



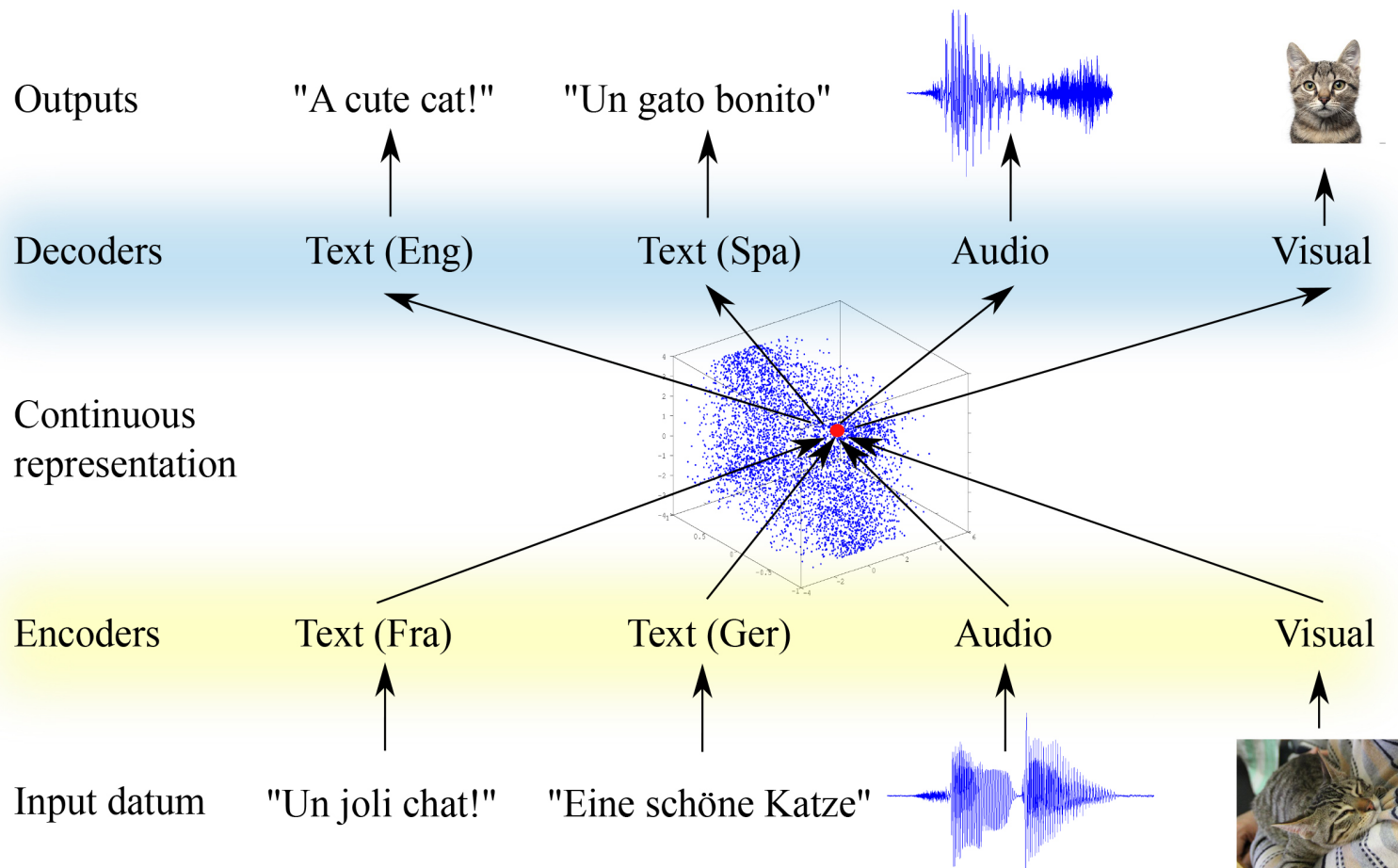
M2CR

Loïc BARRAULT, University of Le Mans

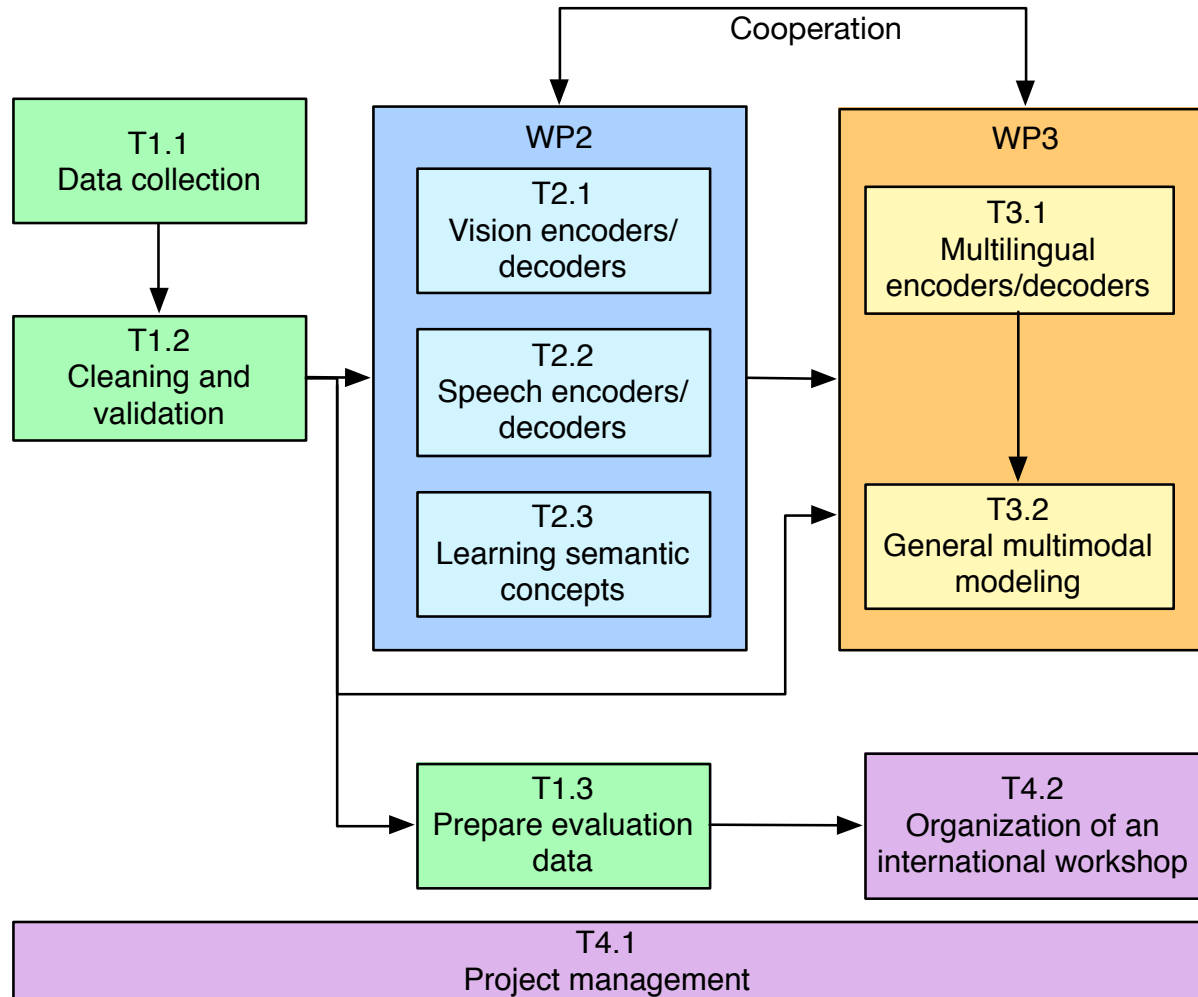
LIUM / LST

www.m2cr.univ-lemans.fr

Unified architecture



Overview



Modality specific encoders/decoders

- ▶ Text:
 - ▶ Encoder-decoder architecture with attention
 - ▶ Enc: bidirectional RNN
 - ▶ Dec: conditional GRU
- ▶ Machine Translation:
 - ▶ Extensive search of hyper training parameters
 - ▶ models need regularisation (e.g. weight decay, dropout)
 - ▶ [Caglayan et al, 2017]

