Doubly-Attentive Decoder for Multi-modal Neural Machine Translation

Iacer Calixto[1] and Qun Liu[1]

August 14, 2018

[1] ADAPT Centre
School of Computing
Dublin City University
{*FirstName.LastName*}*@adaptcentre.ie*

Outline

Introduction

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction

- Machine Translation (MT): the task in which we wish to learn a model to translate text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - text-only task;
  - model is trained on parallel source/target sentence pairs.

- Image description generation (IDG): the task in which we wish to learn a model to describe an image using natural language (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Multi-Modal Machine Translation

- Multi-Modal Machine Translation (MMT): learn a model to translate text and an image that illustrates this text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on source/image/target triplets;
  - can be seen as a form of augmented MT or augmented image description generation.

Multi-Modal Machine Translation

- Multi-Modal Machine Translation (MMT): learn a model to translate text and an image that illustrates this text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on source/image/target triplets;
  - can be seen as a form of augmented MT or augmented image description generation.

Multi-Modal Machine Translation

- Multi-Modal Machine Translation (MMT): learn a model to translate text and an image that illustrates this text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on source/image/target triplets;
  - can be seen as a form of augmented MT or augmented image description generation.

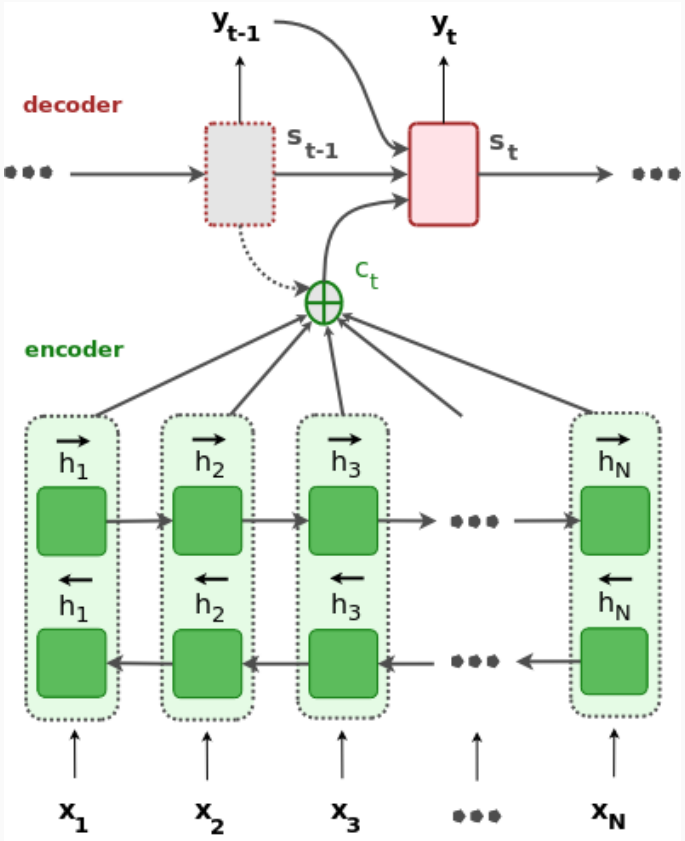
Multi-Modal Machine Translation

- Multi-Modal Machine Translation (MMT): learn a model to translate text and an image that illustrates this text from one natural language (e.g., English) into another (e.g., Brazilian Portuguese).
  - multi-modal task (text and vision);
  - model is trained on source/image/target triplets;
  - can be seen as a form of augmented MT or augmented image description generation.

