# Multimodal Machine Translation

Loïc Barrault

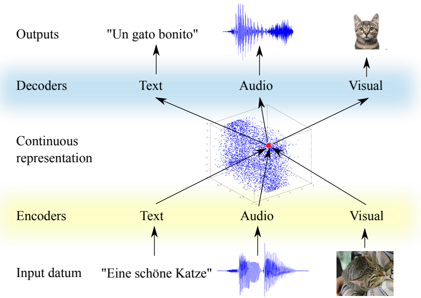
loic.barrault@univ-lemans.fr
Le Mans Université

# Overview

- Introduction / M2CR project
- Multi30k: a multilingual multimodal corpus
- Neural Machine Translation
- Dealing with images
- Multimodal MT

# M2CR project

- Create a unified framework to learn a shared space
- Represent (encode) various modalities
- Decode from it towards any other language or modality.



# Motivation

- Semantics still poorly used in MT systems
    - Embeddings seem to convey such information

- Can meaning be modelled from text only?
    - Can't learn everything from books!

  → Language grounding

  → Use of multiple modalities

- Intermediate step: use visual information to disambiguate translation

## Example 1: morphology

- A baseball player in a black shirt just tagged a player in a white shirt.

# Example 1: morphology

- A baseball player in a black shirt just tagged a player in a white shirt.
- Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.

# Example 1: morphology

- A baseball player in a black shirt just tagged a player in a white shirt.
- Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.
- Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.

# Example 2: semantics

- A woman sitting on a very large rock smiling at the camera with trees in the background.

# Example 2: semantics

- A woman sitting on a very large rock smiling at the camera with trees in the background.
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Felsen und lächelt in die Kamera.
  - Felsen == stone (uncountable)

## Example 2: semantics

- A woman sitting on a very large rock smiling at the camera with trees in the background.
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Felsen und lächelt in die Kamera.
  - Felsen == stone (uncountable)
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Stein und lächelt in die Kamera.
  - Stein == rock (individual stone)

# Multi30k: Multimodal Multilingual Corpus

# Multi30k

- Extension of Flickr30k corpus [Plummer et al., 2017]
- Flickr30k: images from Flickr with crowdsourced English descriptions
- Multi30k: translating English descriptions into German

→ context of Multimodal Machine Translation (MMT'16)

- MMT'17: add French translations
- MMT'18: add Czech translations

- `http://www.statmt.org/wmt18/multimodal-task.html`

# Multi30k: example

**Descriptions**

EN: A ballet class of five girls jumping in sequence.

DE: Eine Ballettklasse mit fünf Mädchen, die nacheinander springen.

FR: Une classe de ballet, composée de cinq filles, sautent en cadence.

CS: Baletní třída pěti dívek skákající v řadě.

# Multi30k: statistics

| Corpus | #sents. | #w. EN | #w. DE | #w. FR | #w. CS |
|---|---|---|---|---|---|
| Train | 29k | 345.0k | 322.4k | 362.0k | 262.5k |
| Val | 1014 | 12.2k | 11.6k | 12.7k | 9.1k |
| Test2016 | 1000 | 11.9k | 10.9k | 12.3k | 9.3k |
| Test2017 | 1000 | 10.5 | 9.6k | 11.2k | - |
| **Total** | 32k | 379.6k | 354.5k | 398.3k | 280.9k |

# Multi30k: some comments

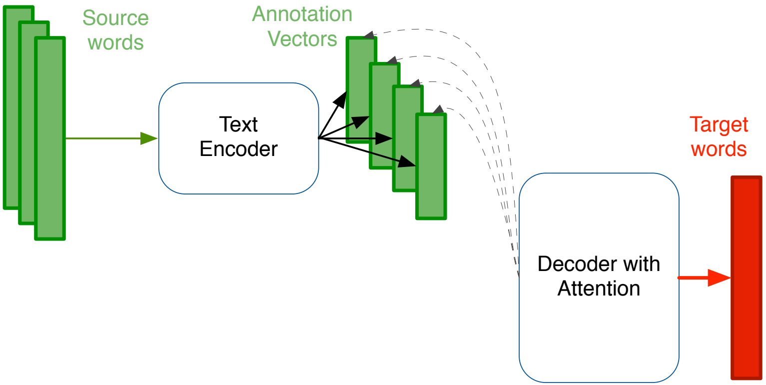
- Descriptions are simple
  - *A man ..., A woman ...,*
- Ongoing:
  - Create more complex examples
  - Visual information should be required to translate the source sentence

  → more ambiguity

  → complex to collect

# Neural Machine Translation

# Neural Machine Translation

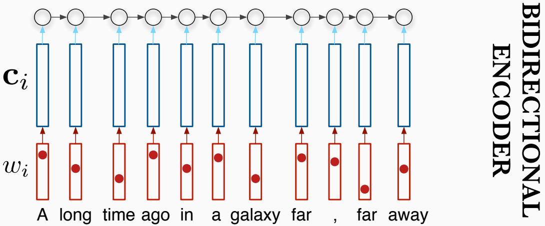



[Bahdanau et al., 2014]

# Bidirectional Encoder

- Previous work → fixed size vector is not be enough to represent a sentence

→ let's use several representations + process the sentence in both directions!

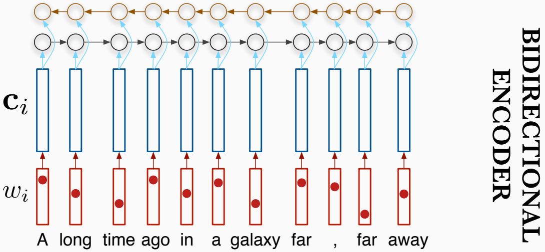



[1.] *1-hot* vector + projection + update forward RNN hidden state

# Bidirectional Encoder

- Previous work → fixed size vector is not be enough to represent a sentence

→ let's use several representations + process the sentence in both directions!

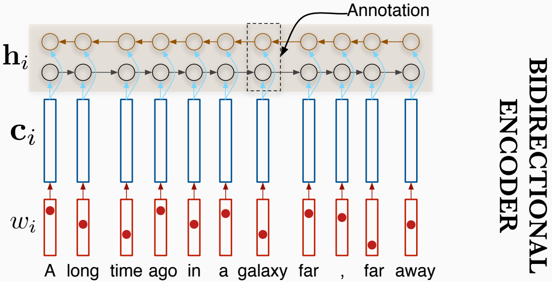



[1bis.] update backward RNN hidden state

# Bidirectional Encoder

- Previous work → fixed size vector is not be enough to represent a sentence

→ let's use several representations + process the sentence in both directions!



