Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions

Yuting Zhao[1], Mamoru Komachi[1], Tomoyuki Kajiwara[2], Chenhui Chu[2]
1. Tokyo Metropolitan University 2. Osaka University

Background

Why multimodal machine translation ?

-- Semantics still poorly used in MT systems.

- 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 system can't learn everything from text only.

Related Work(1/3)

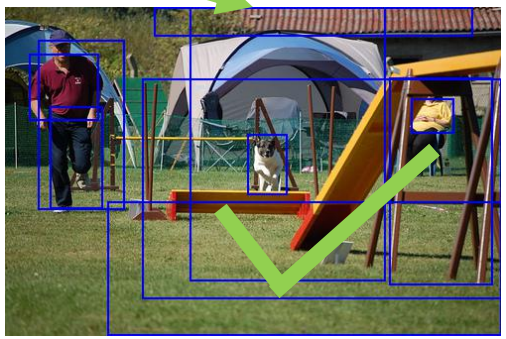
Visual modality in MNMT

Global visual feature

Local visual feature

Whole image

Grid regions

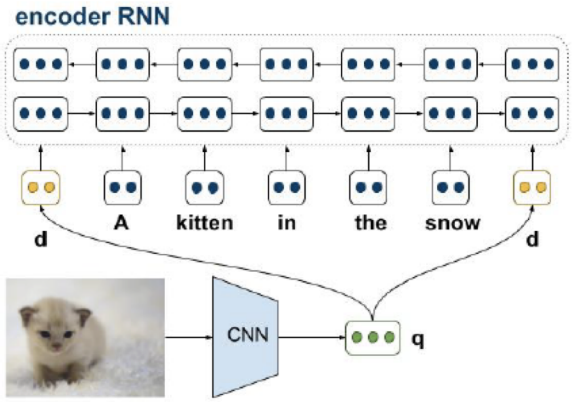


Semantic image regions

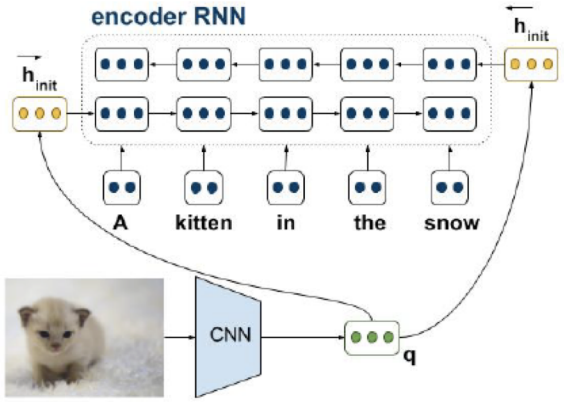
Related Work (2/3)

Global visual feature

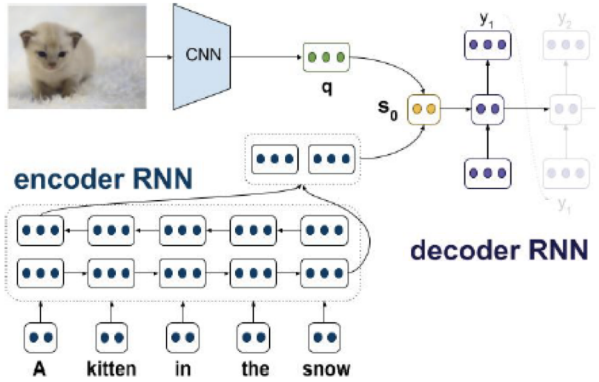
Calixto, Liu, and Campbell 2017:



(a) An encoder bidirectional RNN that uses image features as words in the source sequence.



(b) Using an image to initialise the encoder hidden states.