Model Architecture

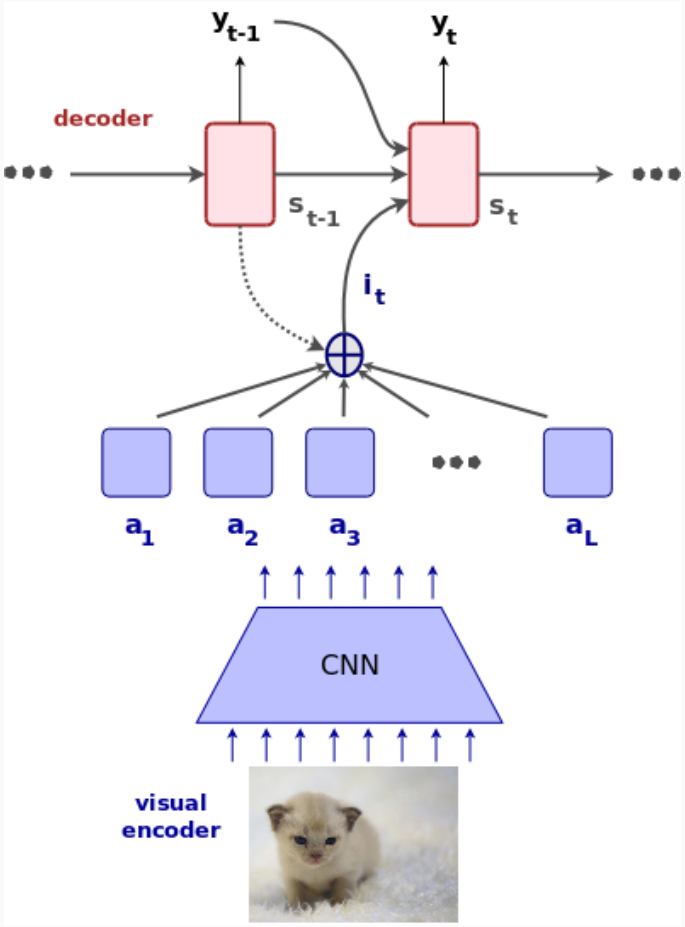
Attentional Neural Machine Translation

The attention mechanism lets the decoder search for the best source words to generate each target word, e.g. Bahdanau et al. (2015).



Attentional Neural Image Description Generation

The attention mechanism lets the decoder look at or attend to specific parts of the image when generating each target word, e.g. Xu et al. (2015).