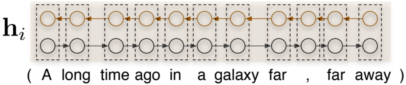

[2.] Annotation = concatenation of forward and backward vectors

Every $\mathbf{h}_i$ encodes the whole source sentence with a focus on the $i^{th}$ word

# Decoder with attention

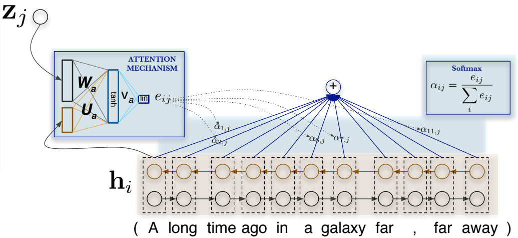
- [2.] Decoder gets the annotations.

DECODER



# Decoder with attention

- [2.] Decoder gets the annotations.
- [3.] Attention weights are computed with feedforward NN.
  $\rightarrow$ *weighted mean* $\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$

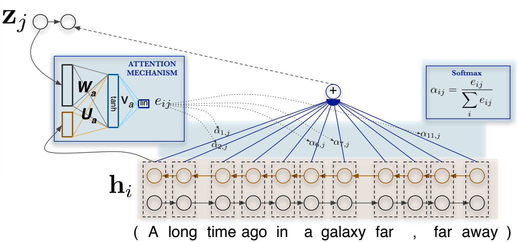



DECODER

# Decoder with attention

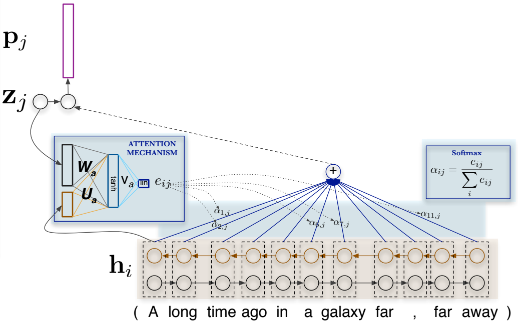
- [2.] Decoder gets the annotations.
- [3.] Attention weights are computed with feedforward NN.
  $\rightarrow$ *weighted mean* $\tilde{\mathbf{h_j}} = \sum_i \alpha_{ij}\mathbf{h}_i$
- [4.] Update hidden state of GRU

DECODER



# Decoder with attention

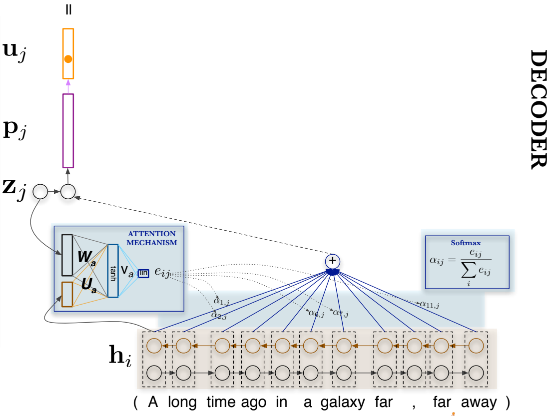
- [2.] Decoder gets the annotations.
- [3.] Attention weights are computed with feedforward NN.
  $\rightarrow$ *weighted mean* $\tilde{\mathbf{h}}_\mathbf{j} = \sum_i \alpha_{ij} \mathbf{h}_i$
- [4.] Update hidden state of GRU
- [5.] Probability distribution for all words



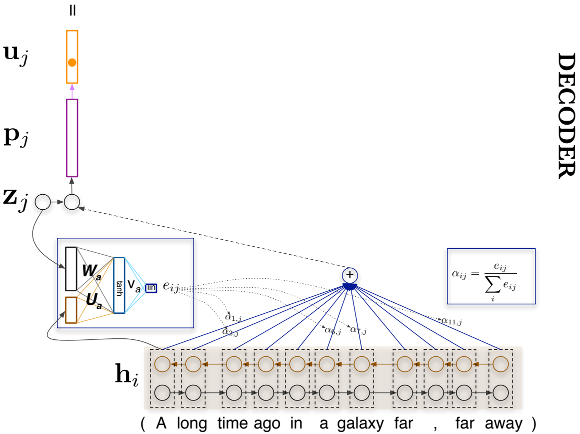

DECODER

# Decoder with attention

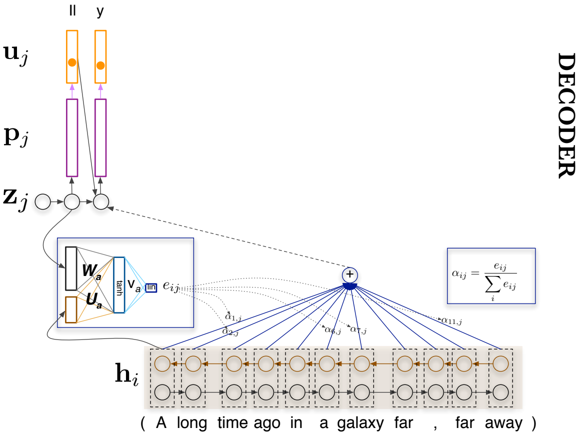
- [2.] Decoder gets the annotations.
- [3.] Attention weights are computed with feedforward NN.
  → *weighted mean* $\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$
- [4.] Update hidden state of GRU
- [5.] Probability distribution for all words
- [6.] Generate next word
  → most probable or beam



