

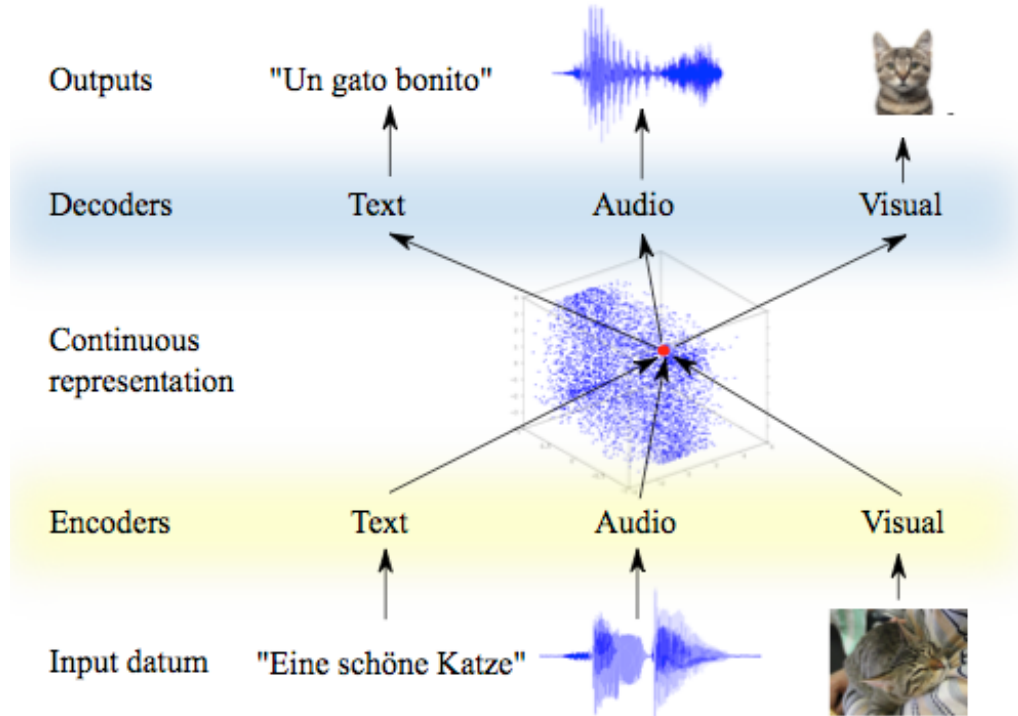


M2CR – Multimodal Multilingual Continuous Representations for HLU

Loïc BARRAULT, University of Le Mans
LIUM / Team Language and Speech Technology

<https://projets-lium.univ-lemans.fr/m2cr/>

Unified end-to-end architecture



Achievements 1/2

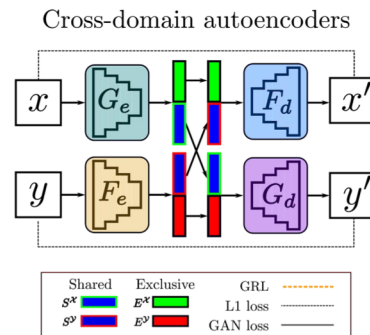
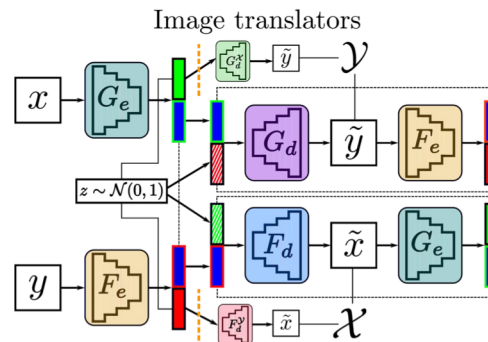
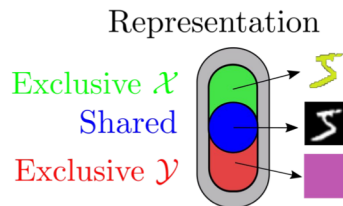
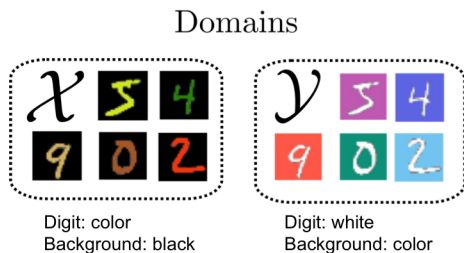
- **New models for NLU tasks**
 - Multimodal MT, ASR, SLT and Speech synthesis
 - Image translation across domains
 - End-to-end SLU system
 - Single task and Multitask learning
- **Datasets**
 - Multi30k
 - HOW2