Doubly-Attentive Multi-Modal NMT Model

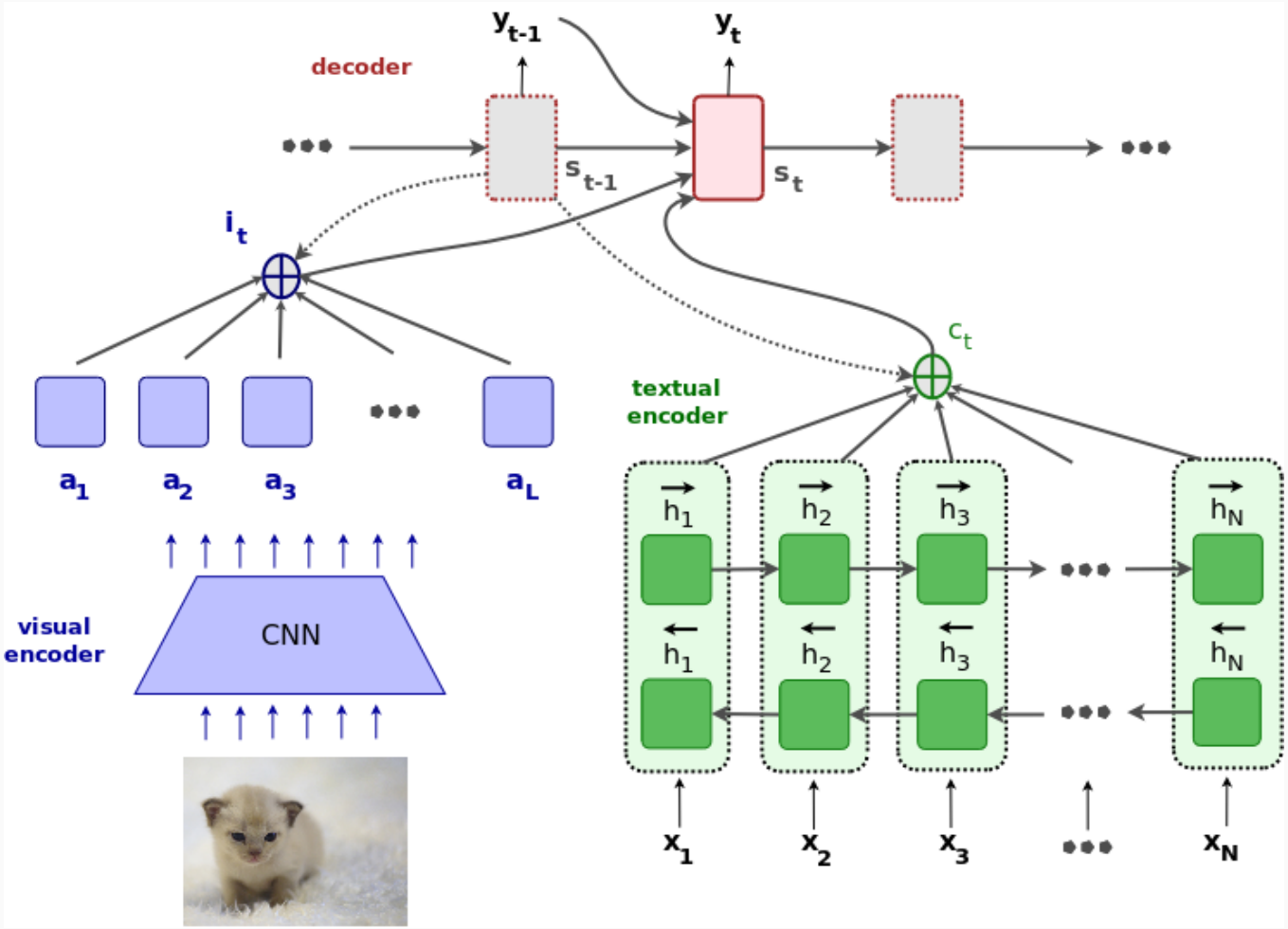


Figure 1: Doubly-Attentive Multi-modal NMT (paper accepted in ACL 2017)