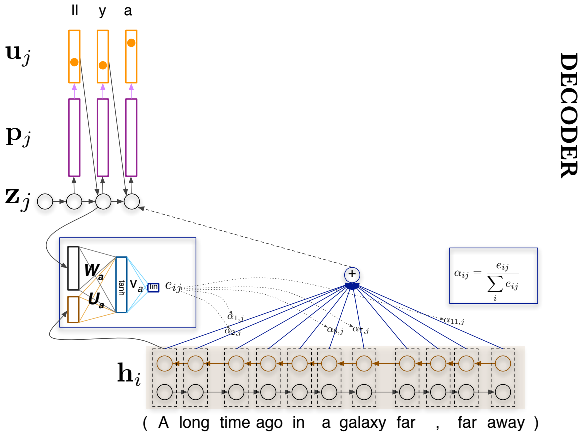
# Decoder with attention



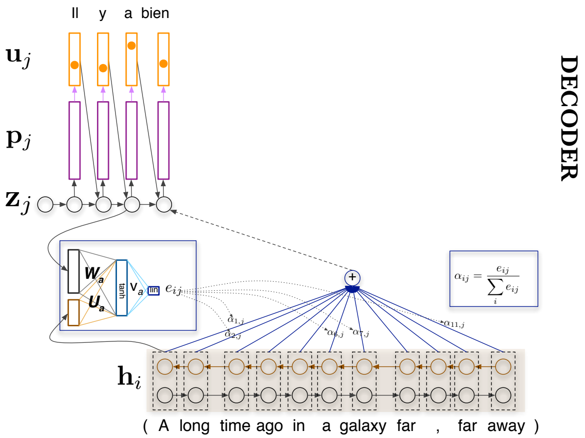
# Decoder with attention



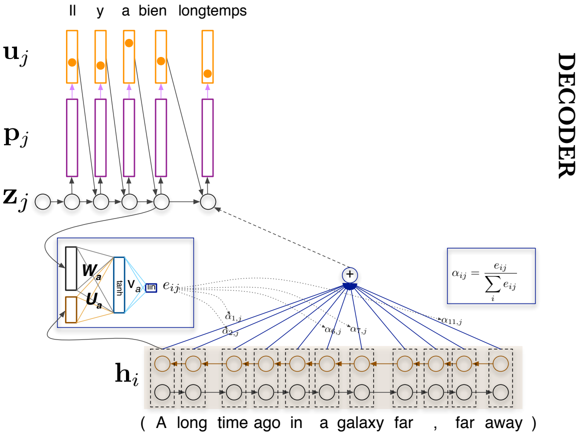
# Decoder with attention



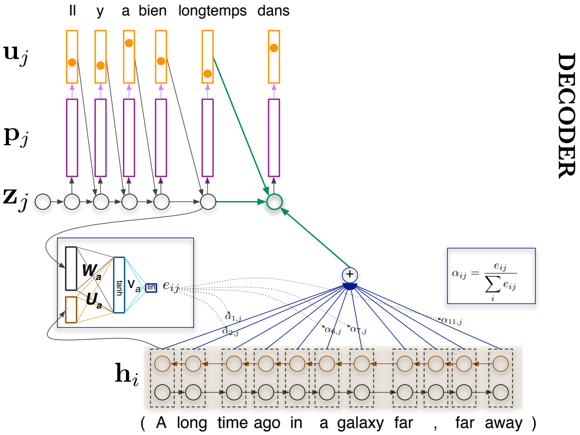
# Decoder with attention



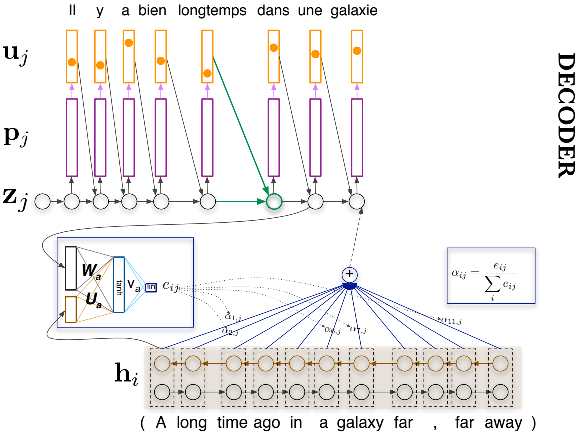
# Decoder with attention



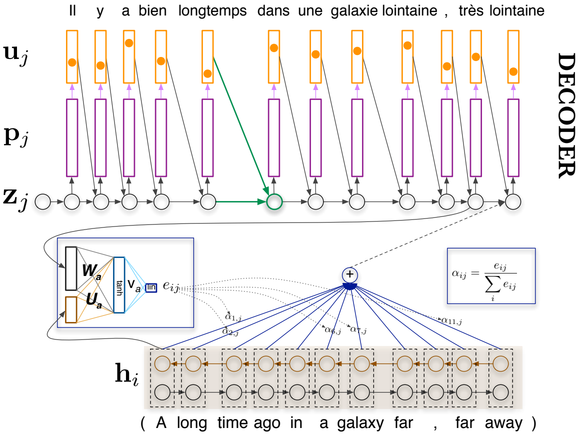
# Decoder with attention



# Decoder with attention



# Decoder with attention



# Multimodal Neural Machine Translation

# Related work

- Re-ranking of MT hypotheses with fixed-size visual vectors [Caglayan et al., 2016a, Shah et al., 2016]
- Fixed-size vector integration into source and/or target
  - → Prepending and/or appending visual vectors to source sequence [Huang et al., 2016]
  - → Decoder initialization [Calixto et al., 2016]
  - → Encoder/decoder initialization, multiplicative interaction schemes [Caglayan et al., 2017, Delbrouck and Dupont, 2017]
  - → ImageNet class probability vector as a feature [Madhyastha et al., 2017]
  - → Prediction of visual vectors as an auxiliary task [Elliott and Kádár, 2017]
- Multimodal Attention
  - → Shared attention [Caglayan et al., 2016a]
  - → Separate attention [Calixto et al., 2016, Caglayan et al., 2016b, Libovický and Helcl, 2017]

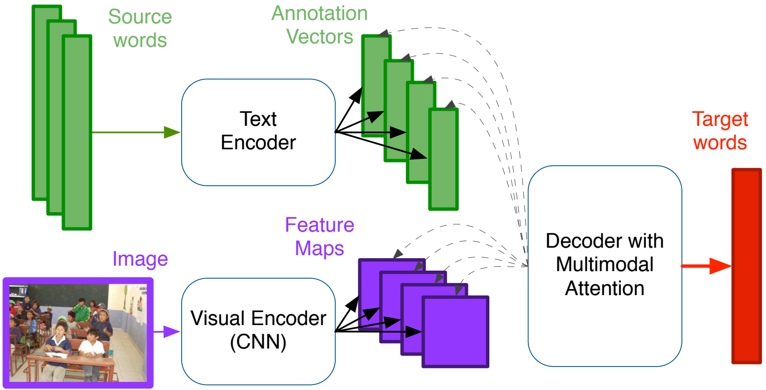
# Overview

- Two approaches will be considered today:

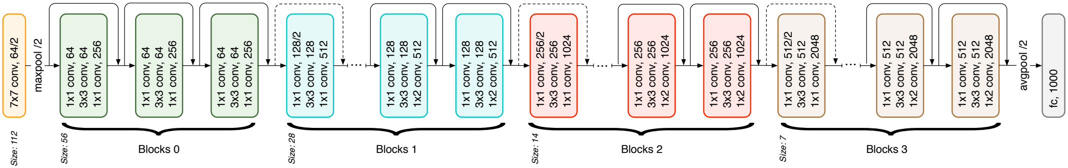
- Fusion of multiple modalities with attention
  $\rightarrow$ Combine image captioning and NMT

- Conditioning over a fixed size image vector
  $\rightarrow$ Integrate visual information at different places in the network

# Merging textual and visual information with attention



# Very deep CNN: Residual Networks

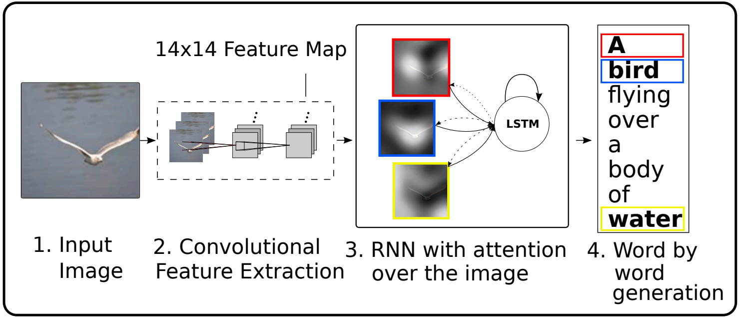



- Different configurations:
  - 50 layers [3,4,6,3]
  - 101 layers [3,4,23,8]
  - 152 layers [3,8,36,3]
- For Multimodal Machine Translation
  - Use convolutional feature maps
  - Use a fixed size representation (final average pooled activations)

# Image captioning

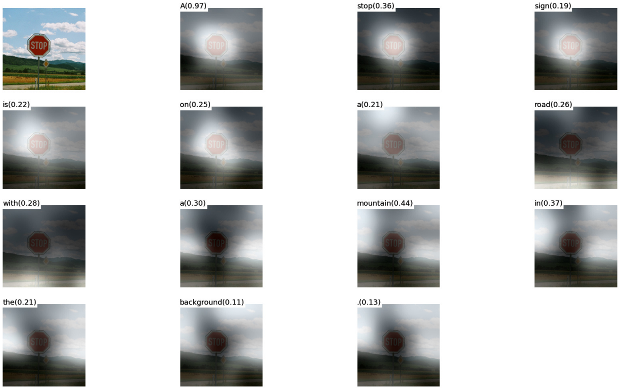
- Show, Attend and Tell, [Xu et al., 2015]

→ Image encoded with a CNN, LSTM decoder with attention



# Image captioning: example

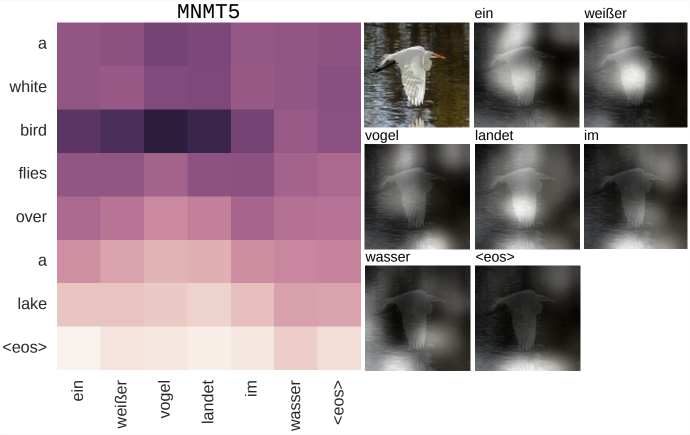
## Image captioning [Xu et al., 2015]




(b) A stop sign is on a road with a mountain in the background.