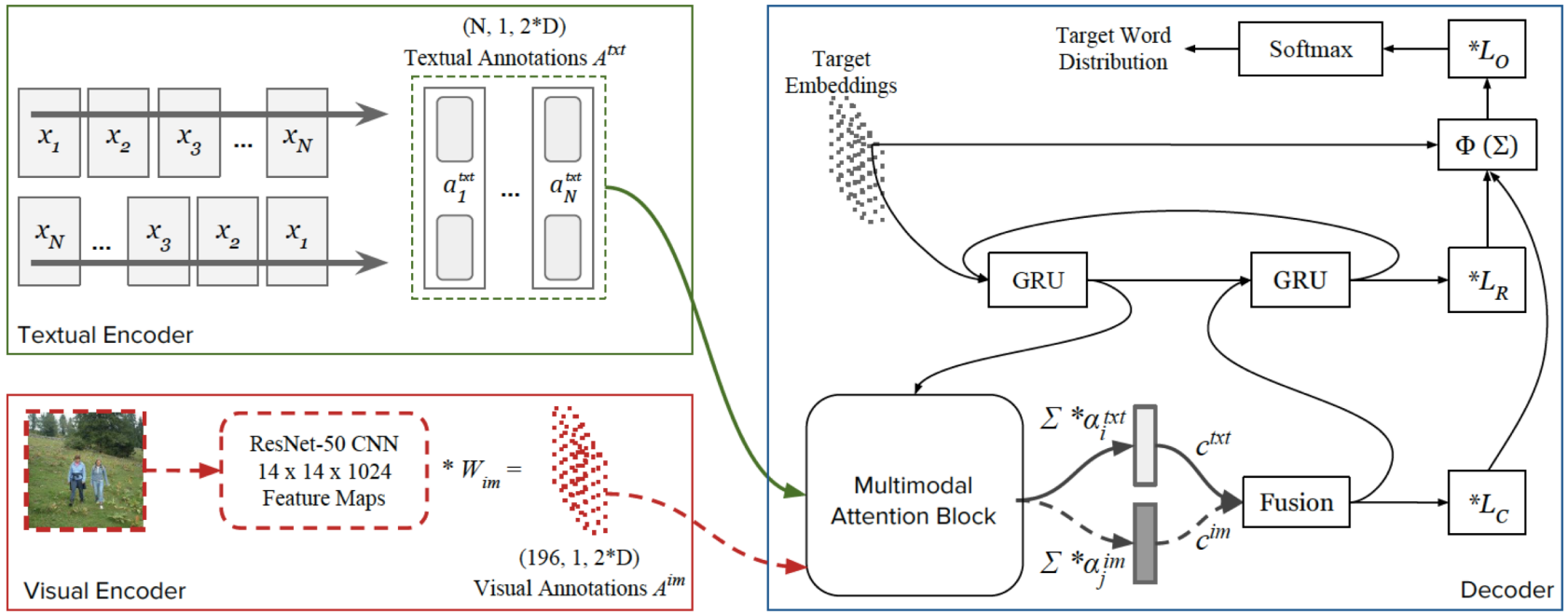
(c) Image as additional data to initialise the decoder hidden state s_0.

Whole image does not contain detailed visual features.

Related Work (3/3)

Local visual feature

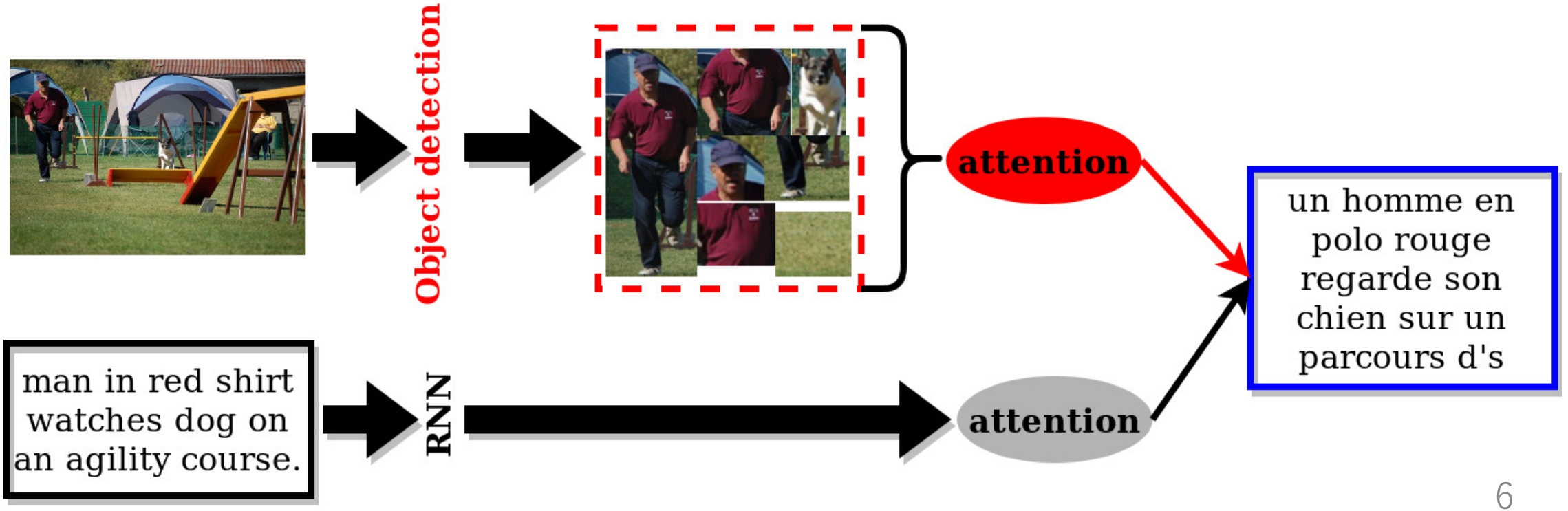
Caglayan, Barrault, and Bougares 2016:



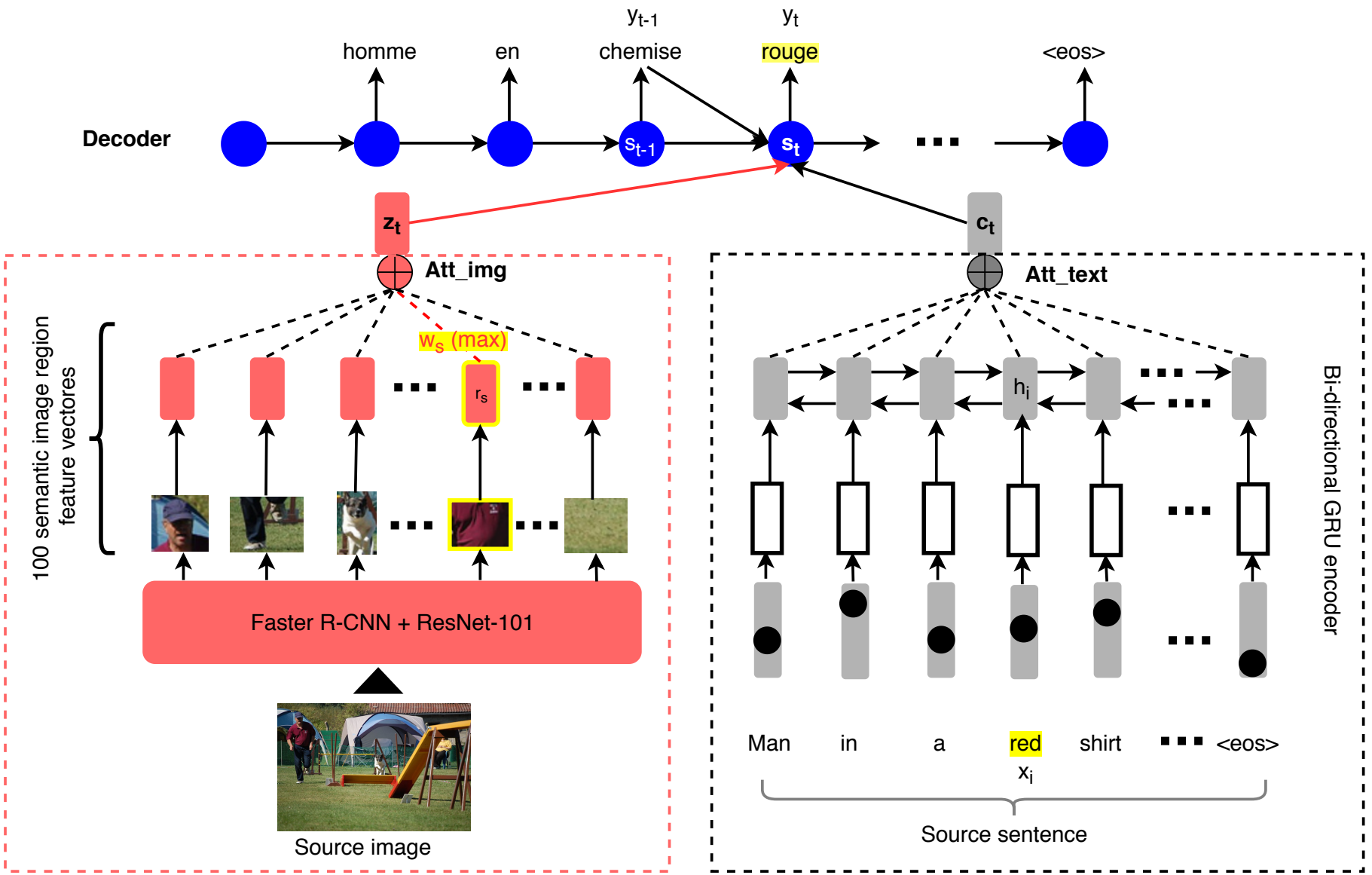
Grid regions does not contain semantic visual features.