Modality specific encoders/decoders

- Vision:
 - Several encoders for vision were evaluated
 - VGG, ResNet-50 and ResNet-152
 - used for Multimodal MT 2016 (details in following slides)
 - New image decoders based on GANs

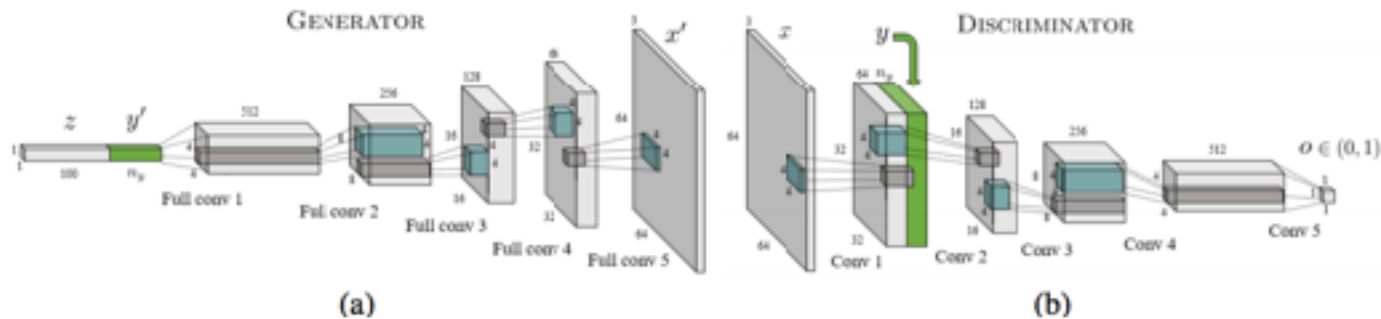
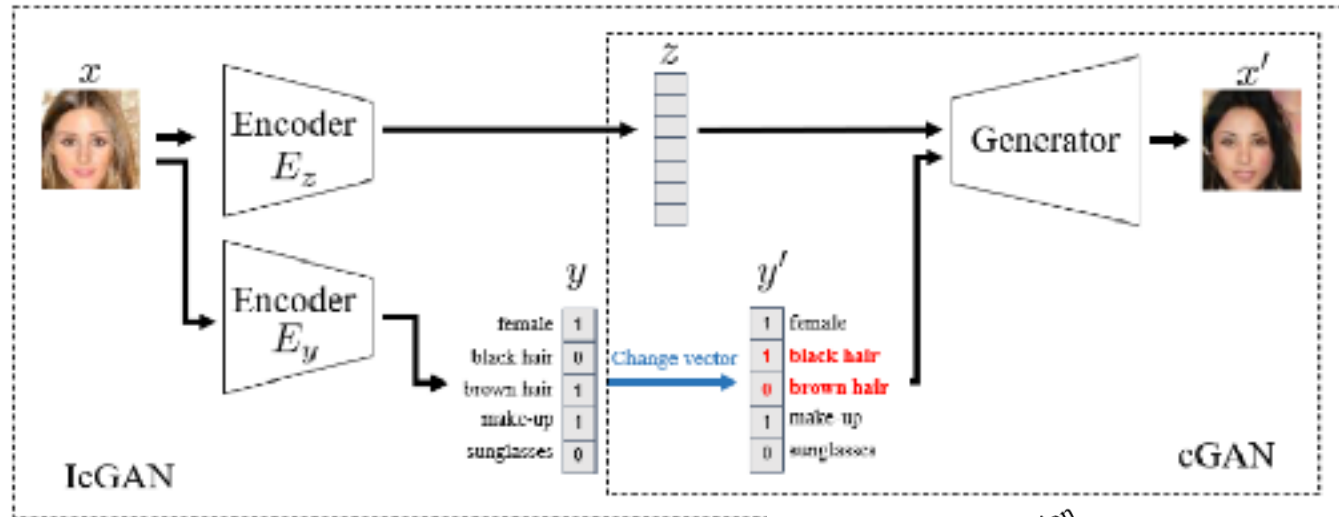


Figure 3: Architecture of the generator (a) and discriminator (b) of our cGAN model. The generator G takes as input both z and y . In the discriminator, y is concatenated in the first convolutional layer.