image gating

Doubly-Attentive Multi-Modal NMT Model

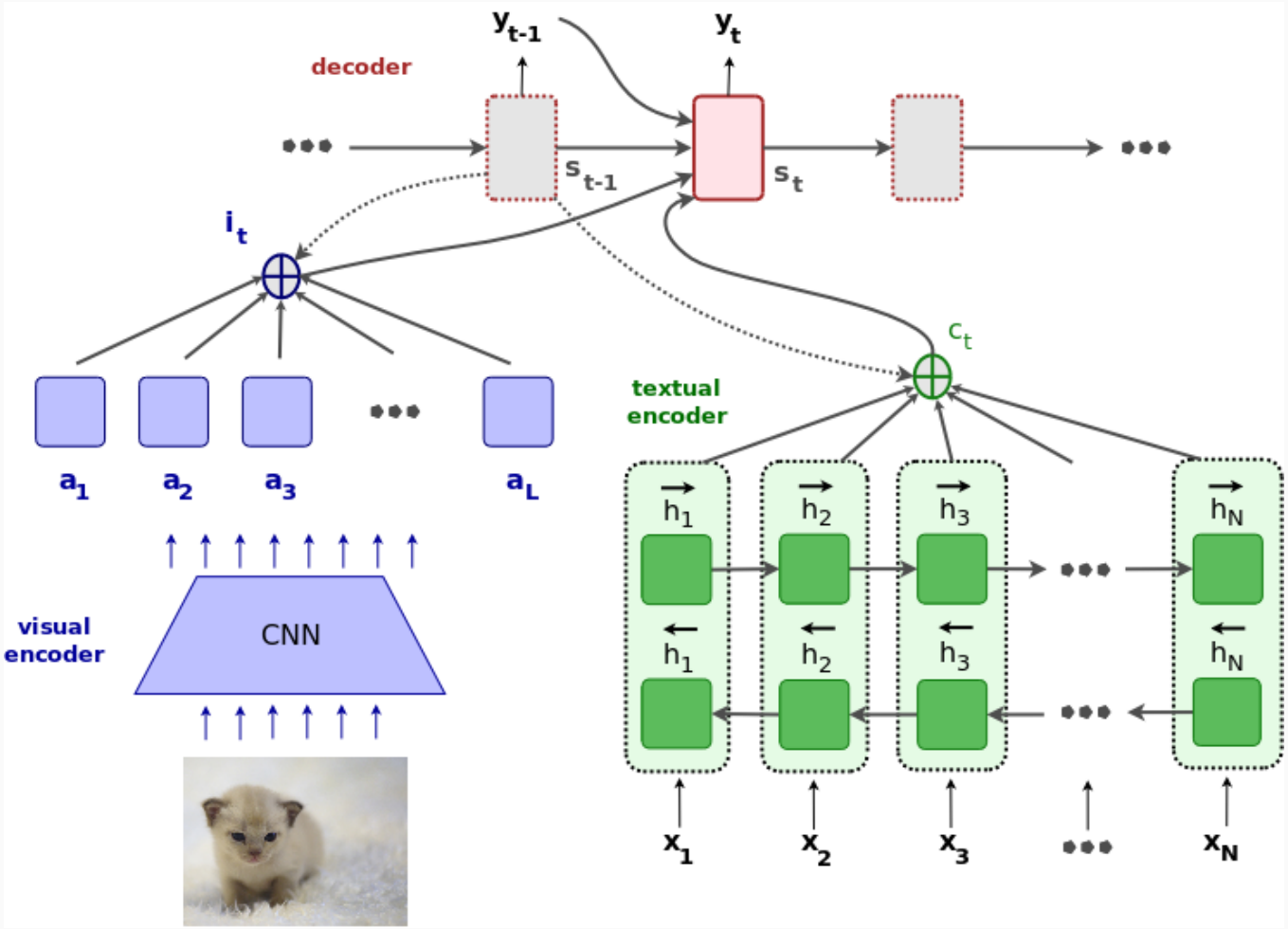


Figure 1: Doubly-Attentive Multi-modal NMT (paper accepted in ACL 2017)

image gating

Data sets

Multi30k

- **29K** images;
- **Translated Multi30k** – **29K** English–German **parallel descriptions** (1 per image);
- **Comparable Multi30k** – **145K** English and **145K** German **comparable descriptions** (5 English and 5 German per image);
- ∼**2M** English and ∼**1.6M** German words.

Multi30k

- **29K** images;
- **Translated Multi30k** – **29K** English–German **parallel descriptions** (1 per image);
- **Comparable Multi30k** – **145K** English and **145K** German **comparable descriptions** (5 English and 5 German per image);
- ∼**2M** English and ∼**1.6M** German words.

Multi30k

- **29K** images;
- **Translated Multi30k** – **29K** English–German **parallel descriptions** (1 per image);
- **Comparable Multi30k** – **145K** English and **145K** German **comparable descriptions** (5 English and 5 German per image);
- ∼**2M** English and ∼**1.6M** German words.

Multi30k

- **29K** images;
- **Translated Multi30k** – **29K** English–German **parallel descriptions** (1 per image);
- **Comparable Multi30k** – **145K** English and **145K** German **comparable descriptions** (5 English and 5 German per image);
- $\sim$**2M** English and $\sim$**1.6M** German words.

WMT 2015

- Corpora:
  - Europarl;
  - News Commentary;
  - Common Crawl;
- ∼**4.3M sentence pairs** in total;
- ∼**103M** English and ∼**103M** German words.

WMT 2015

- Corpora:
  - Europarl;
  - News Commentary;
  - Common Crawl;
- ∼**4.3M sentence pairs** in total;
- ∼**103M** English and ∼**103M** German words.

WMT 2015

- Corpora:
  - Europarl;
  - News Commentary;
  - Common Crawl;
- ∼**4.3M sentence pairs** in total;
- ∼**103M** English and ∼**103M** German words.

WMT 2015

- Corpora:
  - Europarl;
  - News Commentary;
  - Common Crawl;
- **∼4.3M sentence pairs** in total;
- **∼103M** English and **∼103M** German words.

WMT 2015

- Corpora:
  - Europarl;
  - News Commentary;
  - Common Crawl;
- **$\sim$4.3M sentence pairs** in total;
- **$\sim$103M** English and **$\sim$103M** German words.

