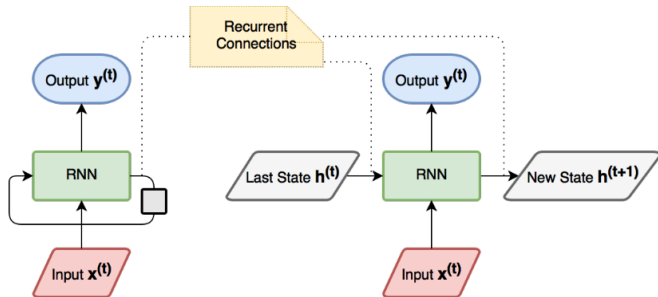


Neural Networks and Deep Learning: Vision & Langage

Nicolas Thome

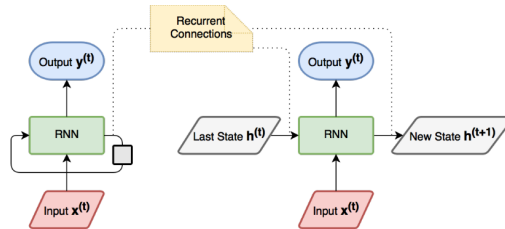
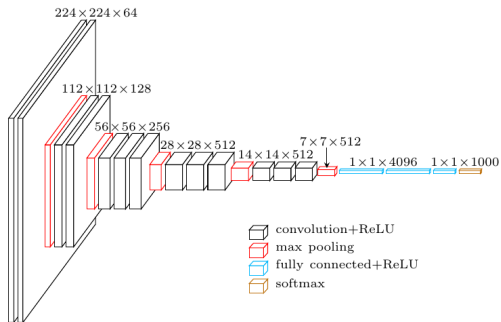
Conservatoire National des Arts et Métiers (Cnam)
Département Informatique

Recurrent Neural Networks (RNNs)



- ▶ **Sequences:** 1d/2d signals (e.g. audio, videos), molecules, text, etc
- ▶ Input vector $x(t)$, e.g. word (text) or image representation (CNN)
- ▶ Input/Output $h(t)$: vector representing model "short-term memory"
- ▶ Output vector $y(t)$: task dependent
- ▶ All parameters trained with backpropagation through time

New Tasks in Artificial Intelligence

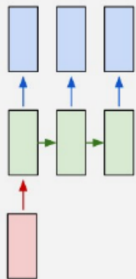


► Intersection of vision and language research

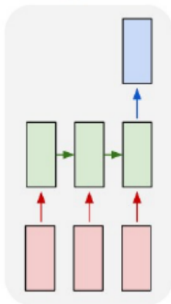
- Improvements in Vision Understanding with ConvNets
- Language (text) Modeling with RNNs

Sequence modeling with RNNs

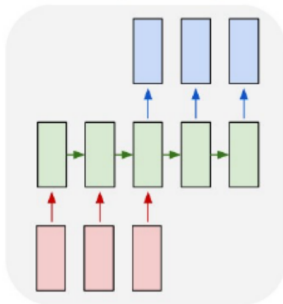
one to many



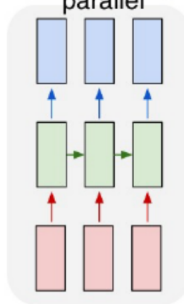
many to one



many to many

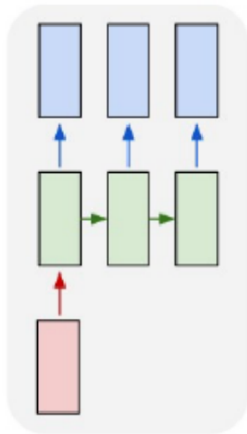


many to many
parallel



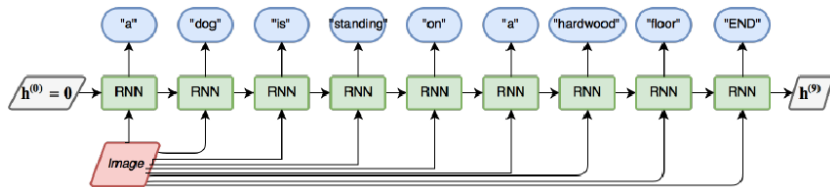
One to Many - Image captioning

one to many

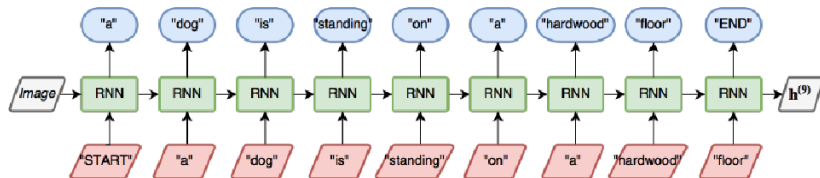


One to Many - Image captioning

- Show, Attend and Tell, Xu *et. al.*, ICML15 [Xu et al., 2015]



- Karpathy, CVPR15 [Karpathy and Li, 2015]



Many to One - Visual Question Answering (VQA)

- ▶ Goal: build a system that can answer questions about images



How many slices of pizza are there?
Is this a vegetarian pizza?



