Multimodal Neural Machine Translation based on Image-Text Semantic Correspondence

画像とテキストの意味対応に基づくマルチモーダルニューラル機械翻訳

Yuting Zhao
東京都立大学　システムデザイン研究科 小町研究室
（現在：九州大学　システム情報科学研究院）

第10回AAMT長尾賞学生奨励賞招待講演
2023年6月21日

Outline

The content is organized as follows:

- Introduces the background and overview of this work.
- Describes existing works of the MNMT task.
- Details the proposed method of the region-attentive MNMT model.
- Details the proposed method of the word-region alignment-guided MNMT model.
- Makes a conclusion of this thesis and describes future directions.
- Introduces the social impacts of this work.

Background

Why multimodal machine translation (MMT) ?

-- Semantics still poorly used in MT.

⭐**An Example of English->German translation task.**

- 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)

MT can’t learn everything from textual context alone.

Overview of the MMT Task

Research questions:
How to integrate images into the MT model?
Can images improve the performance of MT?

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

MT Model

eine frau sitzt vor bäumen im hintergrund auf einem sehr großen ***Stein*** und lächelt in die kamera .

Existing MMT Works (1/2)

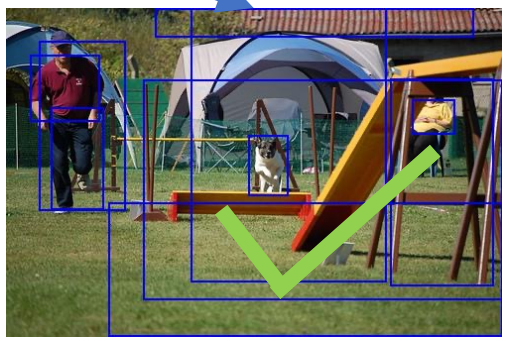
I. How to represent image feature?