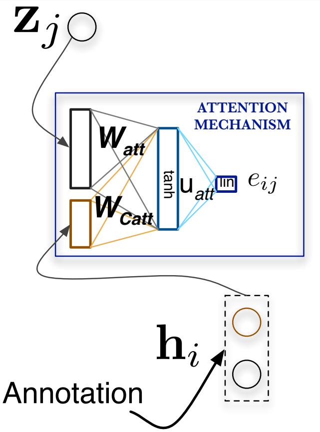
# Multimodal NMT

## MNMT: attention over text and image

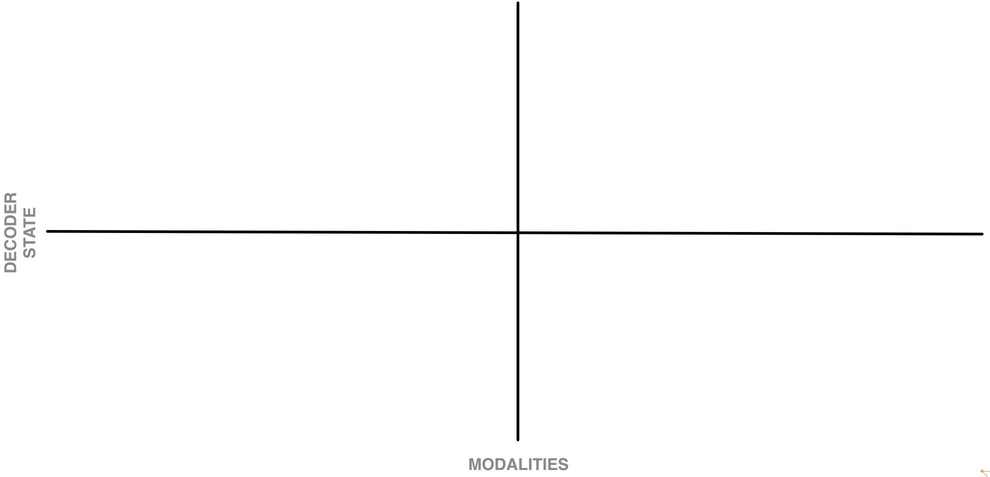


# Multimodal NMT

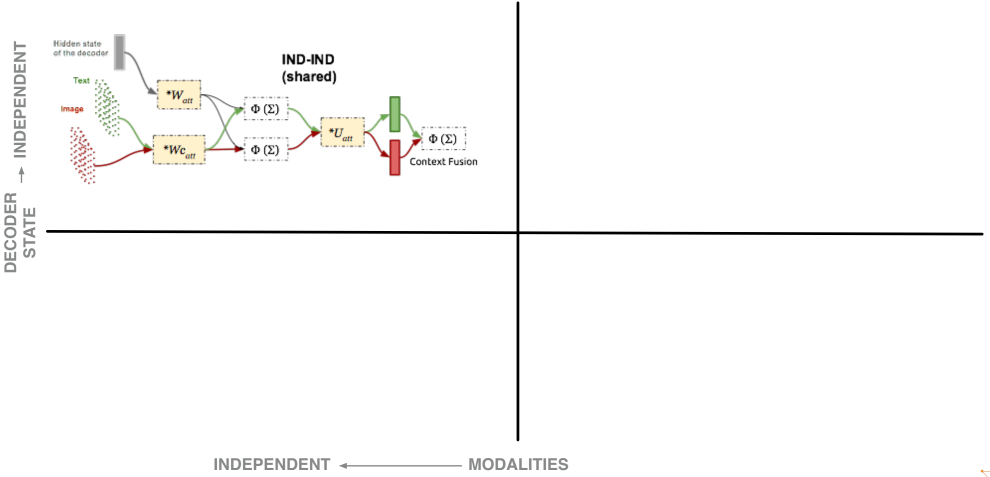
## MNMT: attention mechanism



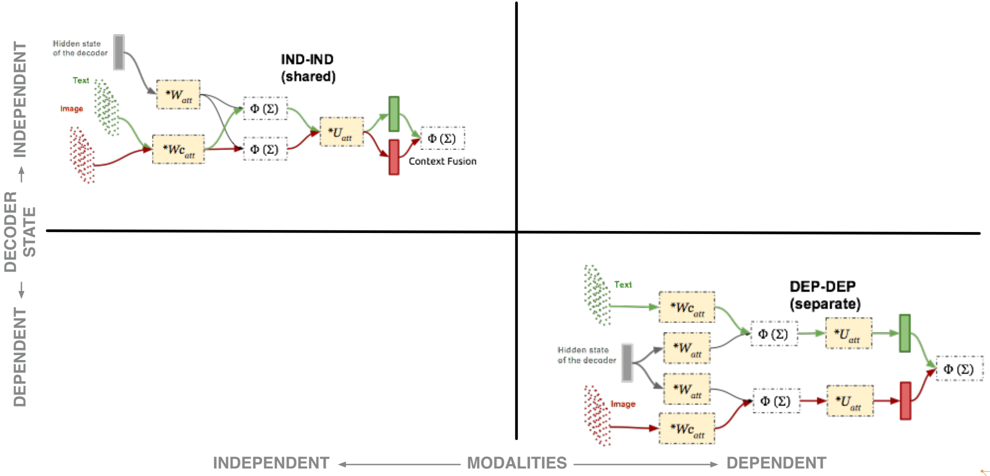
# Multimodal Attention



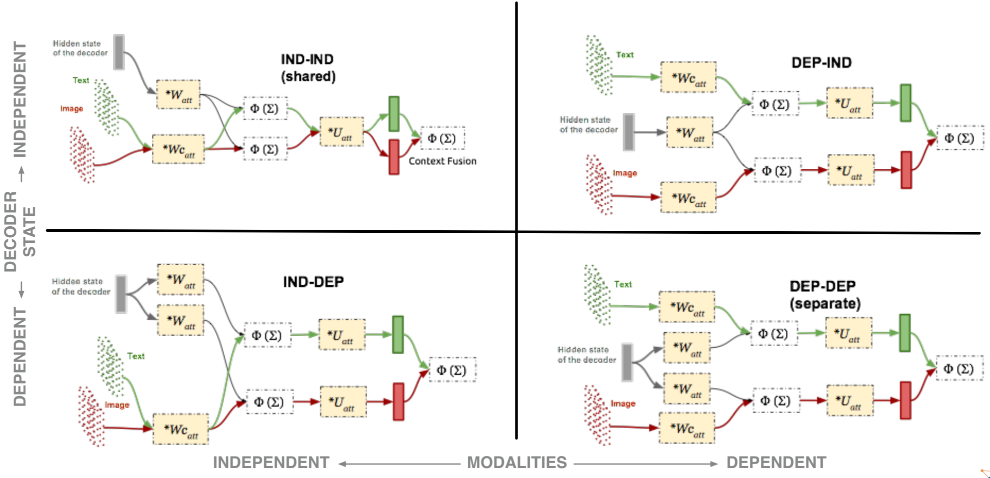
# Multimodal Attention



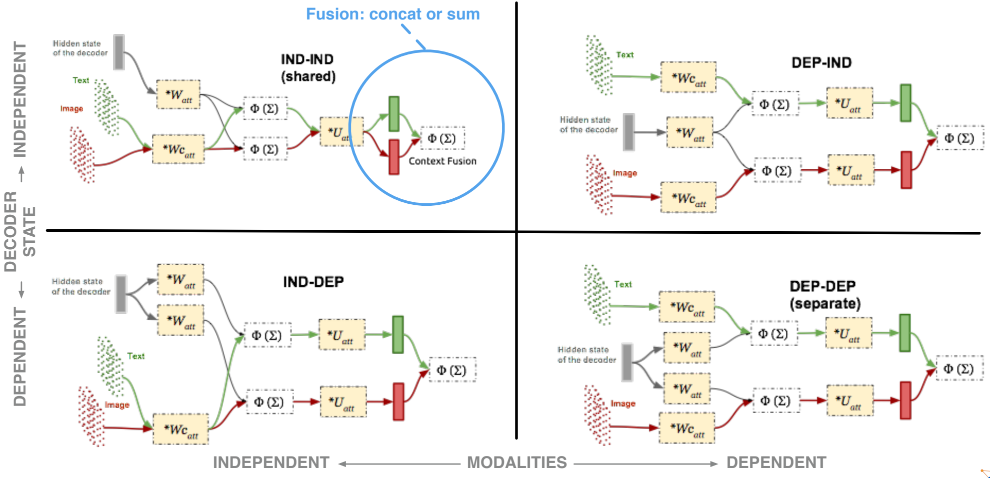
# Multimodal Attention



# Multimodal Attention



# Multimodal Attention



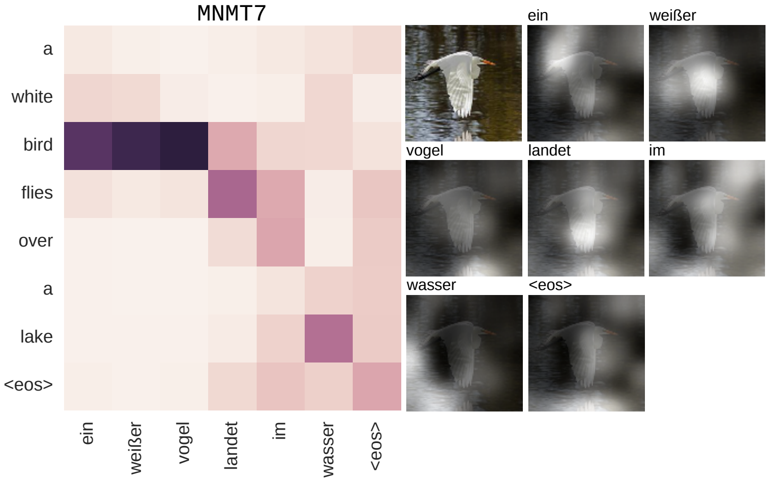
# Multimodal attention

| | Attention Type | | | Validation Scores | | |
|---|---|---|---|---|---|---|
| Model | Fusion | Modality | Decoder | METEOR | BLEU | CIDEr-D |
| NMT | - | - | - | 34.24 (35.59) | 18.64 (21.62) | 58.57 (67.93) |
| IMGTXT | - | - | - | 26.80 | 11.16 | 31.28 |
| MNMT1 | SUM | IND | IND | 33.23 (35.42) | 18.30 (21.24) | 55.45 (65.03) |
| MNMT2 | SUM | IND | DEP | 34.17 (35.48) | 17.70 (20.70) | 53.78 (61.76) |
| MNMT3 | SUM | DEP | IND | 34.38 (35.55) | 18.42 (20.94) | 55.81 (63.37) |
| MNMT4 | SUM | DEP | DEP | 33.67 (34.57) | 17.83 (20.30) | 52.68 (59.63) |
| MNMT5 | CONCAT | IND | IND | 33.31 (34.98) | 17.50 (20.60) | 53.57 (61.46) |
| MNMT6 | CONCAT | IND | DEP | **35.23** (36.79) | 19.30 (22.45) | 60.62 (69.96) |
| MNMT7 | CONCAT | DEP | IND | **35.11 (37.13)** | **19.72*** **(23.24)** | **61.04 (72.16)** |
| MNMT8 | CONCAT | DEP | DEP | 34.80 (**36.98**) | 19.55 (22.78) | 60.20 (70.20) |

[Caglayan et al., 2016b]

- CONCAT is better
- Separate attention is better ( $\neq$ multilingual, [Firat et al., 2017])
- Better results than standard NMT, but ...

# Multimodal attention



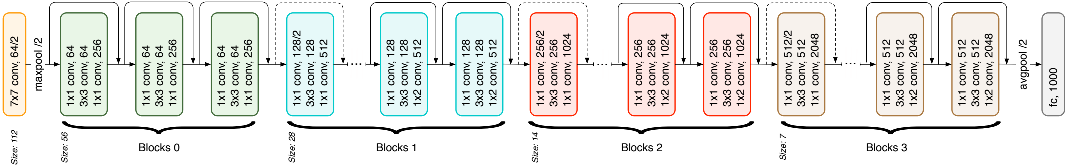