Does it appear to be rainy?
Does this person have 20/20 vision?



What color are her eyes?
What is the mustache made of?

- ▶ Very complex task, that requires :
 - ▶ Precise image and text models
 - ▶ High level interaction modeling
 - ▶ Full scene understanding, reasoning (e.g. spatial ...)

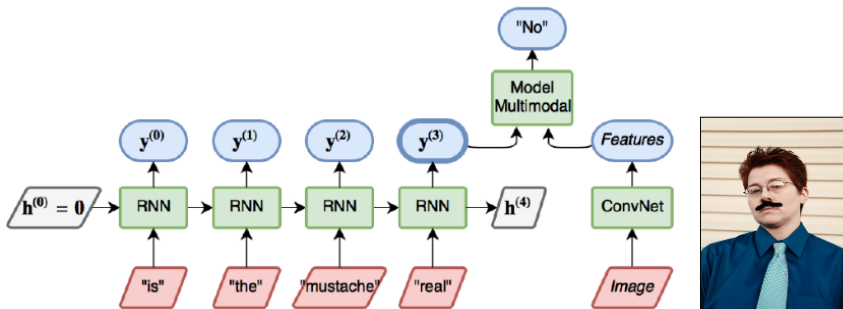


What color is the fire hydrant
on the right? yellow



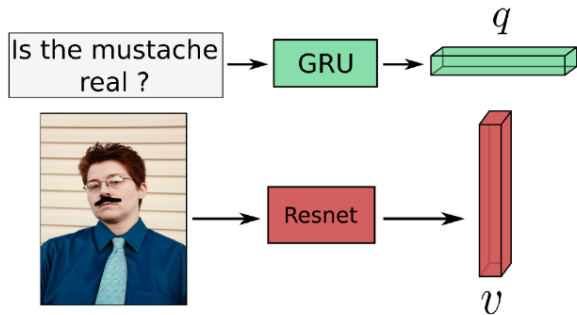
What color is the fire hydrant
on the left? green

Many to One - Visual Question Answering (VQA)



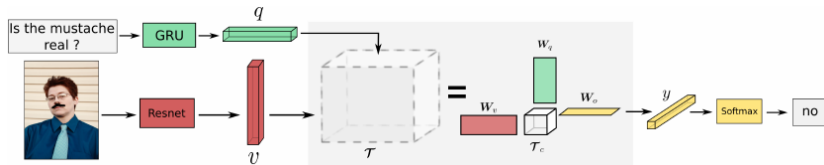
- ▶ Input: question & image
- ▶ Output: answer

Visual Question Answering

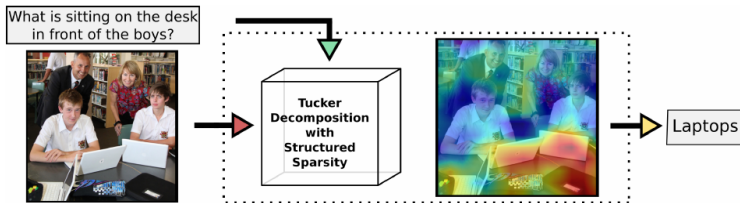


- ▶ State-of-the-art mono-modal representations:
 - ▶ Visual representation: ResNet-152
 - ▶ Question representation: pre-trained GRU (Gated Recurrent Units)

Multimodal Fusion for Visual Question Answering



- ▶ State-of-the-art: bilinear models [Fukui et al., 2016, Kim et al., 2017]
⇒ accurate interactions
- ▶ BUT full bilinear models intractable: factorization based on Tucker decomposition [Ben-younes et al., 2017]



Vision & Langage: Conclusion

References I



Ben-younes, H., Cadène, R., Cord, M., and Thome, N. (2017).
MUTAN: multimodal tucker fusion for visual question answering.
CoRR, abs/1705.06676.



Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016).
Multimodal compact bilinear pooling for visual question answering and visual grounding.
arXiv:1606.01847.



Karpathy, A. and Li, F. (2015).
Deep visual-semantic alignments for generating image descriptions.
In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3128–3137.



Kim, J.-H., On, K.-W., Kim, J., Ha, J.-W., and Zhang, B.-T. (2017).
Hadamard Product for Low-rank Bilinear Pooling.
In 5th International Conference on Learning Representations.



Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In Blei, D. and Bach, F., editors, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 2048–2057. JMLR Workshop and Conference Proceedings.