Global visual feature

Local visual feature

Whole image

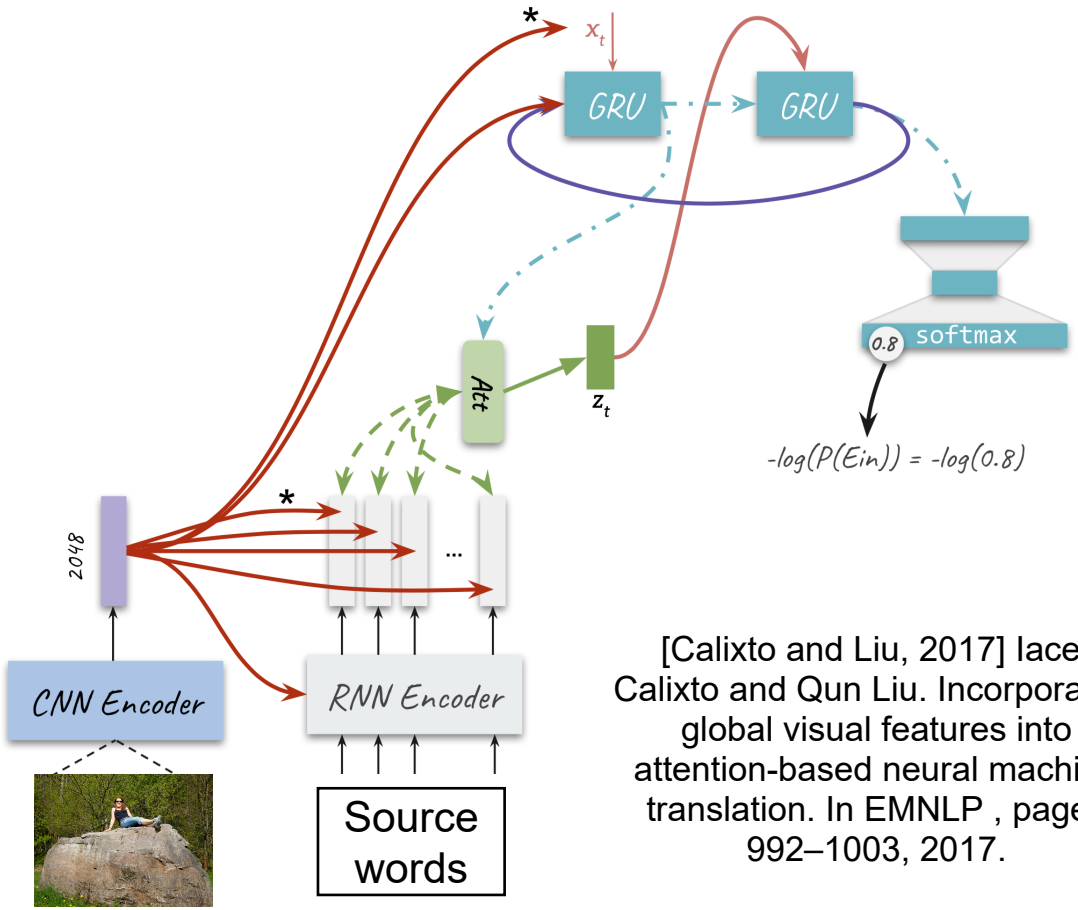
Equally-sized grid



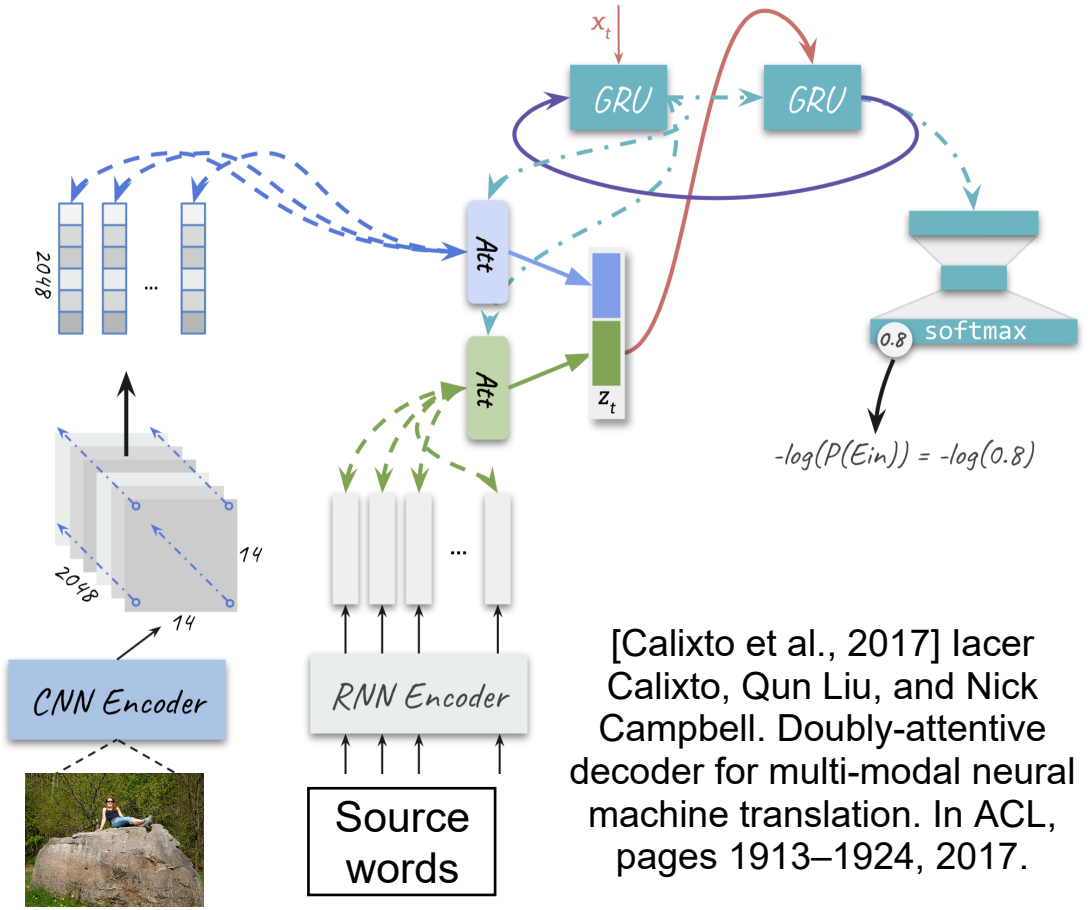
Semantic image regions

Existing MMT Works (2/2)

II. How to integrate image feature?



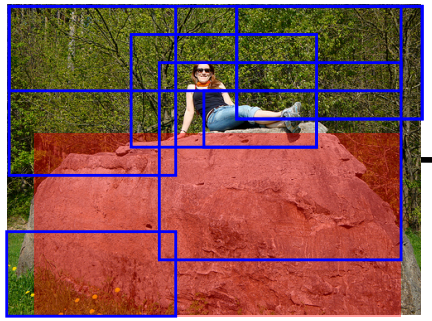
[Calixto and Liu, 2017] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In EMNLP , pages 992–1003, 2017.



[Calixto et al., 2017] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In ACL, pages 1913–1924, 2017.

Challenge in MMT

Research questions:
How to enhance the translation of the text by leveraging their semantic correspondence to the images effectively?



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

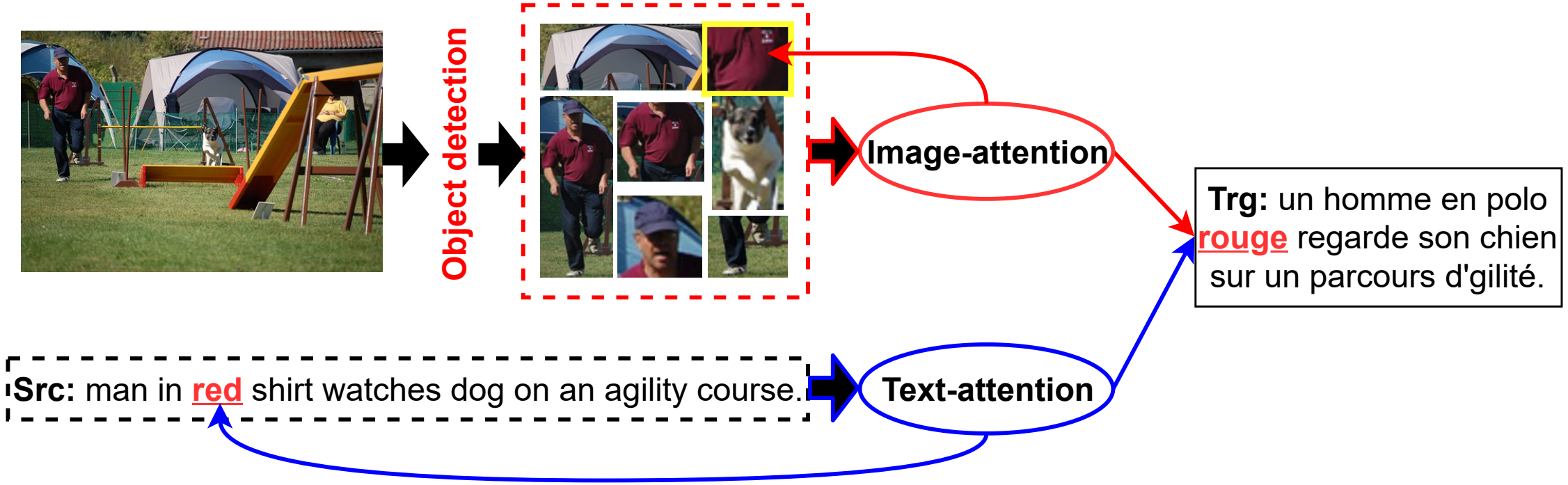
MT Model

eine frau sitzt vor bäumen im hintergrund auf einem sehr großen ***Stein*** und lächelt in die kamera .