Training sets

- **Training**:
  - **Translated Multi30k** (29k triples);
- **Pre-training**:
  - **Translated Multi30k** (29k triples) +
    back-translated Comparable Multi30k (145k triples);
  - **WMT 2015** (4.3M sentence pairs);

Experiments

English–German

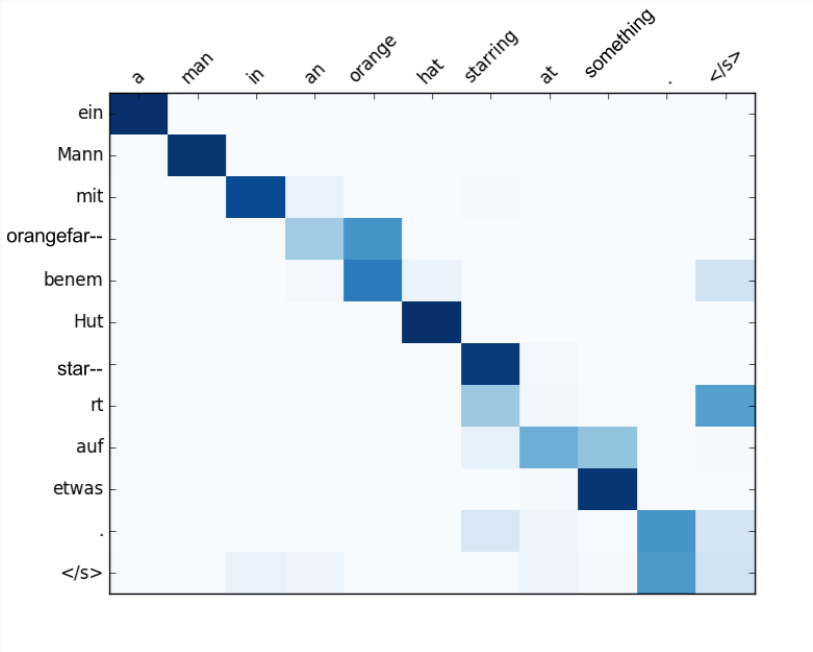
Model	Training data	BLEU4↑	METEOR↑	TER↓	chrF3↑	(prec. / recall)
NMT	$M30k_T$	33.7	52.3	46.7	65.2	(67.7 / 65.0)
PBSMT	$M30k_T$	32.9	54.3†	45.1†	**67.4**	(66.5 / 67.5)
Huang et al. (2016)	$M30k_T$	35.1 (↑ 1.4)	52.2 (↓ 2.1)	—	—	—
	+ RCNN	**36.5** (↑ 2.8)	54.1 (↓ 0.2)	—	—	—
$NMT_{SRC+IMG}$	$M30k_T$	**36.5**†‡	**55.0**†	**43.7**†‡	67.3	(66.8 / 67.4)
Improvements						
$NMT_{SRC+IMG}$ vs. **NMT**		↑ **2.8**	↑ **2.7**	↓ **3.0**	↑ 2.1	↓ 0.9 / ↑ 2.4
$NMT_{SRC+IMG}$ vs. **PBSMT**		↑ **3.6**	↑ 0.7	↓ **1.4**	↓ 0.1	↑ 0.3 / ↓ 0.1
$NMT_{SRC+IMG}$ vs. Huang		↑ 1.4	↑ 2.8	—	—	—
$NMT_{SRC+IMG}$ vs. Huang (+RCNN)		↑ 0.0	↑ 0.9	—	—	—
Pre-training data set: back-translated $M30k_C$ (in-domain)						
PBSMT (LM)	$M30k_T$	34.0 ↑ 0.0	**55.0**† ↑ 0.0	44.7 ↑ 0.0	**68.0**	(66.8 / 68.1)
NMT	$M30k_T$	35.5‡ ↑ 0.0	53.4 ↑ 0.0	43.3‡ ↑ 0.0	65.2	(67.7 / 65.0)
$NMT_{SRC+IMG}$	$M30k_T$	**37.1**†‡	54.5†	**42.8**†‡	66.6	(67.2 / 66.5)
$NMT_{SRC+IMG}$ vs. best PBSMT		↑ **3.1**	↓ 0.5	↓ **1.9**	↓ 1.4	↑ 0.4 / ↓ 1.6
$NMT_{SRC+IMG}$ vs. NMT		↑ **1.6**	↑ **1.1**	↓ **0.5**	↑ 1.4	↓ 0.5 / ↑ 1.5
Pre-training data set: WMT'15 English-German corpora (general domain)						
PBSMT (concat)	$M30k_T$	32.6	53.9	46.1	67.3	(66.3 / 67.4)
PBSMT (LM)	$M30k_T$	32.5	54.1	46.0	67.3	(66.0 / 67.4)
NMT	$M30k_T$	37.8 ↑ 0.0	56.7 ↑ 0.0	41.0 ↑ 0.0	69.2	(69.7 / 69.1)
$NMT_{SRC+IMG}$	$M30k_T$	**39.0**†‡	**56.8** ‡	**40.6** ‡	**69.6**	(69.6 / 69.6)
$NMT_{SRC+IMG}$ vs. best PBSMT		↑ **6.4**	↑ **2.7**	↓ **5.4**	↑ 2.3	↑ 3.3 / ↑ 2.2
$NMT_{SRC+IMG}$ vs. NMT		↑ **1.2**	↑ 0.1	↓ 0.4	↑ 0.4	↓ 0.1 / ↑ 0.5