Overview

- Semantic image regions for MNMT with double attention
- 0.5 and 0.9 BLEU point improvement on English--German and English—French.



The model(1/5)


Decoder
homme
en
y_{t-1}
chemise
y_t
rouge
<eos>
s_{t-1}
s_t
z_t
c_t
Att_img
Att_text
w_s (max)
r_s
h_i
100 semantic image region feature vectors
Faster R-CNN + ResNet-101
Source image
Bi-directional GRU encoder
Man
in
a
red
shirt
<eos>
x_i
Source sentence

The model (2/5)

Source-sentence side: $X = (x_1, x_2, x_3, \cdots, x_n)$

A bi-directional GRU generates annotation vectors:

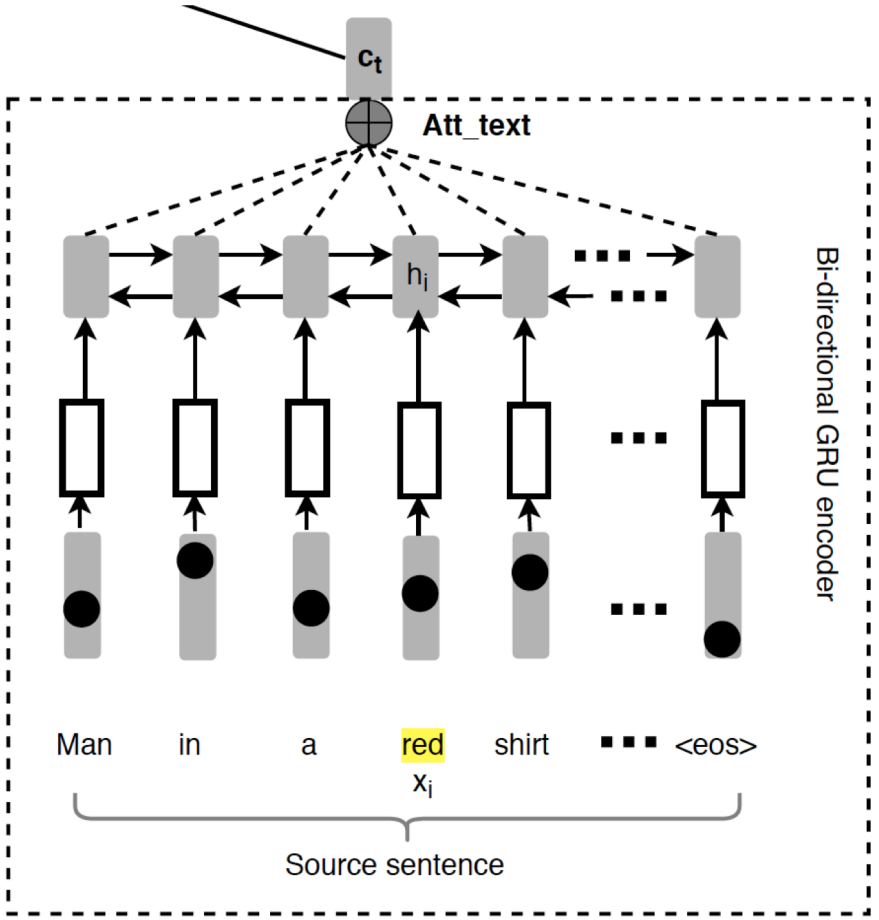
$C = (h_1, h_2, h_3, \cdots, h_n)$

Text attention generates a text context vector $\mathbf{c_t}$:

$$e_{t,i}^{\text{text}} = (V^{\text{text}})^{\mathrm{T}} \tanh(U^{\text{text}} \hat{s}_t + W^{\text{text}} h_i)$$

$$\alpha_{t,i}^{\text{text}} = \frac{\exp(e_{t,i}^{\text{text}})}{\sum_{k=1}^{n} \exp(e_{t,k}^{\text{text}})}$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i}^{\text{text}} * h_i$$