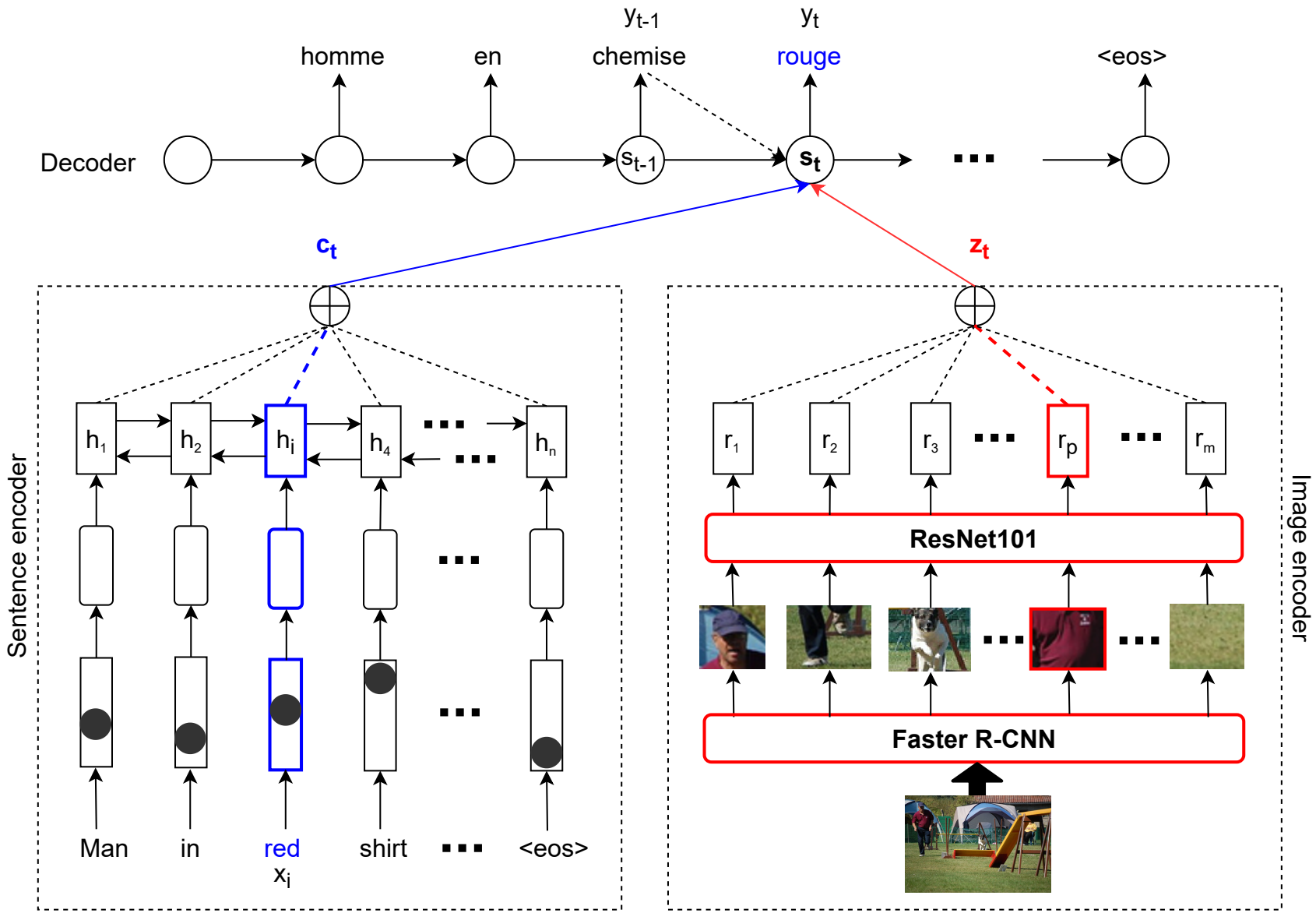
Region-Attentive Multimodal Neural Machine Translation

(Neurocomputing, Vol.476, pp.1-13, 2022.)

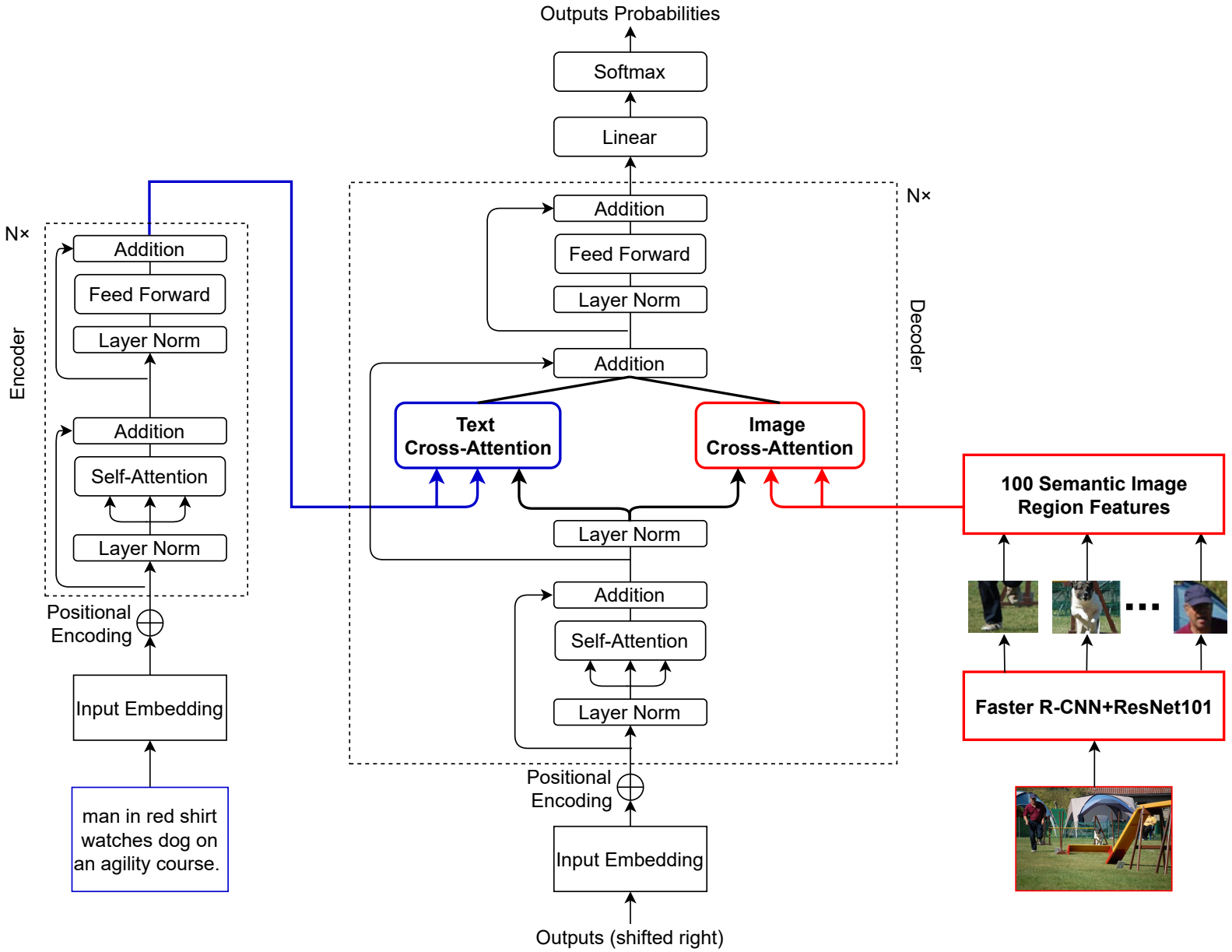
Overview: Region-Attentive MNMT



Methodology: RA-RNN



Methodology: RA-TRANS



Experiments: Datasets

Each image was paired with image descriptions expressed by both the original English sentences and the German and French translations.

- ❖ Multi30k *[1]*:
- *Train set:* 30k training images.
- *Valid set:* 1,014 validation images.
- *Test sets:* 1,000 testing images.

- ❖ Translation tasks:
- English->German (En-De)
- English->French (En-Fr)

❖ Evaluation metrics: BLEU and METEOR.

[1] Elliott, D., Frank, S., Sima'an, K., Specia, L., Multi30k: Multilingual English-German Image Descriptions, in VL@ACL 2016, pp. 70–74.

Experiments: Setup

- **Architectures: RNN-based models and Transformer-based models.**

❖ ***Baselines:***

RNN/TRANS: the text-to-text RNN/Transformer model, wherein only the textual sentences were used.