# Multimodal attention


# Multimodal attention

- Image attention not satisfying, possible causes:
  - sequence length mismatch
  - encoder is pre-trained on ImageNet and never updated
- Text attention is good
  - encoder is jointly trained with the decoder
- MMT task:
  - words contain more specific information than image
  - image is far more ambiguous
- Attention mechanism is not powerful enough to attend both text and image (?).
  - remove attention over the image
  - integrate fixed size vector from image and condition NMT with it.

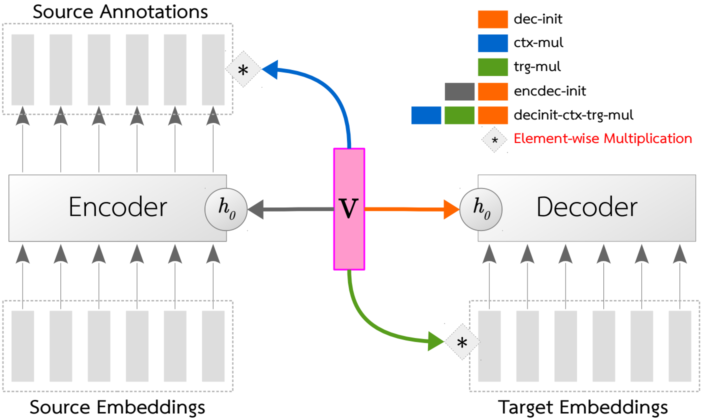
# Integrating visual information at different places in the NMT system

- Integrating fixed size vector: so-called *pool5* vector



# Integrating visual information at different places in the NMT system

- Integrating fixed size vector as visual information [Caglayan et al., 2017]



# Multimodal Machine Translation campaign (MMT'17)

- EN →DE: multiplicative interaction with target embeddings is (marginally) better

| **En→De** | # Params | Test2016 (Ensemble) BLEU | METEOR | Test2017 ($\mu \pm \sigma$/Ensemble) BLEU | METEOR |
|---|---|---|---|---|---|
| Baseline NMT | 4.6M | 40.7 | 59.2 | 30.8 $\pm$ 1.0 / 33.2 | 51.6 $\pm$ 0.5 / 53.8 |
| fusion-conv | 6.0M | 39.9 | 59.1 | 29.8 $\pm$ 0.9 / 32.7 | 51.2 $\pm$ 0.3 / 53.4 |
| dec-init-ctx-trg-mul | 6.3M | 40.2 | 59.3 | 30.9 $\pm$ 1.0 / 33.2 | 51.4 $\pm$ 0.3 / 53.7 |
| dec-init | 5.0M | 41.2 | 59.4 | 31.2 $\pm$ 0.7 / 33.4 | 51.3 $\pm$ 0.3 / 53.2 |
| encdec-init | 5.0M | 40.6 | 59.5 | 31.4 $\pm$ 0.4 / 33.5 | 51.9 $\pm$ 0.4 / 53.7 |
| ctx-mul | 4.6M | 40.4 | 59.6 | 31.1 $\pm$ 0.7 / 33.5 | 51.9 $\pm$ 0.2 / 53.8 |
| **trg-mul** | 4.7M | 41.0 | **60.4** | 30.7 $\pm$ 1.0 / 33.4 | 52.2 $\pm$ 0.4 / **54.0** |

