Using Images to Ground Machine Translation

Iacer Calixto

Institute for Logic, Language and Computation.
University of Amsterdam.

iacer.calixto@uva.nl

May 9, 2019

Outline

Introduction

Machine Translation (MT),
Image Description Generation (IDG), and
Multi-modal Machine Translation (MMT):

	MT	IDG	MMT
learn a model	✓	✓	✓
NLP vs. CV	NLP	NLP+CV	NLP+CV
generate a description	✓ ?	✓	✓✓
generate a translation	✓	✓ ?	✓✓
source/target pairs	✓	✗	✓ ?
source/target/image tuples	s,t	t,i	s,t,i
text-only vs. multi-modal	text-only	multi-modal	multi-modal

Multi-modal MT: Practical use cases

localisation of product information in e-commerce,
e.g. eBay, Amazon, Alibaba;

Multi-modal MT: Practical use cases

localisation of product information in e-commerce,
e.g. eBay, Amazon, Alibaba;

Image	Product Listing
	(en) apple macbook pro 13.3“ laptop - dvd - rw drive / good screen / airport card keyboar (de) apple macbook pro laptop 13.3“ - dvd - rw - laufwerk / gutes display / airport karte tastatur
	(en) modern napkin holder table top stainless steel weighted arm napkins paper towels (de) moderner tischserviettenhalter aus edelstahl mit beschwertem arm für servietten und papiertücher

Multi-modal MT: Practical use cases

localisation of user posts and photos in social media,
e.g. Twitter, Facebook, Instagram;

Multi-modal MT: Practical use cases

localisation of user posts and photos in social media,
e.g. Twitter, Facebook, Instagram;

Multi-modal MT: Practical use cases

translation of subtitles using video stream.

Multi-modal MT: Practical use cases

translation of subtitles using video stream.

Component	Content
Subtitle	Do you know which is our car?
Dialogue	Dad, do you know which is our car?
AD	Father and son walk along the platform in the railway station
Video snapshot	

Multi-modal MT: Practical use cases

and, of course, the most important of it all...

Multi-modal MT: Practical use cases

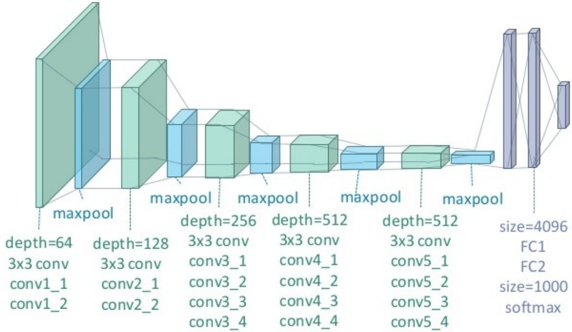
and, of course, the most important of it all...
MEMES' localisation

CNNs

Virtually all MMT and IDG models use pre-trained CNNs for image feature extraction;

CNNs

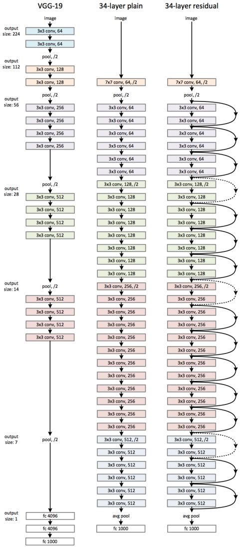
Virtually all MMT and IDG models use pre-trained CNNs for image feature extraction;



VGG 19 network

Simonyan and Zisserman (2014), `https://goo.gl/y0So1l`

CNN examples



https://goo.gl/jqQEvg

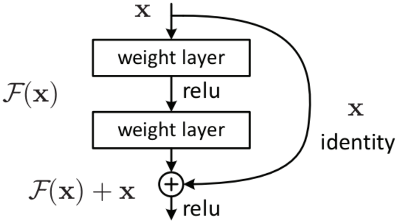
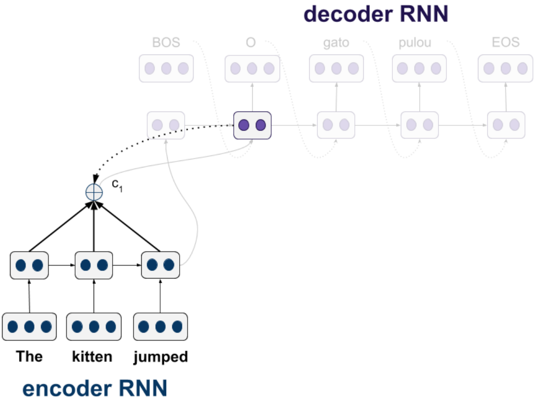


Illustration of a residual connection (He et al., 2015).