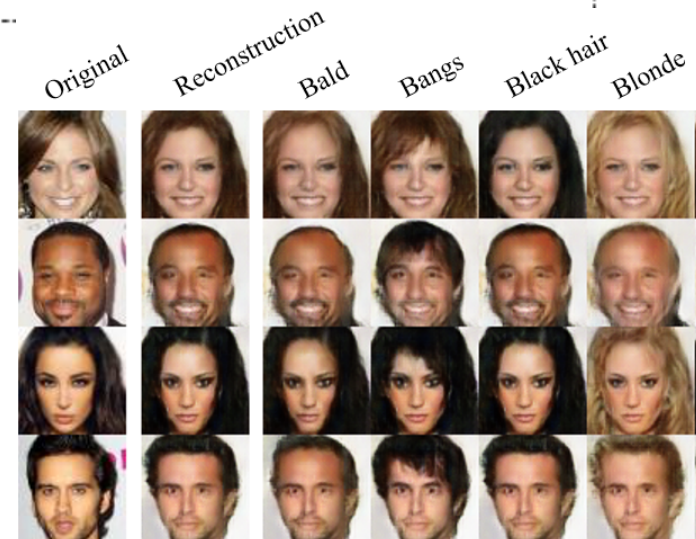
Generative Adversarial Networks

► IcGAN



- explicitly control attributes of generated samples with conditional information

► [Perarnau et al., 2016]



Modality specific encoders/decoders

- ▶ Speech to text:
 - ▶ Fully neural RNN-based ASR system, equipped with an attention mechanism
 - ▶ <https://github.com/rizar/attention-lvcsr>
- ▶ Sequence to sequence with attention mechanism
 - ▶ audio to characters
 - ▶ pure neural systems still not surpass hybrid (HMM+DNN) on LVCSR