Achievements 2/2

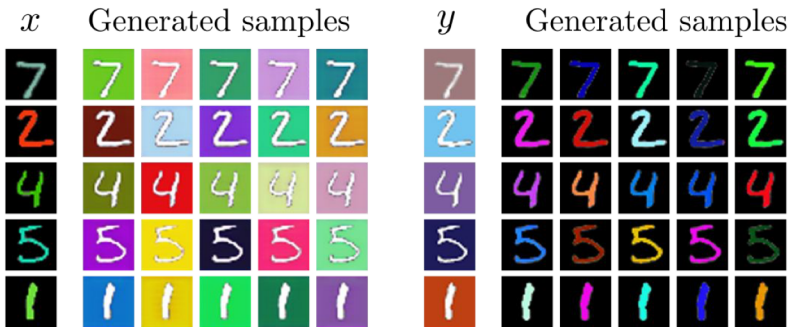
- **Open source tools**
 - nmtpytorch framework
 - domain disentanglement, mix-and-match networks and transfer learning for GANs
 - Theano based developments for e.g. end-to-end ASR
- **Events**
 - Evaluation campaigns
 - HLU MasterClass
 - JHU JSALT summer workshop

Representing modalities

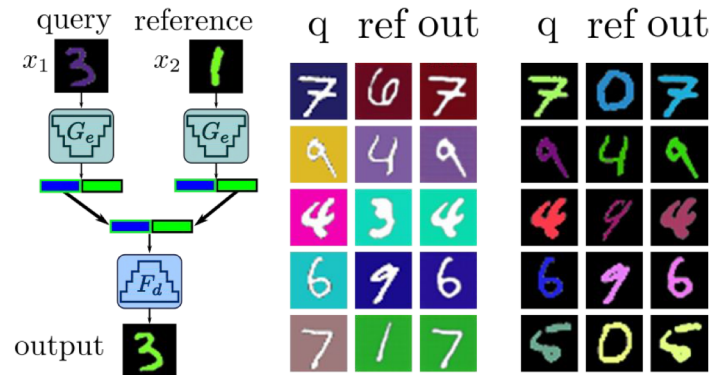
Cross-domain disentanglement



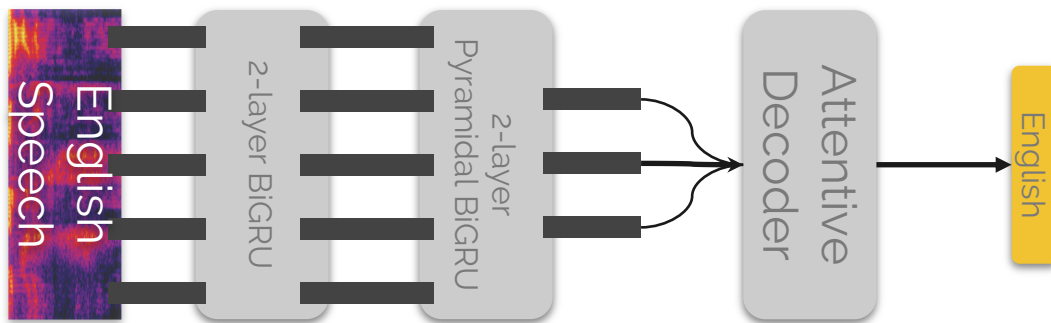
Vision: structuring the continuous space