GA-RNN/GA-TRANS: the Grid-attentive multimodal RNN/Transformer model, using 7*7 equally-sized gird local visual features from each image extracted by a ResNet101 pretrained on ImageNet. Each grid-based local visual feature was represented as a 2,048-dimension vector.

❖ ***Proposed Method:***

RA-RNN/RA-TRANS: the RNN-based/Transformer-based Region-attentive MNMT model, using 100 semantic image region visual features from each image extracted by Faster R-CNN pre-trained on Visual Genome with a ResNet101 pretrained on ImageNet. Each region-based local visual feature was represented as a 2,048-dimension vector.

Results: RNN-based Models

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
RNN	34.8	53.4	56.5	71.9
GA-RNN	36.5	54.8	57.8	72.8
RA-RNN	**36.9**†	**55.5**	**58.1**†	**73.2**
v.s. RNN	(↑ 2.1)	(↑ 2.1)	(↑ 1.6)	(↑ 1.3)
v.s. GA-RNN	(↑ 0.4)	(↑ 0.7)	(↑ 0.3)	(↑ 0.4)

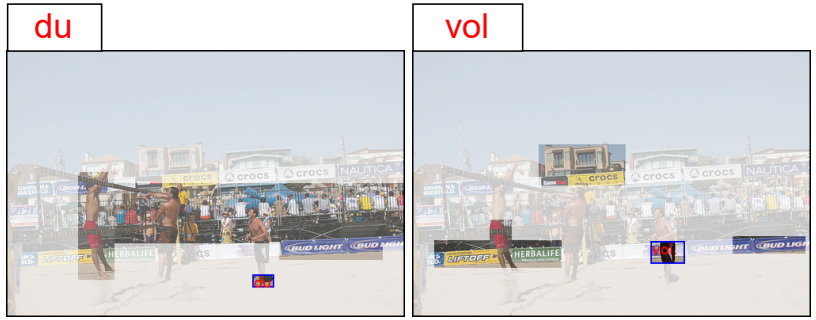
Results: TRANS-based Models

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
TRANS	35.4	52.8	57.4	72.2
GA-TRANS	37.5	55.6	59.5	74.4
RA-TRANS	**38.0**†‡	**56.0**	**60.1**†‡	**74.8**
v.s. TRANS	(↑ 2.6)	(↑ 3.2)	(↑ 2.7)	(↑ 2.6)
v.s. GA-TRANS	(↑ 0.5)	(↑ 0.4)	(↑ 0.6)	(↑ 0.4)

Analyses: Qualitative Analysis

Visualization within RNN-Based Models.

Src (En)	two people are sitting fishing on striped beach chairs in a body of water .
Ref (Fr)	deux personnes sont assises dans des fauteuils de plage rayés , pêchant dans une étendue d's eau .
RNN	deux personnes sont assises sur une structure de plage rayée (a striped beach structure) dans un plan d's eau .
GA-RNN	deux personnes sont assises , pêchent sur une plage de sable (a sandy beach) dans un plan d's eau .
RA-RNN	deux personnes sont assises à pêcher sur des chaises rayées (striped chairs) dans un plan d's eau .



Src (En)	men playing volleyball , with one player missing the ball but hands still in the air .
Ref (Fr)	des hommes jouant au volleyball , avec un joueur ratant le ballon mais avec les mains toujours en l's air .
RNN	des hommes jouant au volleyball , un joueur à l's attraper , mais les autres mains ayant toujours dans les airs (in the air) .
GA-RNN	des hommes jouant au volley-ball , avec un joueur qui le regarde dans les airs (in the air) .
RA-RNN	des hommes jouant au volleyball , avec un joueur qui passer le ballon mais les mains du vol (of the flight).

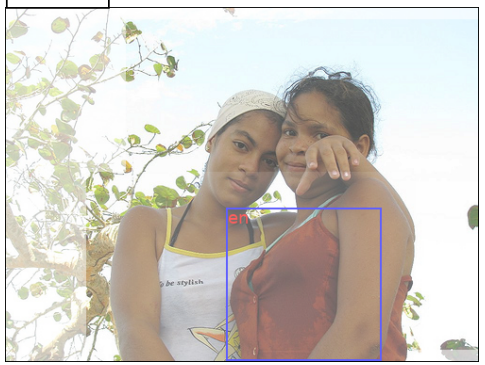
Analyses: Qualitative Analysis

Visualization within Transformer-Based Models.

manipule

Src (En)	the woman in blue is operating a camera in front of two other women .
Ref (Fr)	la femme en bleu manipule un appareil photo devant deux autres femmes .
TRANS	la femme en bleu manie (wields) une caméra en face de deux autres femmes .
GA-TRANS	la femme en bleu fait fonctionner (function) un appareil photo devant deux autres femmes .
RA-TRANS	la femme en bleu manipule (manipulate) un appareil photo devant deux autres femmes .

en



Src (En)	two women wearing tank tops are looking at the camera .
Ref (Fr)	deux femmes portant des débardeurs regardent l's objectif .
TRANS	deux femmes vêtues (wearing) de débardeurs regardent l's objectif .
GA-TRANS	deux femmes portant (wearing) des débardeurs regardent l's objectif .
RA-TRANS	deux femmes en (in) débardeurs regardent l's objectif .

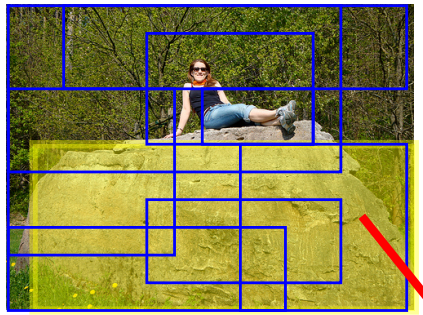
Contributions

- I have proposed region-attentive MNMT method that combines object detection with an additional region-dependent attention mechanism to fully exploit image-text semantic correspondence on NMT architectures, which are named : RA-RNN and RA-TRANS.

- Extensive experimental results show that the proposed method can significantly improve MT performance over different kinds of baselines on both RNN and Transformer architectures.

- Further analysis demonstrates that the proposed method can enhance the translation of text by attending to their semantic corresponding image regions effectively with an additional region-dependent attention mechanism.

Limitation



Limitation:
Too much object-detected image regions to be focused on is difficult.