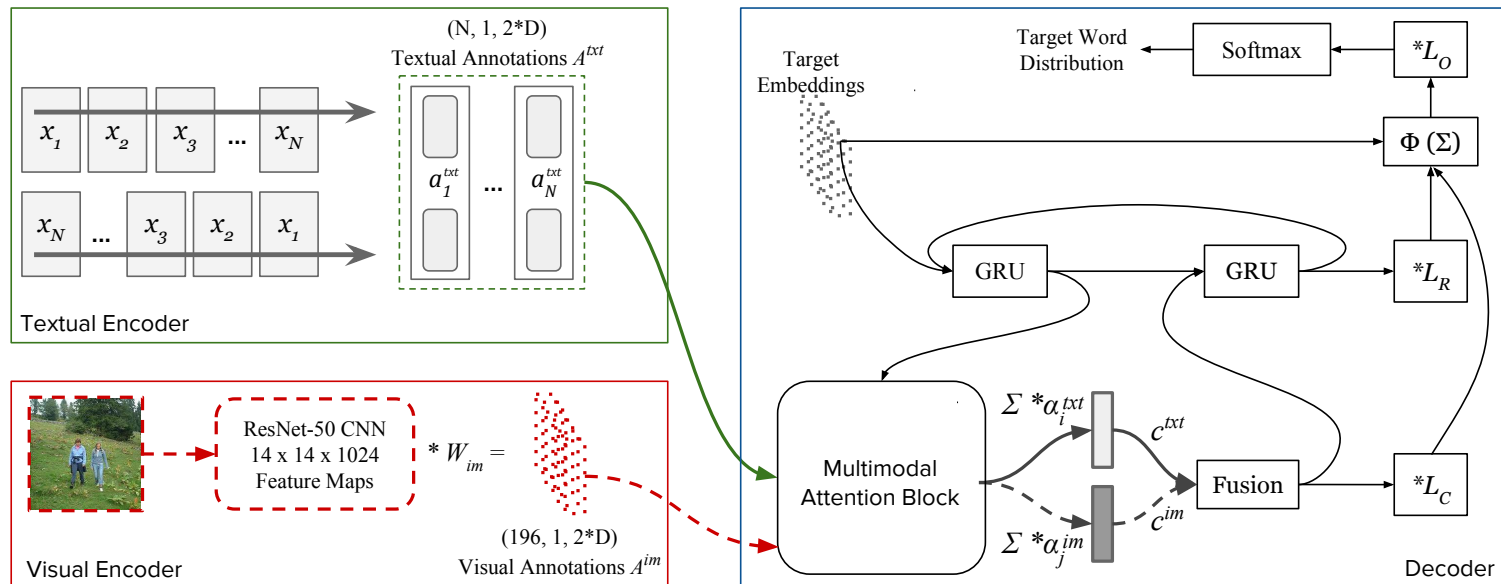
Modality specific encoders/decoders

- ▶ Spoken Language Understanding
 - ▶ Rigorous study of several types of word embeddings
 - ▶ at linguistic, semantic and acoustic level
 - ▶ [Ghannay et al., 2016a, 2016b]
- ▶ Comparison between CRF and Neural Networks
 - ▶ Fully neural system do not surpass CRF-based systems (yet?)
 - ▶ CRF more powerful at modelling sequences
 - ▶ [hot news!] Combination helps!
 - ▶ Both models have some complementarities

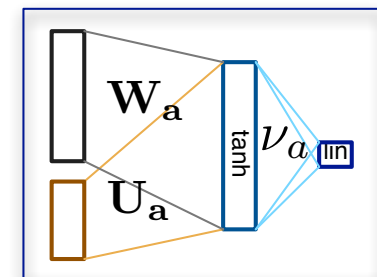
Multimodality

- ▶ Multimodal Machine Translation framework
 - ▶ **nmtpy**, a framework for mono- and multi-modal NMT systems
 - ▶ <https://github.com/lium-lst/nmtpy>
 - ▶ LIUM+CVC won the 1st MMT evaluation campaign
 - ▶ [Caglayan et al., 2016]
- ▶ M2CR is organising MMT17 evaluation campaign
 - ▶ Data creation:
 - ▶ Extension of Multi30k with French Language
 - ▶ 31k image descriptions in English translated into French
 - ▶ tri-lingual (En, De, Fr) bi-modal (image, text) corpus
 - ▶ Image can contextualise and disambiguate translations