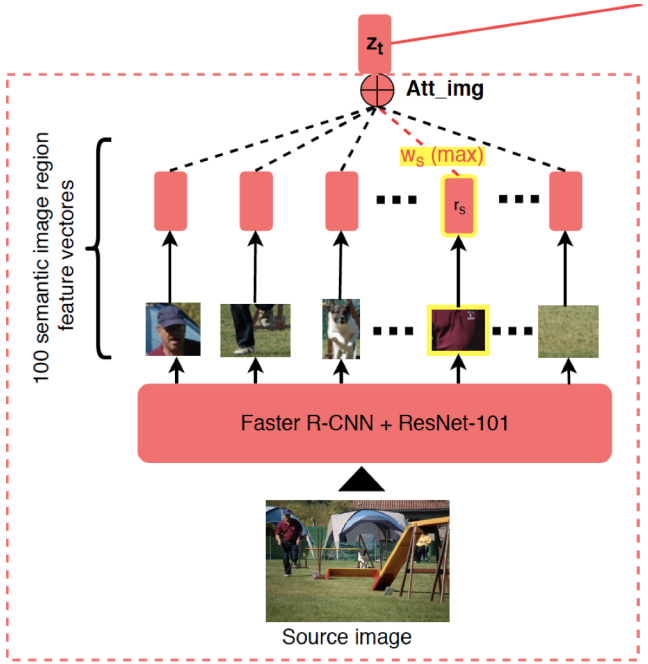
The model (3/5)

Source-image side:

- **Semantic image region extraction: Faster R-CNN + ResNet-101**

Semantic image region visual feature vectors :

$$R = (r_1, r_2, r_3, \cdots, r_{100})$$



- **Image attention: a soft attention**

Calculate an image context vector z_t

The model (4/5)

An image context vector z_t:

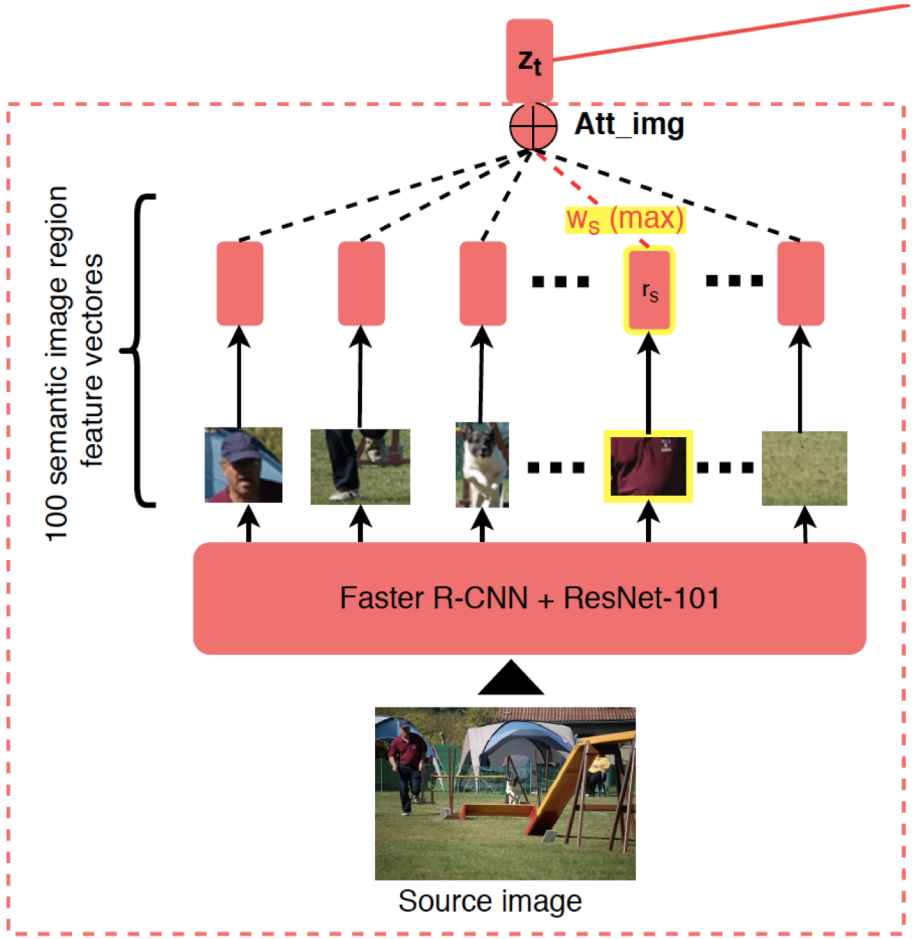
$$e_{t,s}^{\text{img}} = (V^{\text{img}})^{\text{T}} \tanh(U^{\text{img}} \hat{s}_t + W^{\text{img}} r_s)$$

$$\alpha_{t,s}^{\text{img}} = \text{softmax}(e_{t,s}^{\text{img}})$$

$$= \frac{\exp(e_{t,s}^{\text{img}})}{\sum_{k=1}^{100} \exp(e_{t,k}^{\text{img}})}$$