Many-to-many image translation through latent space manipulation



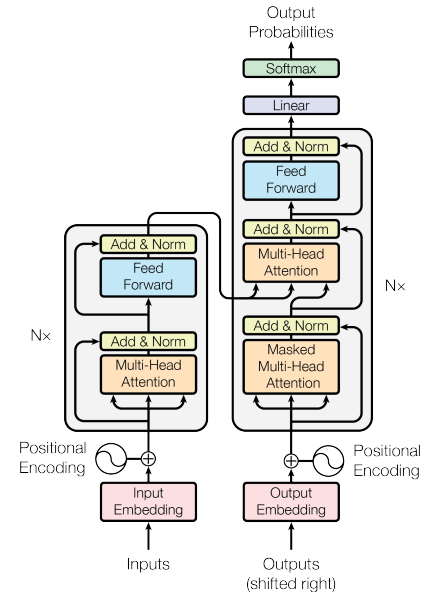
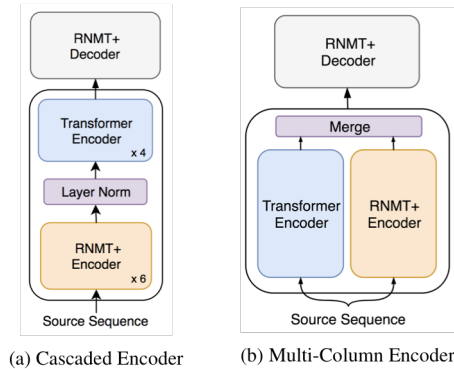
Speech encoder



- ☐ 4-Layer BiGRU Encoder (200D)
- ☐ 200D Embeddings
- ☐ 2-Layer Conditional GRU Decoder
- ☐ MLP Attention
- ☐ Dropout ($p=0.4$)

Text: Recurrent & Transformer models

- Field is evolving very fast!