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

MT Model

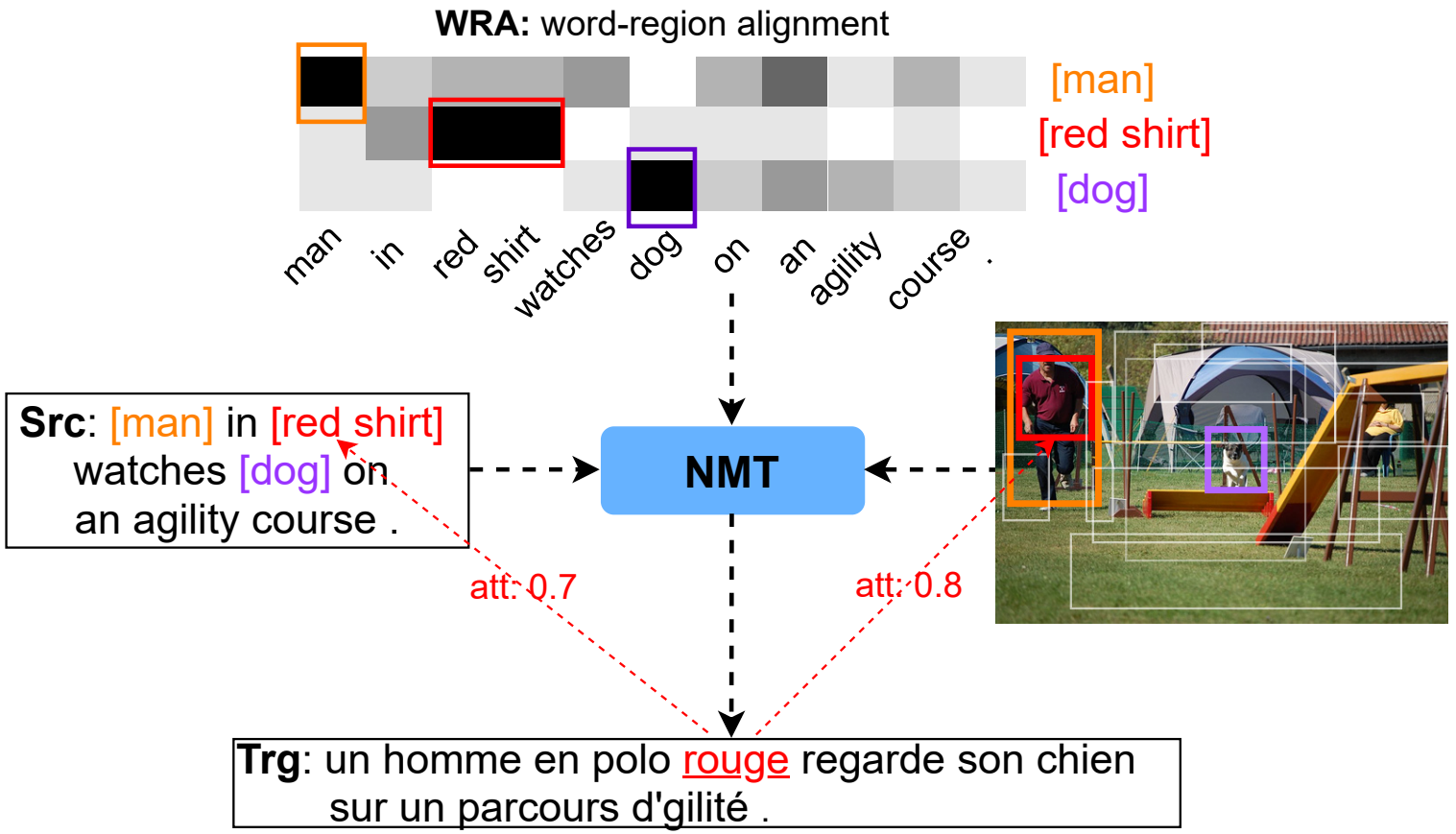
eine frau sitzt vor bäumen im hintergrund auf einem ***sehr großen Stein*** und lächelt in die kamera .

Word-Region Alignment-Guided Multimodal Neural Machine Translation

(IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.30, pp.244-259, 2022)

Overview: WRA-Guided MNMT

WRA acts as *an auxiliary input* to *guide interactions between the textual and visual information* inside the entire multimodal neural machine translation (MNMT) model.



WRA Generation

Two types of explicit WRA:

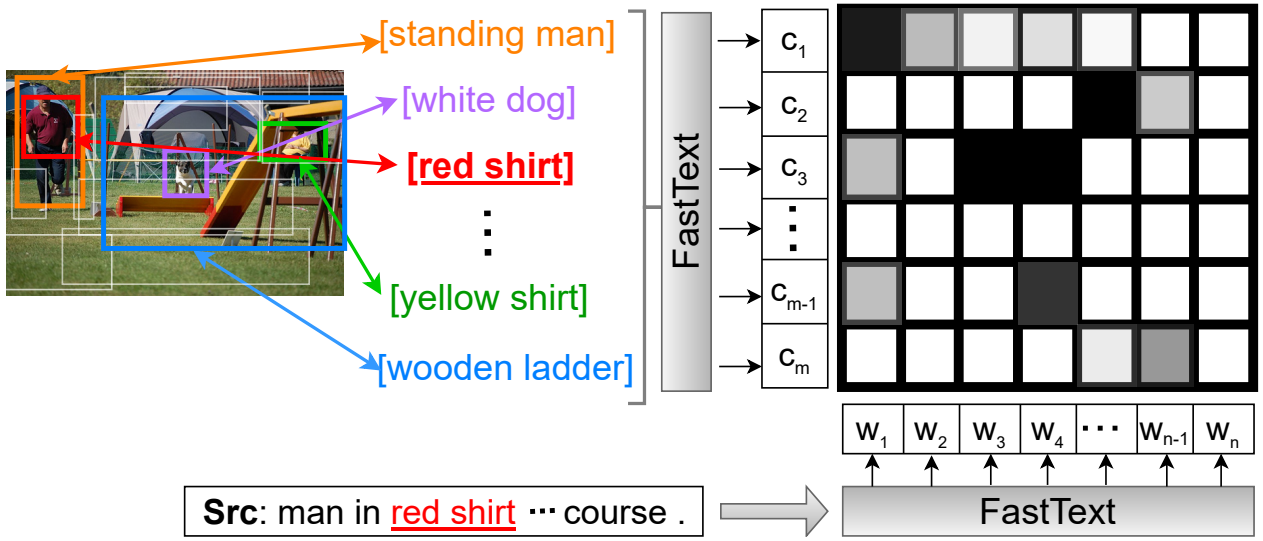
calculate the semantic relevance between the i-th region and the whole source sentence.

- **Soft WRA:** $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \cdots, \mathbf{g}_m\} \in \mathbb{R}^{d_n}$

$$g_{i,j} = \frac{\mathbf{c}_i^T \cdot \mathbf{w}_j}{\|\mathbf{c}_i\| \cdot \|\mathbf{w}_j\|}, i \in [1, m], j \in [1, n]$$

- **Hard WRA:** $\{\mathbf{g'}_1, \mathbf{g'}_2, \mathbf{g'}_3, \cdots, \mathbf{g'}_m\} \in \mathbb{R}^{d_n}$

$$g'_{i,j} = \begin{cases} 1, & \text{if } \arg\max_{j \in [1,n]}(\mathbf{g}_i) = j, \\ 0, & \text{otherwise} \end{cases}$$



Source words: $\mathrm{W} = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \cdots, \mathbf{w}_n\} \in \mathbb{R}^{d_{300}}$

Image regions: $\mathrm{C} = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \cdots, \mathbf{c}_m\} \in \mathbb{R}^{d_{300}}$

WRA containing explicit semantic interactions between the source words and image regions.