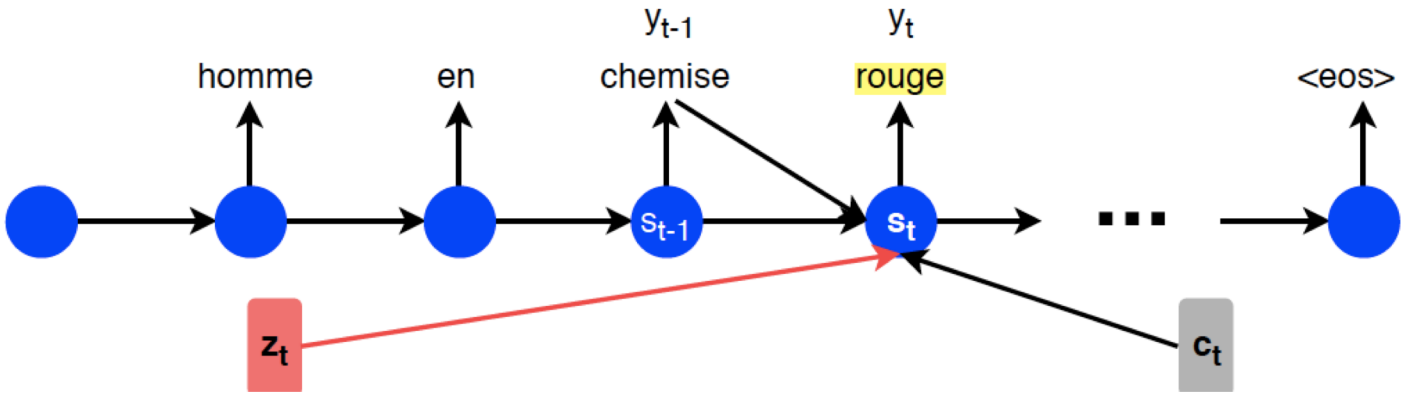
$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta)$$

$$z_t = \beta_t \sum_{s=1}^{100} \alpha_{t,s}^{\text{img}} * r_s$$



The model (5/5)

Decoder:



A 2-layer conditional GRU generates decoder hidden state $\mathbf{s_t}$:

$$\xi_t = \sigma(W_\xi^{\text{text}} c_t + W_\xi^{\text{img}} z_t + U_\xi \hat{s}_t)$$

$$\gamma_t = \sigma(W_\gamma^{\text{text}} c_t + W_\gamma^{\text{img}} z_t + U_\gamma \hat{s}_t)$$

$$\bar{s}_t = \tanh\left(W^{\text{text}} c_t + W^{\text{img}} z_t + \gamma_t \odot (U \hat{s}_t)\right)$$

$$s_t = (1 - \xi_t) \odot \bar{s}_t + \xi_t \odot \hat{s}_t$$

Calculate conditional probability of generating a target word $\mathbf{y_t}$:

$$\text{softmax}(L_o \tanh(L_s s_t + L_c c_t + L_z z_t + L_w E_Y[y_{t-1}]))$$

Experiments (1/2)

Dataset: Multi30K

29k training, 1,014 validation and 1,000 test2016 images.

Each image is paired with 5 descriptions in multiple languages.

Baselines:

Text-only OpenNMT: En-De and En-Fr textual part of Multi30k.
2-layer bidirectional GRU encoder
a 2-layer conditional GRU decoder with attention.

Doubly-attentive MNMT: Multi30k

Two attention mechanisms.
Visual features: 7 × 7 image grids by CNNs.