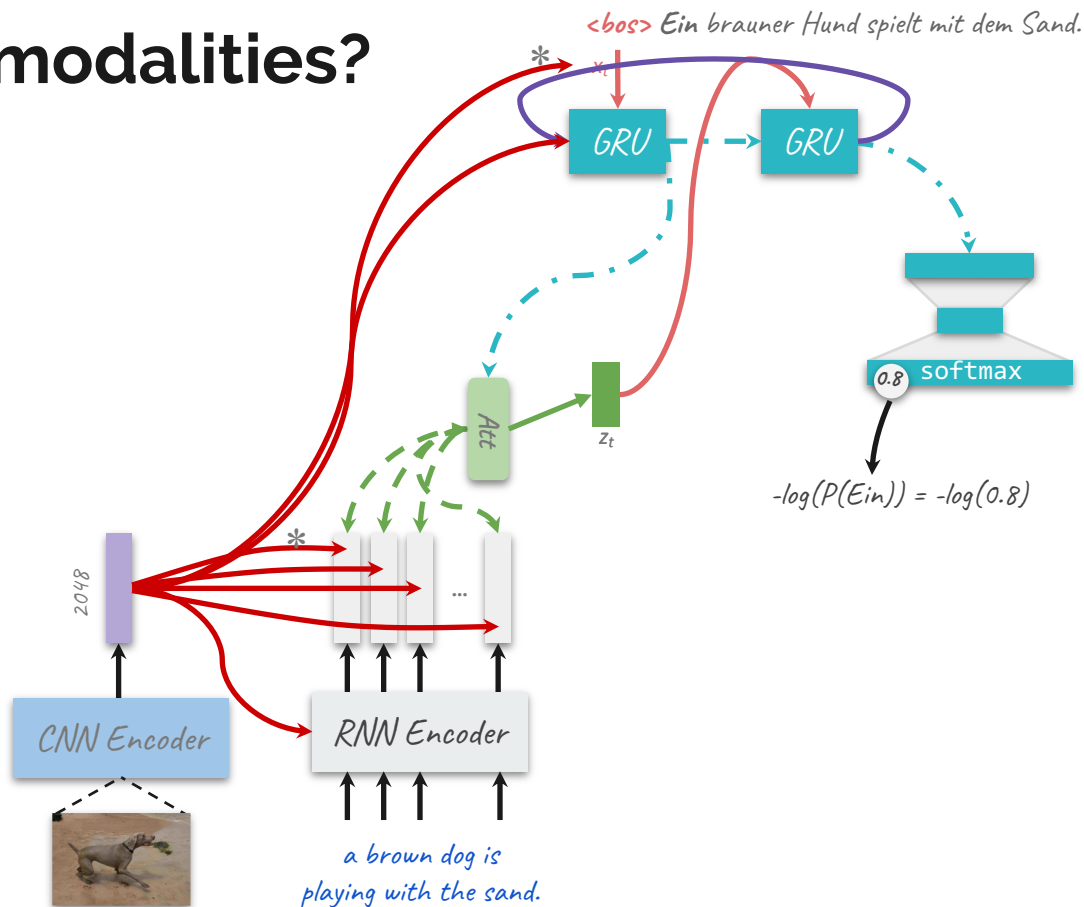


- Chen et al. 2018, The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation
- Vaswani et al. 2017. Attention is all you need.

Dealing with multiple modalities

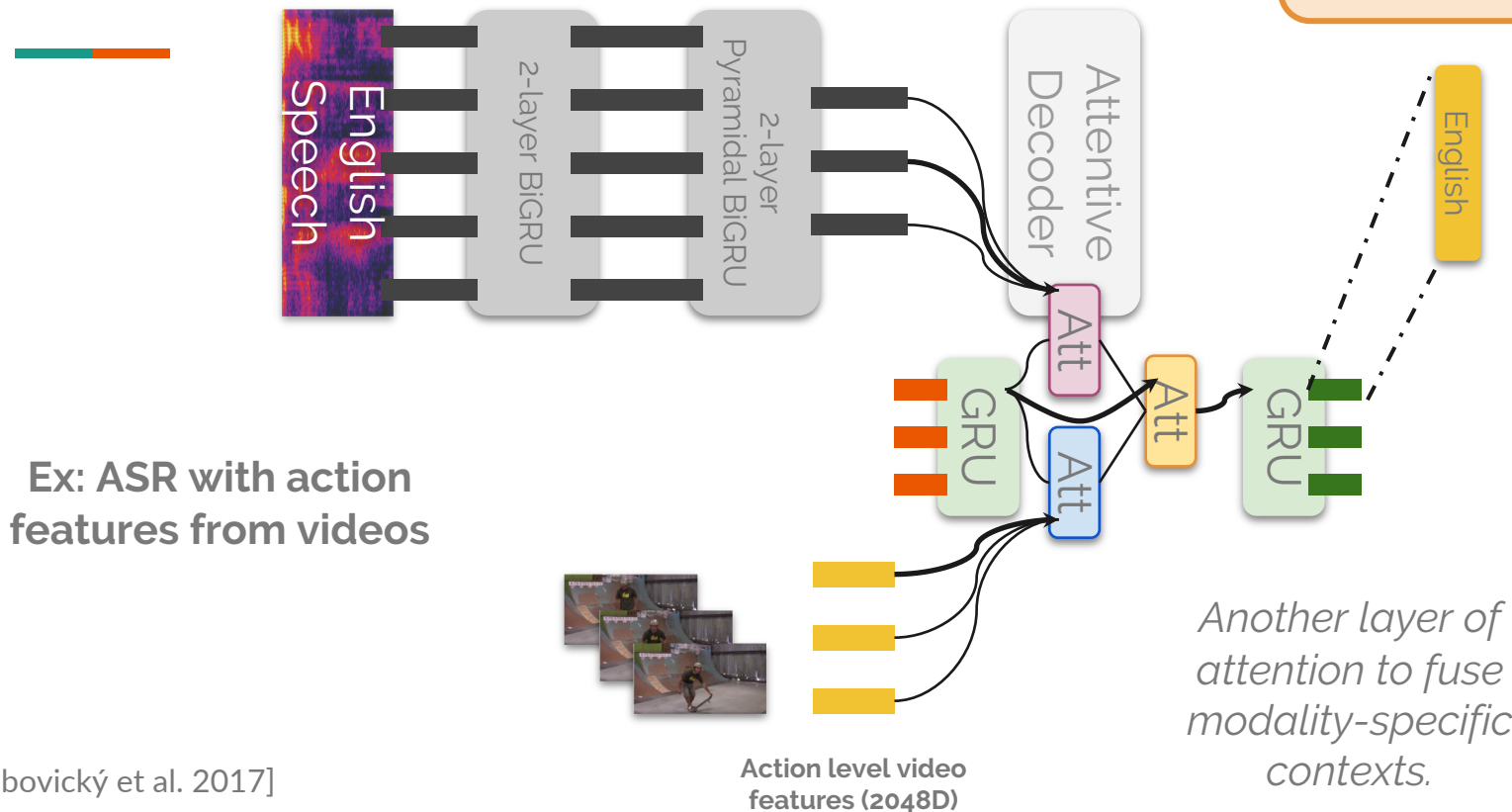
How to fuse two modalities?