Integration Strategy: W2R (1/2)

🌟Word-to-Region (W2R):

- **Generate WRA-guided textual representations:**

$$\mathrm{H}' = (\mathbf{h}'_1, \mathbf{h}'_2, \mathbf{h}'_3, \cdots, \mathbf{h}'_m) \in \mathbb{R}^{d_\mathrm{r}}$$

  - Under the guidance of the soft WRA:

$$\mathbf{h}'_i = \mathrm{T}(\frac{1}{n} \odot (\mathbf{g}_i \cdot \mathrm{H}))$$

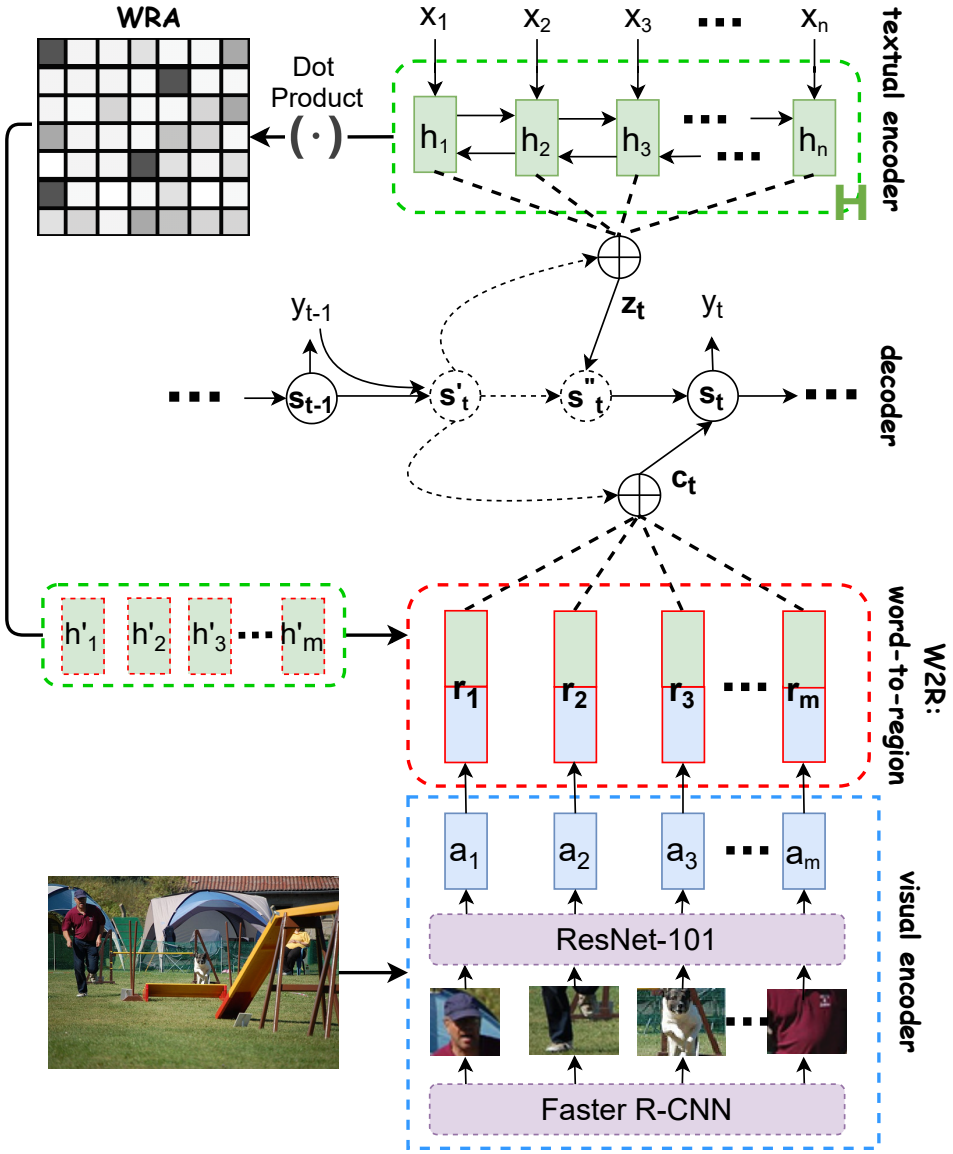
  - Under the guidance of the hard WRA:

$$\mathbf{h}'_i = \mathrm{T}(\mathbf{g}'_i \cdot \mathrm{H})$$

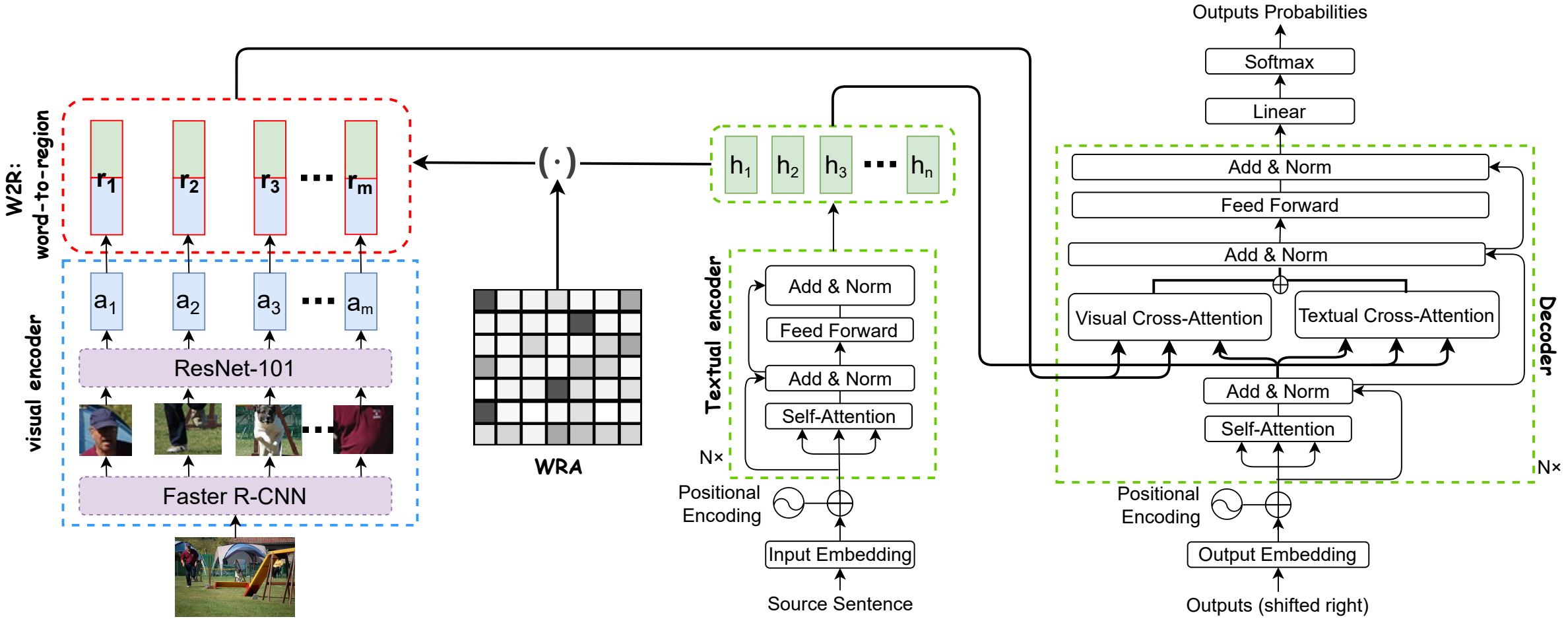
- **Generate multimodal representations:**