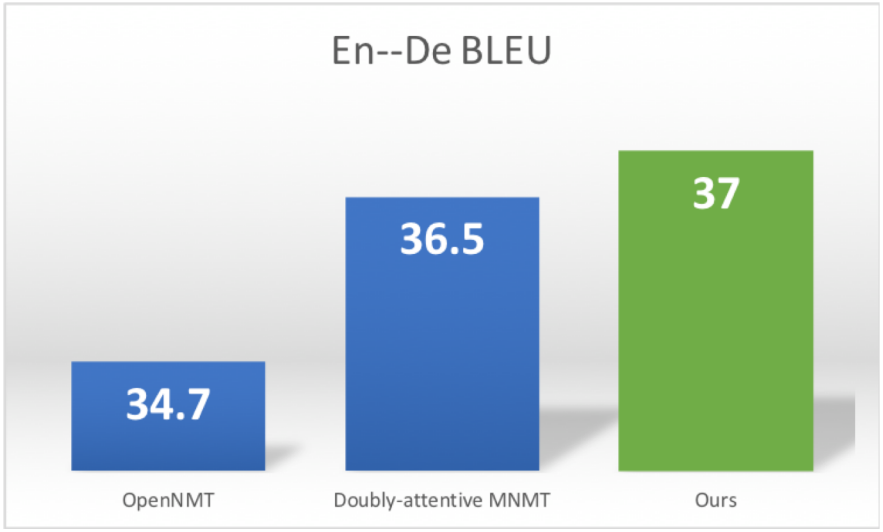
Experiments (2/2)

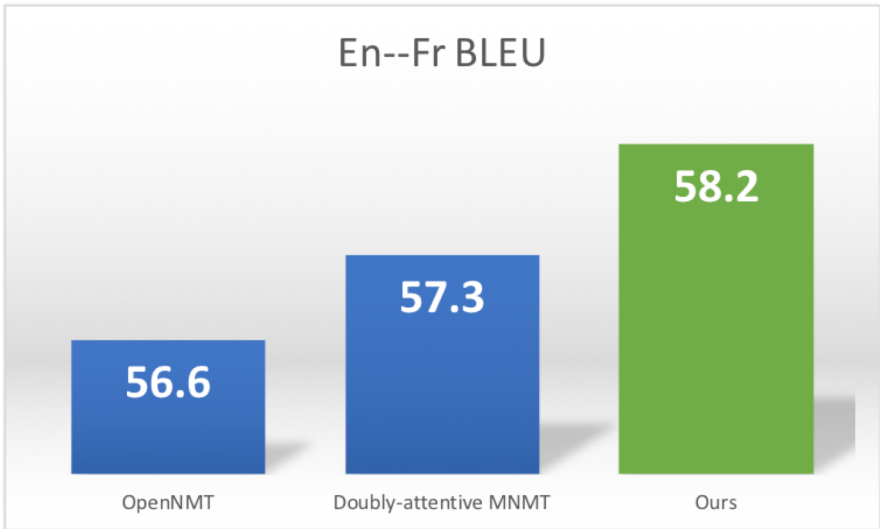
Settings:

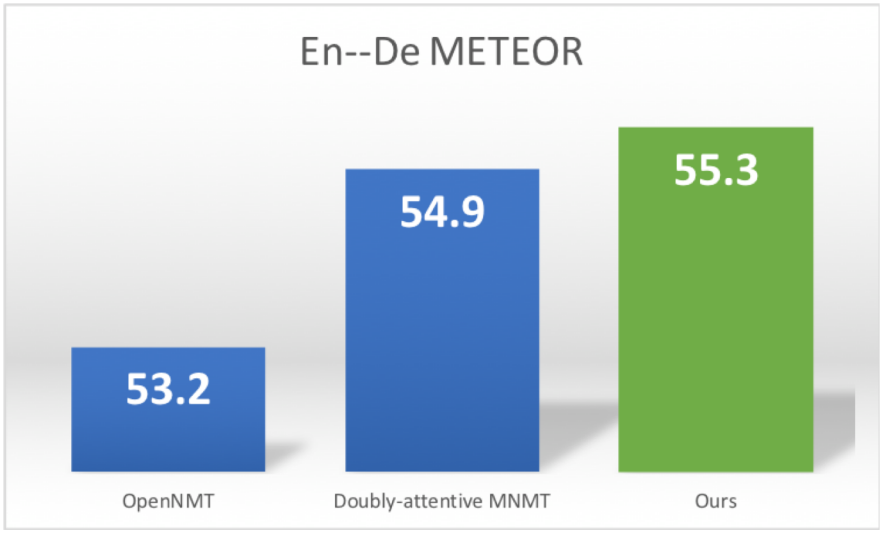
- The hidden state dimension: 500
- word embedding dimension: 500
- batch size: 40
- beam size: 5
- text dropout to 0.3
- image region dropout to 0.5.
- ADADELTA with a learning rate of 0.002
- 25 epochs