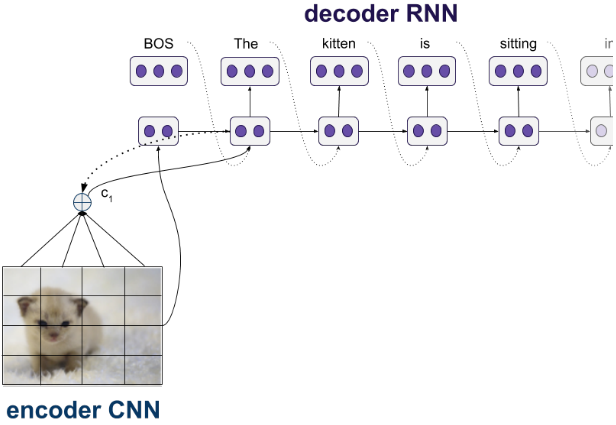
Neural Machine Translation

An attention mechanism lets the decoder search for the best source words to generate each target word, e.g. Bahdanau et al., 2015.

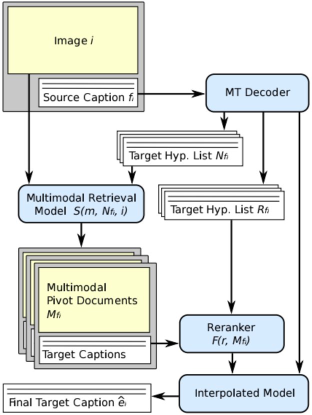


Neural Image Description Generation

An attention mechanism lets the decoder look at specific parts of the image when generating each target word, e.g. Xu et al., 2015.

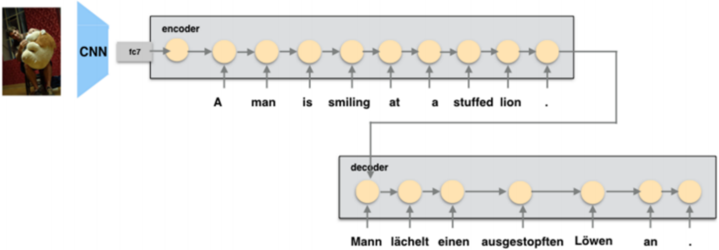


Heidelberg University



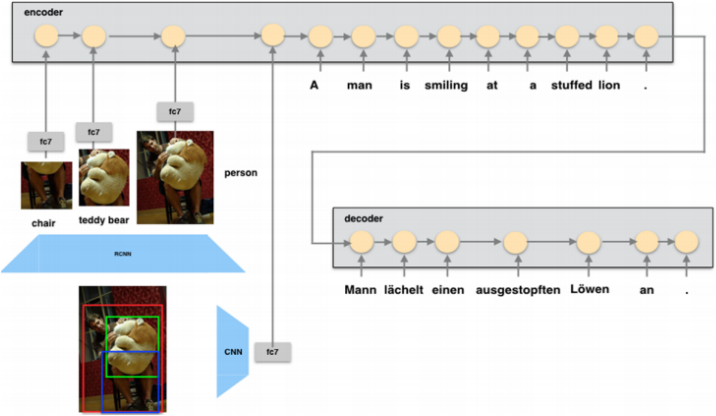
(Hitschler et al., 2016)

CMU [1/3]



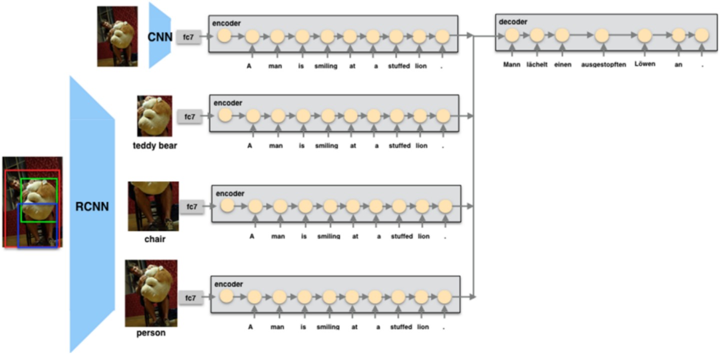
(Huang et al., 2016)

CMU [2/3]