$$\mathrm{R} = (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \cdots, \mathbf{r}_m) \in \mathbb{R}^{d_\mathrm{r}}$$

$$\mathbf{r}_i = \mathrm{CONCAT}(\mathbf{h}'_i, \mathbf{a}_i) = \begin{bmatrix} \mathbf{h}'_i \\ \mathbf{a}_i \end{bmatrix}$$



Integration Strategy: W2R (2/2)



Experiments: Datasets

❖ Multi30k:

- *Train set:* 30k training images.
- *Valid set:* 1,014 validation images.
- *Test sets:*
  - Test2016: 1,000 testing images.
  - Test2017: 1,000 testing images.

❖ Translation tasks:

- English->German (En-De)
- English->French (En-Fr)

❖ Evaluation metrics: BLEU and METEOR.

Experiments: Setup

- **Architectures: RNN-based models and Transformer-based models.**

❖ ***Baselines:***

NMT: the text-to-text NMT model, wherein only the textual sentences were used.

$MNMT_R$: the doubly attentive MNMT model using regional visual features, without integrating WRA to process W2R strategy.

❖ ***Proposed Method:***

$MNMT_{W2R(sa)}$: the proposed MNMT model incorporating soft WRA to guide W2R stage.

$MNMT_{W2R(ha)}$: the proposed MNMT model incorporating hard WRA to guide W2R stage.

Results: En-De Task

Multi30k En→De				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing MNMT Models				
VAG-NMT [Zhou et al., 2018]	N/A	N/A	31.6	52.2
VMMT$_F$ [Calixto et al., 2019]	37.7	56.0	30.1	49.9
Del+Obj [Ive et al., 2019]	38.0	55.6	N/A	N/A
Trans+VR [Zhang et al., 2020]	36.9	N/A	28.6	N/A
VAR-S2S (hard) [Yang et al., 2020]	N/A	N/A	29.3	51.2
VAR-TF (hard) [Yang et al., 2020]	N/A	N/A	29.3	50.2
MNMT+SVA [Nishihara et al., 2020]	**39.9**	**58.1**	N/A	N/A
GMFE-NMT [Yin et al., 2020]	39.8	57.6	32.2	51.9
MTF [Yao and Wan, 2020]	38.7	55.7	N/A	N/A
OVC+L$_m$ [Wang and Xiong, 2021]	N/A	N/A	**32.3**	**53.4**
ImagiT [Long et al., 2021]	38.5	55.7	32.1	52.4

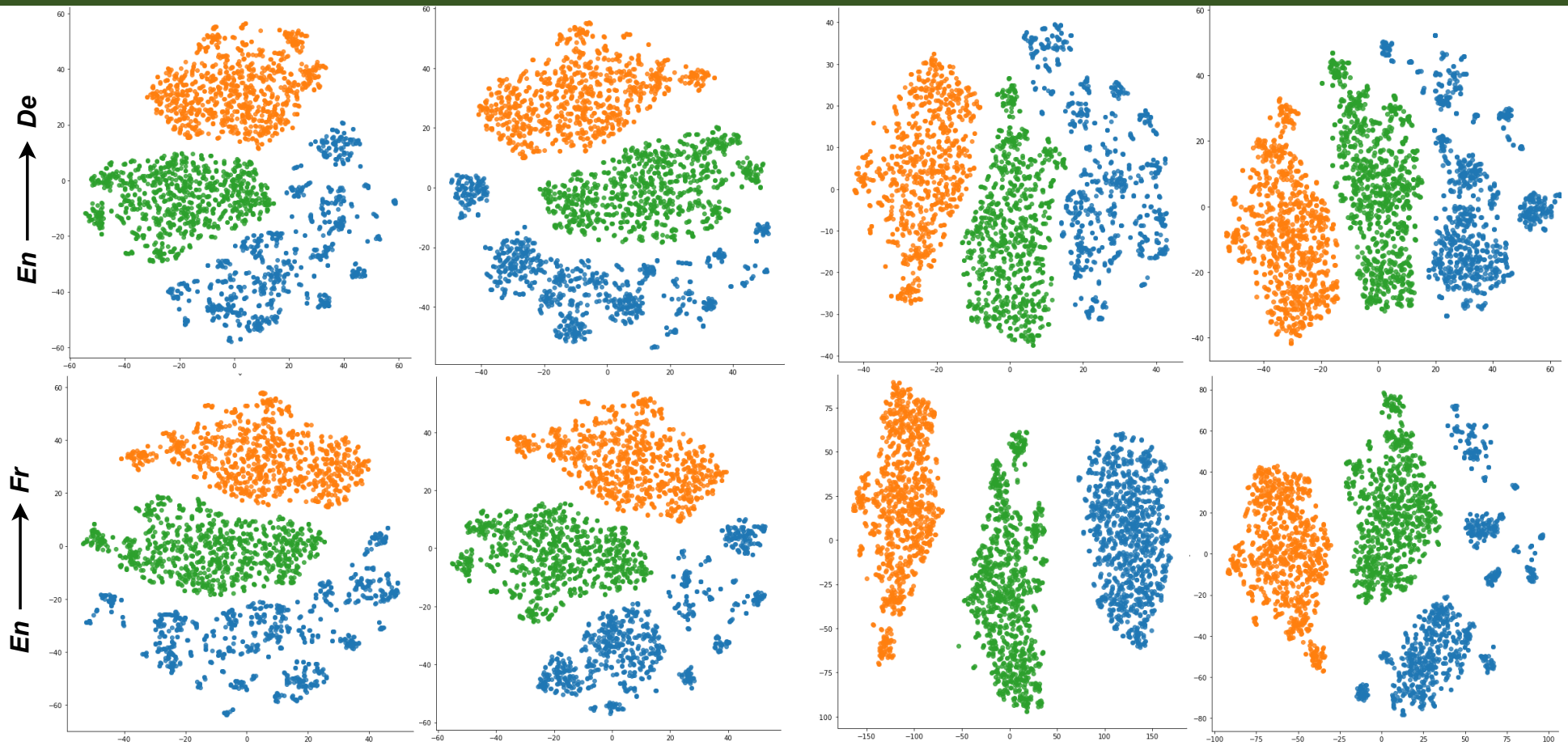
Multi30k En→De				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based Models				
NMT	37.4	57.5	29.6	51.3
MNMT$_R$	37.5	57.7	30.1	51.6
MNMT$_{W2R(sa)}$	**38.4**†‡	**58.1**	30.2†	51.9
MNMT$_{W2R(ha)}$	**38.4**†‡	58.0	**31.2**†‡	**52.2**
Transformer-Based Models				
NMT	38.4	57.5	31.5	51.9
MNMT$_R$	38.4	57.6	31.1	51.5
MNMT$_{W2R(sa)}$	**39.3**†‡	**58.3**	**32.3**†‡	**52.8**
MNMT$_{W2R(ha)}$	39.0†‡	58.2	31.8‡	52.6