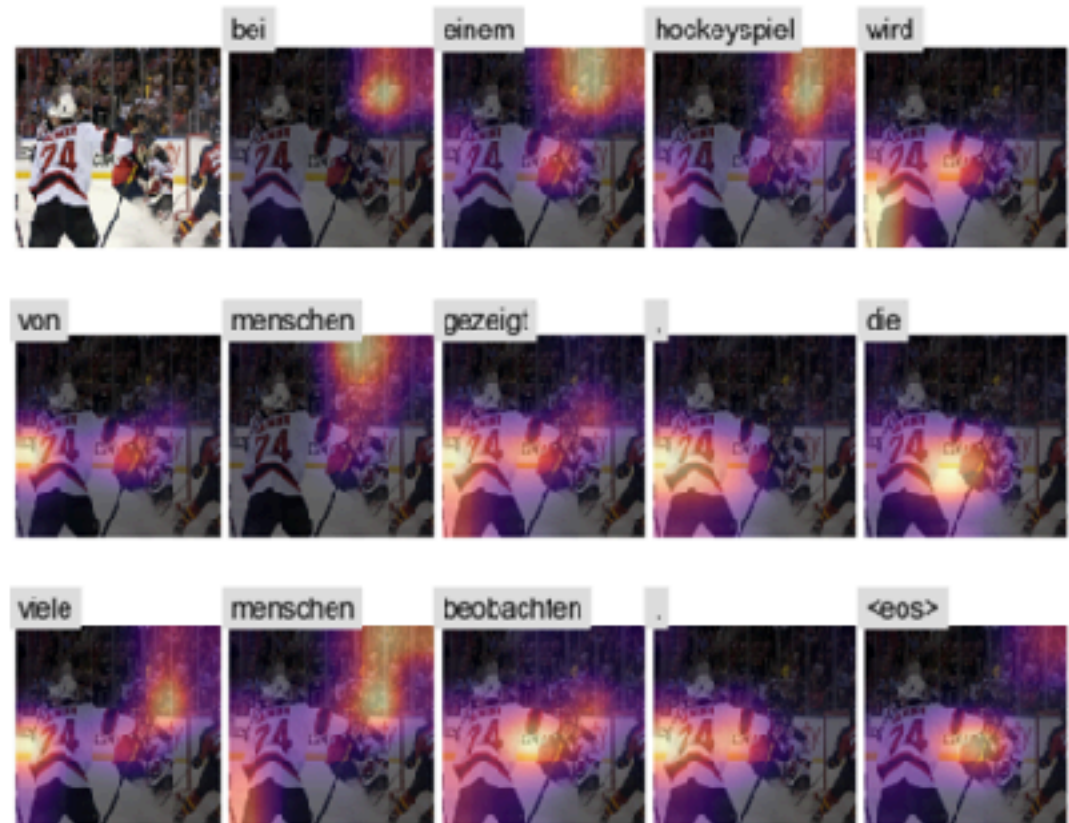
Multimodality



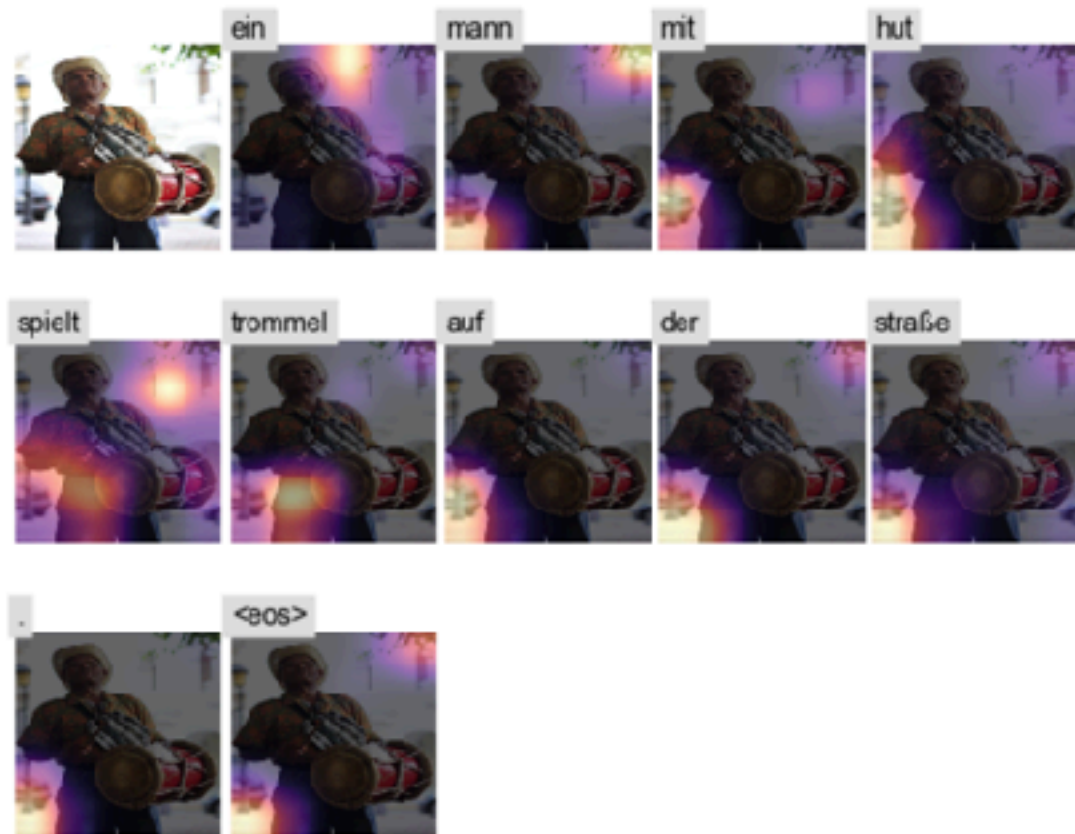
- ▶ Enc. image : ResNet-50 (precomputed)
- ▶ Enc. text: biRNN (trained jointly)
- ▶ Att mechanism: simple MLP
- ▶ Dec: conditional GRU
 - ▶ feedback ($\sim$ LM), context fusion



