (pp. 2227-2237).

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers) (pp. 4171-4186).

- Mickus, T., Paperno, D., Constant, M., & van Deemeter, K. (2019). What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. arXiv preprint arXiv:1911.05758.

- Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In Proceedings of the 13th International Conference on Computational Semantics - Long Papers, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics

- Barsalou, L. W. (2008). Grounded cognition. Annu. Rev. Psychol., 59, 617-645.

- Harnad, S. 1990. The symbol grounding problem. Physica D: Nonlinear Phenomena 42(1–3). 335–46.

W2 REFERENCES

- Beinborn, L., Botschen, T., & Gurevych, I. (2018, August). Multimodal Grounding for Language Processing. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2325-2339).

- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

- Cohn-Gordon, R., Goodman, N., & Potts, C. (2018, June). Pragmatically Informative Image Captioning with Character-Level Inference. In Proceedings of NAACL-HLT, Volume 2 (Short Papers) (pp. 439-443).

- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).

- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., ... & Batra, D. (2017). Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 326-335).

W2 REFERENCES

- De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5503-5512).

- Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems (pp. 13-23).

- Tan, H., & Bansal, M. (2019, November). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of EMNLP-IJCNLP (pp. 5103-5114).

- Murahari, V., Batra, D., Parikh, D., & Das, A. (2019). Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline. arXiv preprint arXiv:1912.02379.