Results: En-Fr Task

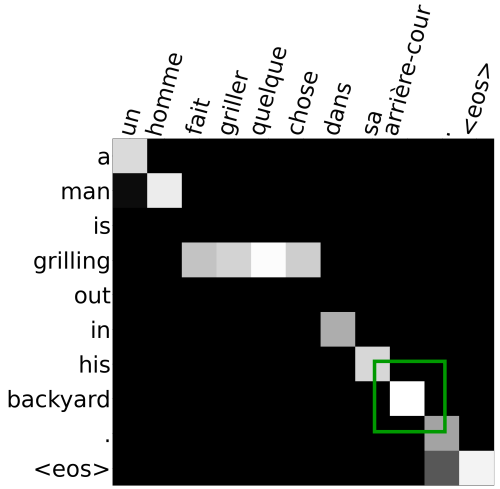
Multi30k En→Fr				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing MNMT Models				
VAG-NMT [Zhou et al., 2018]	N/A	N/A	53.8	70.3
Del+Obj [Ive et al., 2019]	59.8	74.4	N/A	N/A
VAR-S2S (hard) [Yang et al., 2020]	N/A	N/A	52.6	69.9
VAR-TF (hard) [Yang et al., 2020]	N/A	N/A	53.3	70.4
Trans+VR [Zhang et al., 2020]	57.5	N/A	48.5	N/A
GMFE-NMT [Yin et al., 2020]	**60.9**	**74.9**	53.9	69.3
OVC+L_m [Wang and Xiong, 2021]	N/A	N/A	**54.1**	**70.5**
ImagiT [Long et al., 2021]	59.7	74.0	52.4	68.3

Multi30k En→Fr				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based Models				
NMT	59.3	74.6	51.6	69.2
$MNMT_R$	59.5	74.7	51.6	69.0
$MNMT_{W2R(sa)}$	59.7	75.0	$52.2^{\dagger\ddagger}$	**69.6**
$MNMT_{W2R(ha)}$	$\mathbf{60.3}^{\dagger\ddagger}$	**75.5**	$\mathbf{52.3}^{\dagger\ddagger}$	**69.6**
Transformer-Based Models				
NMT	60.7	75.2	53.1	69.6
$MNMT_R$	60.6	75.4	52.7	69.2
$MNMT_{W2R(sa)}$	$61.7^{\dagger\ddagger}$	**76.3**	$\mathbf{54.1}^{\dagger\ddagger}$	**70.6**
$MNMT_{W2R(ha)}$	$\mathbf{61.8}^{\dagger\ddagger}$	**76.3**	$54.0^{\dagger\ddagger}$	70.4

Analyses: Visualization



Analyses: Case Study



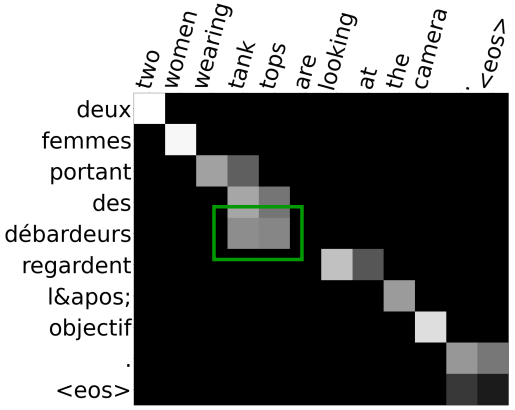
Src (En): a man is grilling out in his ***backyard*** .

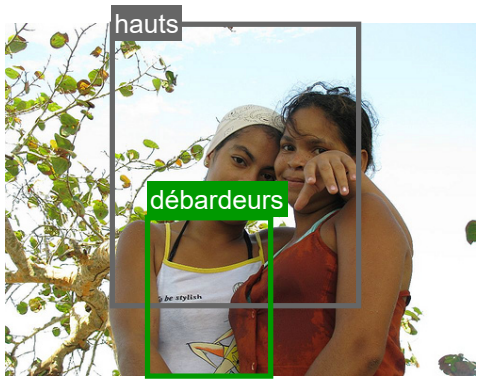
Ref (Fr): un homme fait un barbecue dans son ***arrière-cour*** .

MNMT$_{W2R(ha)}$: un homme fait griller quelque chose dans sa ***arrière-cour*** .

MNMT$_{R}$: un homme fait griller quelque chose dans sa ***cour (yard)*** .

NMT: un homme fait griller quelque chose dans sa ***cour (yard)*** .





Src (En): two women wearing ***tank tops*** are looking at the camera .

Ref (Fr): deux femmes portant des ***débardeurs*** regardent l' objectif .

MNMT$_{W2R(sa)}$: deux femmes portant des ***débardeurs*** regardent l' objectif .

MNMT$_{R}$: deux femmes vêtues de ***hauts (tops)*** regardent l' objectif .

NMT: deux femmes portant des ***hauts (tops)*** regardent l' objectif .

Contributions

- I have proposed the *WRA* to bridge multimodal inputs based on semantic correlation.
- I have proposed a *novel integration strategy W2R* to guide MNMT to translate certain source words attending to semantically corresponding image regions.
- Experiments validated the *consistent efficacy* of the proposed method and revealed that it significantly improved baselines based on different evaluation metrics.
- Further analysis demonstrates that the proposed method can achieve *better translation performance* owing to better visual information use.

Thanks for your listening!