Multimodal alignment



Multimodal alignment



Multimodal Machine Translation

- ▶ Image attention still not satisfying
 - ▶ possible causes:
 - ▶ sequence length
 - ▶ encoder pre-trained
- ▶ Text attention is good
 - ▶ encoder is jointly trained
 - ▶ MT task: word contain more **specific** information than image
 - ▶ image is far more ambiguous

Ongoing/next work

- ▶ Improving multimodal attention
 - ▶ sometimes we don't want to give attention to a specific modality
 - ▶ e.g. tool words are not shown in an image
 - ▶ Actual attention mechanism always provide weights
 - ▶ between 0 and 1, summing to 1 (prob. distribution)
 - ▶ Enhance attention with additional data
- ▶ Exploring image feature capturing the 'scene'
 - ▶ corpus "Places"
 - ▶ closer to Flickr data?
- ▶ Extract objects from image related to the input words
 - ▶ only use relevant part of image

Ongoing/next work

- ▶ Towards multitask learning
 - ▶ multi-source (languages and modalities)
 - ▶ multiple tasks
 - ▶ MT, language modelling, image description, SLU
 - ▶ Goal: multi-purpose representations

Publications

- ▶ [Caglayan et al, 2016] Caglayan, O., Barrault, L., and Bougares, F. Cross-modal Attention for Image Captioning, 2016
- ▶ [Ghannay et al. 2016a] Ghannay, S., Favre, B., Estève, Y., Camelin, N. Word embedding evaluation and combination. 10th edition of the Language Resources and Evaluation Conference (LREC 2016), Portorož (Slovenia), 2016
- ▶ [Ghannay et al. 2016b] Ghannay, S., Estève, Y., Camelin, N., Deléglise, P. Evaluation of acoustic word embeddings. RepEval@ACL 2016: The 1st Workshop on Evaluating Vector-Space Representations for NLP , Berlin (Germany), 2016
- ▶ [Simonnet et al, 2016] Simonnet, E., Deléglise, P., Camelin, N., Estève, Y. Des Réseaux de Neurones avec Mécanisme d'Attention pour la Compréhension de la Parole. 31ème Journées d'Études sur la Parole, Paris (France), 2016

Publications

- ▶ [Caglayan et al., 2016] Caglayan, O., Aransa, W., Wang, Y., Masana, M., Garcia-Martinez, M., Bougares, F., Barrault, L., and van de Weijer, J. Does multimodality help human and machine for translation and image captioning?, in First conference on machine translation (WMT), 2016.
- ▶ [Perarnau et al., 2016] Perarnau, G., van de Weijer, J., Raducanu, B. and Álvarez M. J. Invertible Conditional GANs, in Workshop on Adversarial Training, NIPS 2016
- ▶ [Wang et al., 2016] Wang, Y., Zhang, L., van de Weijer, J., Ensembles of Generative Adversarial Networks, Workshop on Adversarial Training, NIPS 2016, Barcelona, Spain.
- ▶ [Caglayan et al, 2017] Caglayan, O., Barrault, L., and Bougares, F. An Extensive Hyperparameter Search Towards Smaller Neural Machine Translation Models, 2017

Partners

- ▶ LIUM, University of Le Mans, France, coordinator
 - ▶ L. Barrault, F. Bougares, Y. Esteve, P. Deléglise, N. Camelin
 - ▶ W. Aransa, A. Bardet, O. Caglayan, M. Garcia-Martinez, S. Ghannay, E. Simonnet

- ▶ MILA, Montreal, Quebec, Canada
 - ▶ Y. Bengio, P. Brakel

- ▶ CVC-LAMP, Barcelona, Spain
 - ▶ J. Van de Weijer, Y. Wang, M. Masana, L. Herranz