- EN →FR: no clear difference

| **En→Fr** | Test2016 (Ensemble) BLEU | METEOR | Test2017 ($\mu \pm \sigma$/Ensemble) BLEU | METEOR |
|---|---|---|---|---|
| Baseline NMT | 54.3 | 71.3 | 50.4 $\pm$ 0.9 / 53.0 | 67.5 $\pm$ 0.7 / 69.8 |
| fusion-conv | 56.5 | 72.8 | 51.6 $\pm$ 0.9 / 55.5 | 68.6 $\pm$ 0.7 / 71.7 |
| dec-init | 56.7 | 73.0 | 52.7 $\pm$ 0.9 / 55.5 | 69.4 $\pm$ 0.7 / 71.9 |
| ctx-mul | 56.7 | 73.0 | 52.6 $\pm$ 0.9 / 55.7 | 69.5 $\pm$ 0.7 / 71.9 |
| trg-mul | 56.7 | 73.0 | 52.7 $\pm$ 0.9 / 55.5 | 69.5 $\pm$ 0.7 / 71.7 |
| ens-nmt-7 | 54.6 | 71.6 | 53.3 | 70.1 |
| **ens-mmt-6** | 57.4 | **73.6** | 55.9 | **72.2** |

## Conclusion: integrate a fixed-size image vector

- For text: Prof Ray Mooney (U. Texas):

→ "You can't cram the meaning of a whole *$#*! sentence into a single *$#*! vector!"
  - went to matrix representation + attention

- Can we summarise the whole content of an image into a single vector?
  - Probably not what we want
  - Parsimony: extract only relevant parts of the image
  - e.g. objects related to the input words
  - from coarse to fine visual information

# What's next?

- Jointly (re)-train the CNN for image encoding

$\rightarrow$ learn better features suitable for a generation task

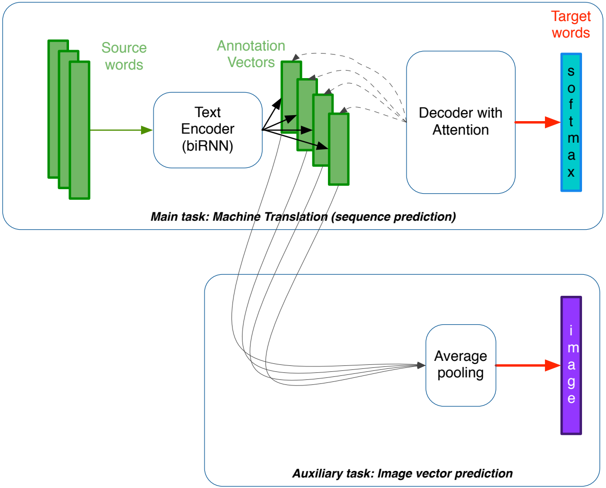
- Multi-task learning

$\rightarrow$ provide grounded word representations by introducing an auxiliary task involving image and text.

- Can be done on source or target words.

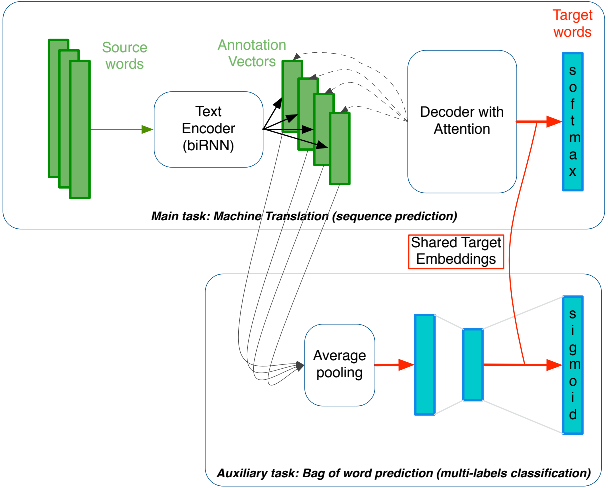
- Various auxiliary tasks can be considered
  - Predicting the image vector from source sequences
  - Predicting bag-of-words (BOW) from image ($\sim$ captioning) or from source sequence

# Multi-task learning

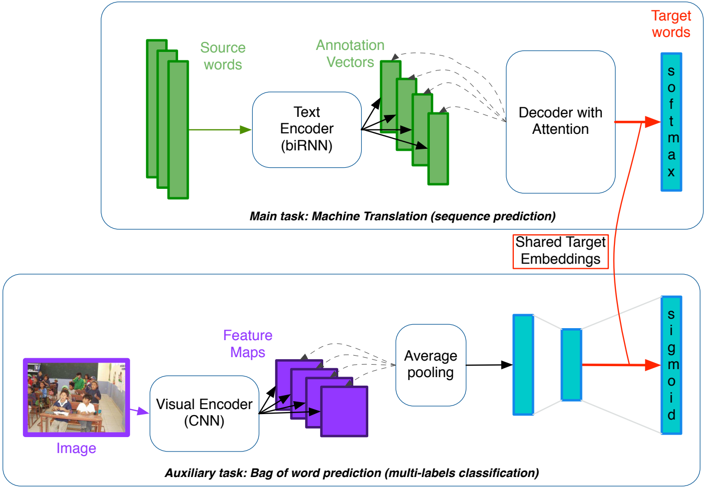



- Imagination: [Elliott and Kádár, 2017]