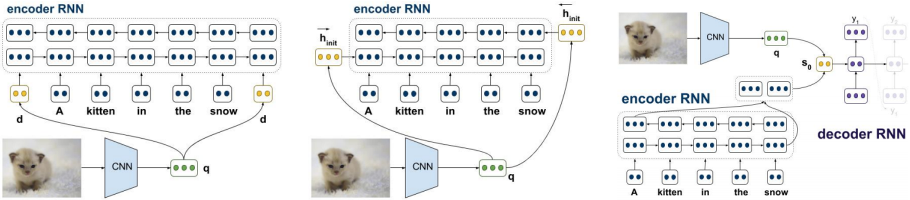
(Huang et al., 2016)

CMU [3/3]



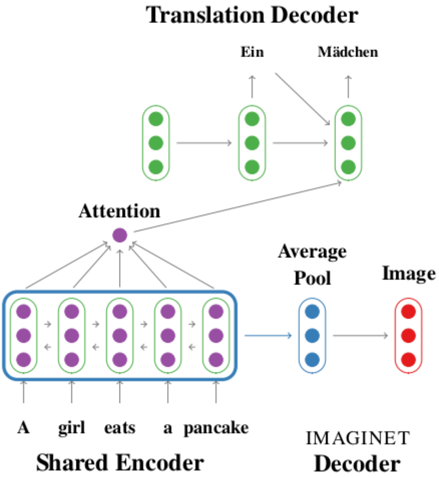
(Huang et al., 2016)

Global visual features



(Calixto et al., 2017)

UvA-TiCC



(Elliott and Kádár, 2017)

LIUM-CVC (Caglayan et al., 2017)

- Global visual features, i.e. 2048D `pool5` features from a ResNet-50 network, are multiplicatively interacted with the target word embeddings;

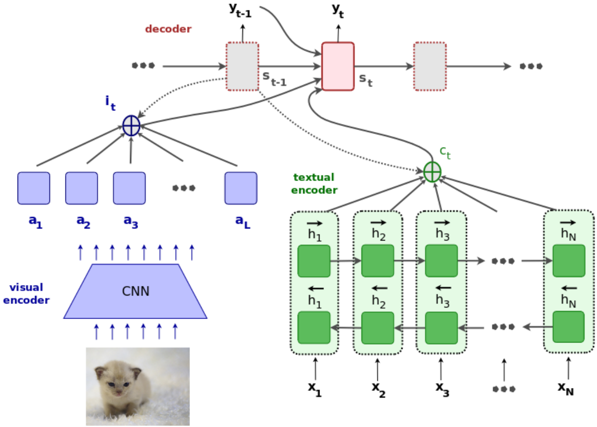
(Elliott et al., 2017)

LIUM-CVC (Caglayan et al., 2017)

- With 128D embeddings and 256D recurrent layers, their resulting models have ∼5M parameters.

(Elliott et al., 2017)

Doubly-Attentive MMT



(Calixto et al., 2017)

Conclusions

- multi-modal neural MT **use cases**;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;
- **visual attention** can be used as a tool for **model interpretability**;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;
- **visual attention** can be used as a tool for **model interpretability**;
- what's next?
  - **multi-task learning**, e.g. visual question answering;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;
- **visual attention** can be used as a tool for **model interpretability**;
- what's next?
    - **multi-task learning**, e.g. visual question answering;
    - **generative multi-modal MT** models;

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;
- **visual attention** can be used as a tool for **model interpretability**;
- what's next?
  - **multi-task learning**, e.g. visual question answering;
  - **generative multi-modal MT** models;
  - use images to ground models while being able to **translate sentences without images**?

Conclusions

- multi-modal neural MT **use cases**;
- **visually grounded** MT models;
- models that **efficiently exploit additional data** in pre-training;
- **visual attention** can be used as a tool for **model interpretability**;
- what's next?
  - **multi-task learning**, e.g. visual question answering;
  - **generative multi-modal MT** models;
  - use images to ground models while being able to **translate sentences without images**?
  - using **external knowledge** (i.e. multi-modal knowledge bases) in **end-to-end learning**;

References I

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations. ICLR 2015.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., and van de Weijer, J. (2017). LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pages 432–439.

Calixto, I., Liu, Q., and Campbell, N. (2017a). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In Proceedings of the 55th Conference of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1913–1924, Vancouver, Canada.

Calixto, I. and Liu, Q. (2017b). Incorporating Global Visual Features into Attention-based Neural Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1003–1014, Copenhagen, Denmark.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. In Proceedings of the 5th Workshop on Vision and Language, VL@ACL 2016, Berlin, Germany.

Elliott, D., Kádár, Á. (2017). Imagination improves Multimodal Translation. arXiv preprint arXiv:1705.04350.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385.

Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2399–2409, Berlin, Germany.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In Proceedings of the First Conference on Machine Translation, pages 639–645, Berlin, Germany.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

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.

Thank you!

Questions?