- Ex: Multimodal MT



Hierarchical Attention

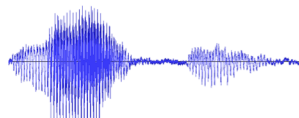
Motivation: Can we benefit from selective multimodal attention?

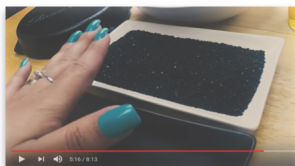


[Libovický et al. 2017]

Multitask learning (MTL)

*So as you can see I added some
sesame seed, some black sesame
seed here in my plate*
Subtitle


Speech signal



Keyframe / Video

*Text
Encoder*

*Speech
Encoder*

*Visual
Encoder*

Translation

*Como vocês podem ver, eu
coloquei no meu prato o
gergelim preto*

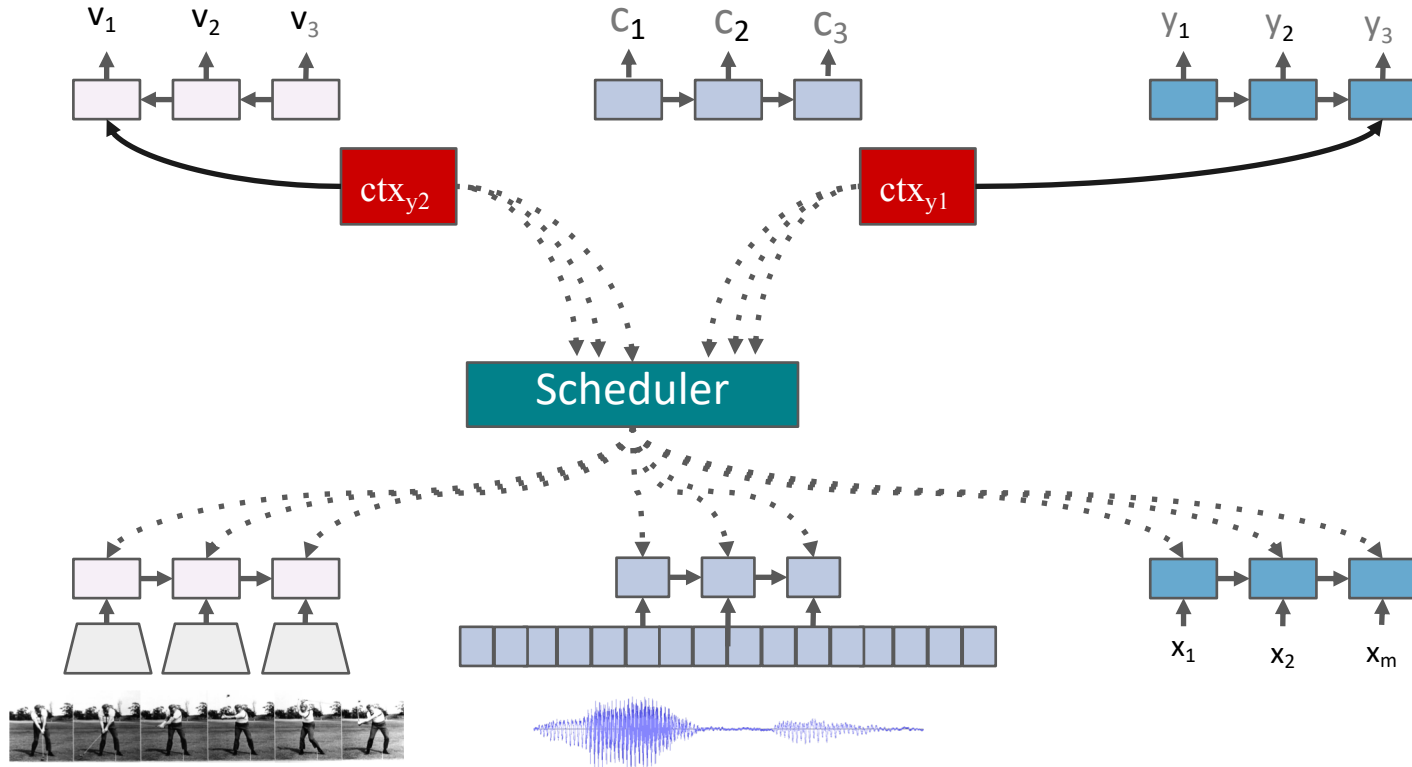
Transcription

*So as you can see I added some
sesame seed, some black sesame
seed here in my plate*

Summary

*A cooking recipe for Seared
Sesame Crusted Tuna with Wild
Rice*

MTL by scheduling tasks



Datasets and open source tools

Tools, data and shared task

- Multimodal Machine Translation framework
 - **nmtpy** -> **nmtpytorch**: a framework for mono- and multi-modal NMT, ASR and SLT systems

nmtpytorch

MMT data: Multi30k

- 30k images from Flickr
- 4 languages and 2 modalities
 - All are description of images
- Pb.: task is (too) easy
- Need more challenge!



Descriptions

EN: A boy dives into a pool near a water slide.

DE: Ein Junge taucht in der Nähe einer Wasserrutsche in ein Schwimmbecken.

FR: Un garçon plonge dans une piscine près d'un toboggan.

CS: Chlapec skáče do bazénu poblíž skluzavky.

HOW2 dataset

- 2000 hours of **how-to** videos
 - 480 for MT, all for ASR
- Manually transcribed
- Metadata available
 - Nb. of likes/dislikes
 - Nb. of visualizations
 - Uploader, date, tags
- 80k video descriptions (i.e. “summaries”)
- 300k segments translated into Portuguese
- Very different topics
 - Cooking, fixing things, playing instruments, etc.



MOstrar más

Dissemination

HLU Masterclass & Evaluation campaigns

- HLU masterclass: share knowledge among projects in the call
- Multimodal Machine translation
 - 2016, 2017, 2018,

HLU Masterclass & Evaluation campaigns

- This year:
 - ICML'19 : HOW2 challenge
 - Define new challenging tasks based on HOW2 data
 - Evaluation campaign on ASR, MT, SLT etc..
 - <https://srvk.github.io/how2-challenge/>
 - IWSLT'19 : Multimodal Spoken Language Translation
 - To be announced

JSALT Grounded Seq2seq transduction summer workshop

- M2CR participated in the F. Jelinek JHU summer workshop
- Developed Multimodal and/or Multitask learning for NLP
 - Using the nmtpytorch framework
- Related publications:
 - Caglayan O., Madhyasta P., Specia L. and Barrault, L., **Probing the Need for Visual Context in Multimodal Machine Translation**, NAACL'19
 - Caglayan O. Sanabria R., Palaskar S., Barrault L. and Metze F. **Multimodal Grounding for Sequence-to-Sequence Speech Recognition**, ICASSP'19