# Multi-task learning

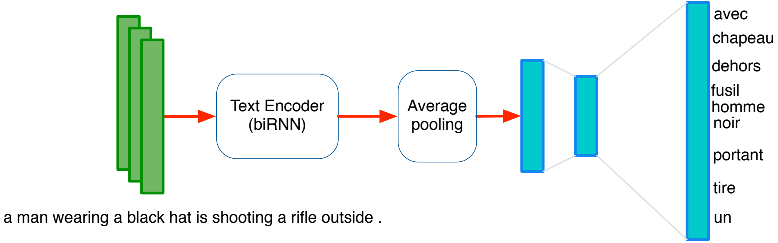


# Multi-task learning



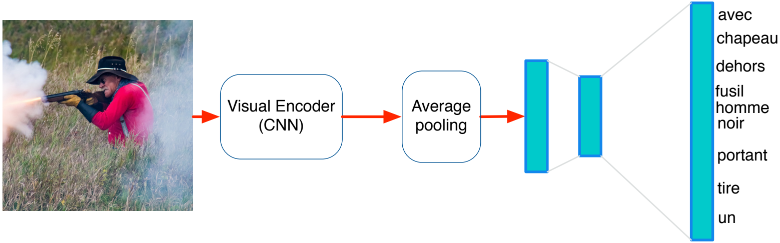
# Multi-label prediction: example

| | |
|---|---|
| SRC | a man wearing a black hat is shooting a rifle outside . |
| REF | un homme portant un chapeau noir tire avec un fusil dehors . |



# Multi-label prediction: example

| | |
|---|---|
| SRC | a man wearing a black hat is shooting a rifle outside . |
| REF | un homme portant un chapeau noir tire avec un fusil dehors . |



# Multi-task learning

- Predict words in the description (multi-label classification task)

| En→De Flickr | # Params | Test2017 ($\mu \pm \sigma$) BLEU | METEOR | TER | R@100 | LRAP |
|---|---|---|---|---|---|---|
| NMT17 | 4.6M | 30.8 $\pm$ 1.0 | 51.6 $\pm$ 0.5 | - | - | - |
| MMT17 (trgmul) | 4.7M | 30.7 $\pm$ 1.0 | 52.2 $\pm$ 0.4 | - | - | - |
| Baseline NMT | 4.6M | 31.4 $\pm$ 0.4 | 52.1 $\pm$ 0.2 | 50.4 $\pm$ 1.1 | - | - |
| NMT WP-lastctx | 5.6M | 32.2 $\pm$ 0.2 | 52.7 $\pm$ 0.5 | 49.9 $\pm$ 0.4 | 0.52 | 0.31 |
| NMT WP-lastctx-tied | 4.6M | 31.7 $\pm$ 0.8 | 52.3 $\pm$ 0.1 | 50.2 $\pm$ 0.8 | 0.51 | 0.30 |
| Visual WP-Res152 | | 31.2 $\pm$ 0.6 | 51.9 $\pm$ 0.3 | 50.7 $\pm$ 0.3 | 0.49 | 0.28 |
| Visual WP-Res50 | | 31.2 $\pm$ 0.6 | 52.6 $\pm$ 0.1 | 51.0 $\pm$ 1.6 | 0.48 | 0.28 |
| +ftune-lastblock | | 31.4 $\pm$ 0.3 | 52.3 $\pm$ 0.2 | 51.0 $\pm$ 0.5 | 0.49 | 0.27 |

- LRAP: Label Ranking Average Precision

# Multi-task learning: some conclusions / ongoing work and perspectives

- Integrating multiple tasks
  - Adding an auxiliary task seems to provide better results
  
  $\rightarrow$ try with more tasks
- Address specific "language games"
  - create a test suite dedicated to a language problem
  - e.g. gender agreement
  
  $\rightarrow$ Prof Moens relative location $\Rightarrow$ relate to textual input

# Some advertisement

nmtpytorch

## Multimodal Machine Translation framework

- framework for mono- and multi-modal NMT systems
- `https://github.com/lium-lst/nmtpytorch`

# Some advertisement

## MMT'18 @ WMT

- Organizing MMT18 evaluation campaign
- http://www.statmt.org/wmt18/multimodal-task.html
- Tasks:
  1. MMT
  2. Multi-source MMT En, De, Fr, Img $\rightarrow$ Cs
- Data:
  - Multi30k: 31k image descriptions
  - quadri-lingual (En, De, Fr, Cs) bi-modal (image, text) corpus
- Past events:
  - http://www.statmt.org/wmt17/multimodal-task.html
  - http://www.statmt.org/wmt16/multimodal-task.html

Photo Credits: Cité Plantagenêt, Le Mans Tourisme

Come visit us in Le Mans!

# Questions?

# References I

Bahdanau, D., Cho, K., and Bengio, Y. (2014).
Neural machine translation by jointly learning to align and translate.
In *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, Copenhagen, Denmark. Association for Computational Linguistics.

# References II

Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016a).
Does multimodality help human and machine for translation and image captioning?
In *Proceedings of the First Conference on Machine Translation*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Caglayan, O., Barrault, L., and Bougares, F. (2016b).
Multimodal attention for neural machine translation.
*CoRR*, abs/1609.03976.

Calixto, I., Elliott, D., and Frank, S. (2016).
Dcu-uva multimodal mt system report.
In *Proceedings of the First Conference on Machine Translation*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

# References III


Delbrouck, J. and Dupont, S. (2017).
Multimodal compact bilinear pooling for multimodal neural machine translation.
*CoRR*, abs/1703.08084.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017).
Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description.
In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Elliott, D. and Kádár, A. (2017).
Imagination improves multimodal translation.
In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

# References IV

Firat, O., Cho, K., Sankaran, B., Yarman Vural, F. T., and Bengio, Y. (2017).
Multi-way, multilingual neural machine translation.
*Computer Speech and Language.*, 45(C):236–252.

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. Association for Computational Linguistics.

Libovický, J. and Helcl, J. (2017).
Attention strategies for multi-source sequence-to-sequence learning.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202. Association for Computational Linguistics.

# References V


Madhyastha, P. S., Wang, J., and Specia, L. (2017).
Sheffield multimt: Using object posterior predictions for multimodal machine translation.
In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 470–476, Copenhagen, Denmark. Association for Computational Linguistics.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017).
Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.
*International Journal of Computer Vision*, 123(1):74–93.

# References VI

Shah, K., Wang, J., and Specia, L. (2016).
Shef-multimodal: Grounding machine translation on images.
In *Proceedings of the First Conference on Machine Translation*, pages 660–665, Berlin, Germany. Association for Computational Linguistics.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
*CoRR*, abs/1502.03044.