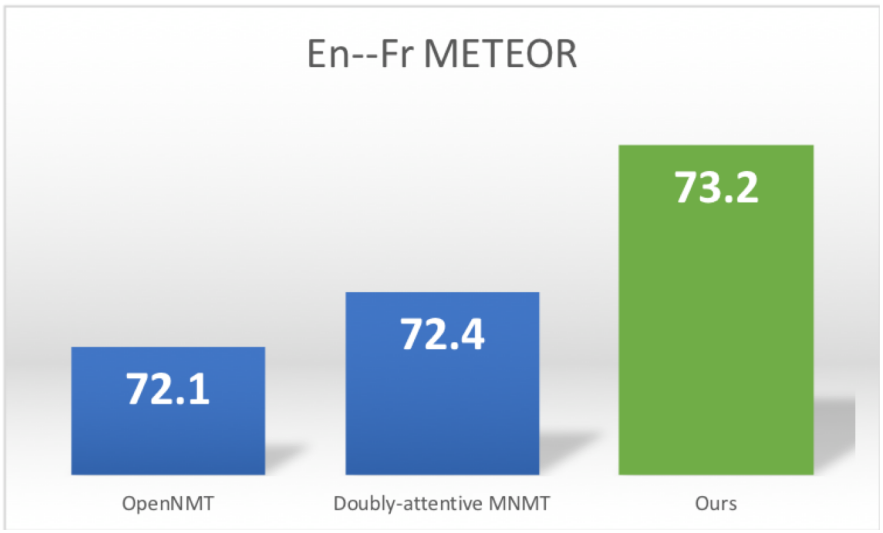
Evaluation: BLEU and METEOR metrics

Results


En--De BLEU
34.7
36.5
37
OpenNMT
Doubly-attentive MNMT
Ours


En--Fr BLEU
56.6
57.3
58.2
OpenNMT
Doubly-attentive MNMT
Ours


En--De METEOR
53.2
54.9
55.3
OpenNMT
Doubly-attentive MNMT
Ours


En--Fr METEOR
72.1
72.4
73.2
OpenNMT
Doubly-attentive MNMT
Ours

Analysis (1/2)

Comparative observation of translations from 50 examples on English-French

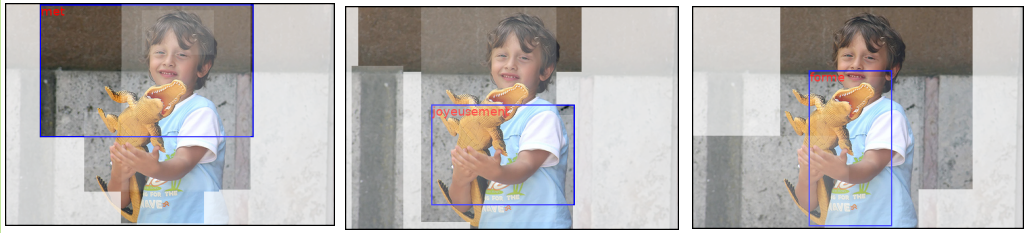
Better than baselines	Better than MNMT baseline	Better than NMT baseline	No change	Worsen
8 (16%)	6 (12%)	10 (20%)	24 (48%)	2 (4%)

Source(En)	2 blond girls are sitting on a ledge in a crowded plaza .
Reference(Fr)	deux filles blondes sont assises sur un rebord , sur une place bondée .
NMT	deux filles blondes sont assises sur une corniche dans une place très fréquentée(very busy) .
MNMT	deux filles blondes sont assises sur un rebord dans une place très fréquentée .
Ours	deux filles blondes sont assises sur un rebord dans une place bondée(crowded) .

Analysis (2/2)

Shortcomings

- The image attention mechanism does not work efficiently.
- The visual feature quality is not good enough.



Source(En)	a small child wearing a blue and white t-shirt happily holding a yellow plastic alligator .
Reference(Fr)	un petit enfant avec un t-shirt bleu et blanc tenant joyeusement un alligator en plastique jaune .
NMT	un petit enfant vêtu d's un t-shirt bleu et blanc brandissant(brandishing) une bouteille(bottle) en plastique jaune .
MNMT	un petit enfant vêtu d's un t-shirt bleu et blanc tenant(holding) un fusil(rifle) en plastique jaune .
Ours	un petit enfant vêtu d's un t-shirt bleu et blanc met(put) joyeusement(happily) une forme(shape) en plastique jaune .

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

Conclusion

Sum up:

- Integrate semantic image regions with separated attention mechanisms.
- Significantly improved translation performance.
- Verified the effect of semantic image regions.
- Analyzed the advantages and shortcomings.

Future work:

- Enhance the image region to convey a clearer semantic visual feature.
- Integrate the image attention mechanism with stronger attention function.