JSALT Grounded Seq2seq transduction summer workshop



Carnegie Mellon University
Language Technologies Institute



multiMT

amazon

Le Mans
Université



Christian Fuegen



facebook
Google

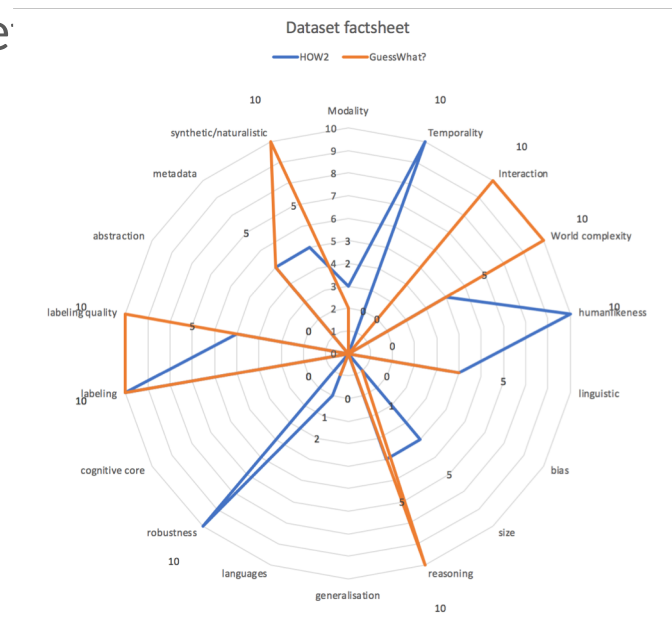
Ongoing/next work

Ongoing/next work

- Project end is in June.
- Transfer learning for multilingual NMT (Ongoing work)
- Focus on even more challenging tasks
 - Cf. M. Lieberman talk this morning
 - HOW2 corpus and tasks
 - Video understanding, video summarization
 - Language Grounding

Ongoing/next work

- Initiative to create datasets factsheet
 - Corpus characteristics
 - Design future NLU tasks



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, K. Kastner, P. Brakel
- CVC-LAMP, Barcelona, Spain
 - J. Van de Weijer, Y. Wang, M. Masana, L. Herranz



QUESTIONS?

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
- **[Caglayan2017b]** O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez , F. Bougares, L. Barrault, M. Masana, L. Herranz, J. van de Weijer, LIUM-CVC Submissions for WMT17 Multimodal Translation Task, Second Conference on Machine Translation, 2017

Publications

- R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, “**How2: A Large-scale Dataset For Multimodal Language Understanding**”, in Workshop on Visually Grounded Interaction and Language (ViGIL)}, NeurIPS, 2018, <http://arxiv.org/abs/1811.00347>
- S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent and E. Morin. ***End-to-end named entity and semantic concept extraction from speech***. IEEE Spoken Language Technology Workshop, Dec 2018, Athens, Greece. <https://hal.archives-ouvertes.fr/hal-01987740/document>
- L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott and S. Frank, ***Findings of the Third Shared Task on Multimodal Machine Translation***, WMT, 2018 <http://aclweb.org/anthology/W18-6402>
- A. Gonzalez-Garcia, J. van de Weijer, Y. Bengio, ***Image-to-image translation for cross-domain disentanglement***, Advances in Neural Information Processing Systems (NeurIPS), NIPS, 2018. <https://arxiv.org/abs/1805.09730> [Publication of two partners: CVC and MILA]
- X Liu, J. van de Weijer, A.D. Bagdanov, ***Leveraging Unlabeled Data for Crowd Counting by Learning to Rank***, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) <https://arxiv.org/abs/1803.03095>

Publications

- O. Caglayan, A. Bardet, F. Bougares, L. Barrault, K. Wang, M. Masana, L. Herranz, J. van de Weijer, **LIUM-CVC Submissions for WMT18 Multimodal Translation Task**, Third Conference on Machine Translation, 2018 [Publication of two partners: LIUM and CVC]
- O. Caglayan, R. Sanabria, S. Palaskar, L. Barrault and F. Metze, **Multimodal grounding for sequence-to-sequence speech recognition**, ICASSP 2019. <https://arxiv.org/abs/1811.03865>
- O. Caglayan, P. Madhyastha, L. Specia and L. Barrault, **Probing the Need for Visual Context in Multimodal Machine Translation**, NAACL, 2019
- R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, “**How2: A Large-scale Dataset For Multimodal Language Understanding**”, in Workshop on Visually Grounded Interaction and Language (ViGIL), NeurIPS, 2018, <http://arxiv.org/abs/1811.00347>
- Y. Wang, J. van de Weijer, L. Herranz, **Mix and match networks: encoder-decoder alignment for zero-pair image translation**, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) <https://arxiv.org/abs/1804.02199>
- Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, B. Raducanu, **Transferring GANs: generating images from limited data**, European Conference of Computer Vision, 2019. <https://arxiv.org/abs/1805.01677>
- C. Wu, L. Herranz, X. Liu, J. van de Weijer, B. Raducanu, **Memory Replay GANs: Learning to Generate New Categories without Forgetting**, Advances in Neural Information Processing Systems, 2018. <https://arxiv.org/abs/1809.02058>