German–English

Model	BLEU4↑	METEOR↑	TER↓	chrF3↑
PBSMT	32.8	34.8	43.9	61.8
NMT	38.2	35.8	40.2	62.8
$\text{NMT}_{\text{SRC+IMG}}$	**40.6**†‡	**37.5**†‡	**37.7**†‡	**65.2**
Improvements				
Ours vs. NMT	↑ **2.4**	↑ **1.7**	↓ **2.5**	↑ **2.4**
Ours vs. PBSMT	↑ **7.8**	↑ **2.7**	↓ **6.2**	↑ **3.4**
Pre-training data set: back-translated M30k$_{\text{C}}$ (in-domain)				
PBSMT	36.8	36.4	40.8	64.5
NMT	42.6	38.9	36.1	67.6
$\text{NMT}_{\text{SRC+IMG}}$	**43.2**‡	**39.0**‡	**35.5**‡	**67.7**
Improvements				
Ours vs. PBSMT	↑ **6.4**	↑ **2.6**	↓ **5.3**	↑ **3.2**
Ours vs. NMT	↑ 0.6	↑ 0.1	↓ 0.6	↑ 0.1

Example

(a) Image–target word alignments.



(b) Source–target word alignments.

Conclusions

Conclusions

- **multi-modal neural MT** model with **two separate attention mechanisms**;
- **consistent improvements** when translating from English into German and vice-versa;
- model can **efficiently exploit additional data** in pre-training, e.g. **back-translated** and **text-only**;
- **visual attention** seems to learn to **focus on one important aspect** of the image, and the model chooses when to use it in generating a word via the **image gate**;

Conclusions

- **multi-modal neural MT** model with **two separate attention mechanisms**;
- **consistent improvements** when translating from English into German and vice-versa;
- model can **efficiently exploit additional data** in pre-training, e.g. **back-translated** and **text-only**;
- **visual attention** seems to learn to **focus on one important aspect** of the image, and the model chooses when to use it in generating a word via the **image gate**;

Conclusions

- **multi-modal neural MT** model with **two separate attention mechanisms**;
- **consistent improvements** when translating from English into German and vice-versa;
- model can **efficiently exploit additional data** in pre-training, e.g. **back-translated** and **text-only**;
- **visual attention** seems to learn to **focus on one important aspect** of the image, and the model chooses when to use it in generating a word via the **image gate**;

Conclusions

- **multi-modal neural MT** model with **two separate attention mechanisms**;
- **consistent improvements** when translating from English into German and vice-versa;
- model can **efficiently exploit additional data** in pre-training, e.g. **back-translated** and **text-only**;
- **visual attention** seems to learn to **focus on one important aspect** of the image, and the model chooses when to use it in generating a word via the **image gate**;

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations. ICLR 2015*.

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.

References II

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057. JMLR Workshop and Conference Proceedings.

Thank you!