Using Images to Ground Machine Translation

Iacer Calixto

December 7, 2017

ADAPT Centre, School of Computing, Dublin City University
Dublin, Ireland.
iacer.calixto@adaptcentre.ie

Outline

Introduction

Introduction [1/2]

- 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., German).
  - 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., German).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction [1/2]

- 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., German).
  - 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., German).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction [1/2]

- 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., German).
  - 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., German).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction [1/2]

- 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., German).
  - 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., German).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction [1/2]

- 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., German).
  - 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., German).
  - multi-modal task (text and vision);
  - model is trained on image/target sentence pairs.

Introduction [1/2]

- 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., German).
    - 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., German).
    - multi-modal task (text and vision);
    - model is trained on image/target sentence pairs.

Introduction [2/2]

- 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., German).
    - 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.

Introduction [2/2]

- 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., German).
  - 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.

Introduction [2/2]

- 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., German).
  - 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.

Introduction [2/2]

- 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., German).
  - 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.

Use Cases

- Multi-Modal Machine Translation (MMT) use-cases:
  - localisation of product information in e-commerce, e.g. eBay, Amazon;
  - localisation of user posts and photos in social networks, e.g. Facebook, Instagram, Twitter;
  - translation of image descriptions in general;
  - translation of subtitles (video), etc.

Use Cases

- Multi-Modal Machine Translation (MMT) use-cases:
  - localisation of product information in e-commerce, e.g. eBay, Amazon;
  - localisation of user posts and photos in social networks, e.g. Facebook, Instagram, Twitter;
  - translation of image descriptions in general;
  - translation of subtitles (video), etc.

Use Cases

- Multi-Modal Machine Translation (MMT) use-cases:
  - localisation of product information in e-commerce, e.g. eBay, Amazon;
  - localisation of user posts and photos in social networks, e.g. Facebook, Instagram, Twitter;
  - translation of image descriptions in general;
  - translation of subtitles (video), etc.

Use Cases

- Multi-Modal Machine Translation (MMT) use-cases:
  - localisation of product information in e-commerce, e.g. eBay, Amazon;
  - localisation of user posts and photos in social networks, e.g. Facebook, Instagram, Twitter;
  - translation of image descriptions in general;
  - translation of subtitles (video), etc.

Use Cases

- Multi-Modal Machine Translation (MMT) use-cases:
  - localisation of product information in e-commerce, e.g. eBay, Amazon;
  - localisation of user posts and photos in social networks, e.g. Facebook, Instagram, Twitter;
  - translation of image descriptions in general;
  - translation of subtitles (video), etc.

Convolutional Neural Networks (CNN)

- Virtually all MMT and IDG models use pre-trained CNNs for image feature extraction;
- Illustration of the VGG19 network (Simonyan and Zisserman, 2014):

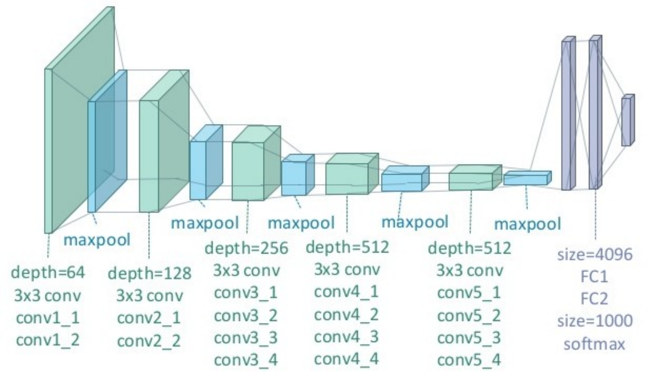
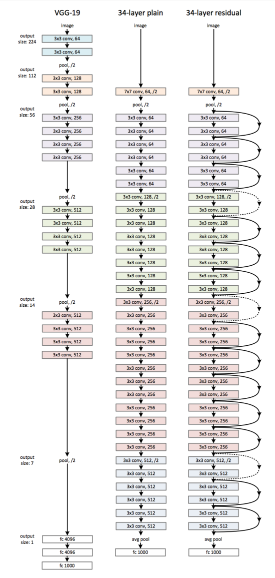
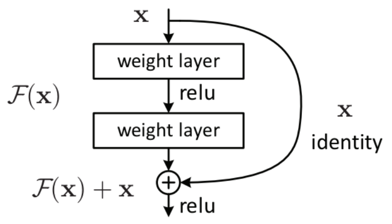


Figure 1: `https://goo.gl/y0So1l`

Example CNNs



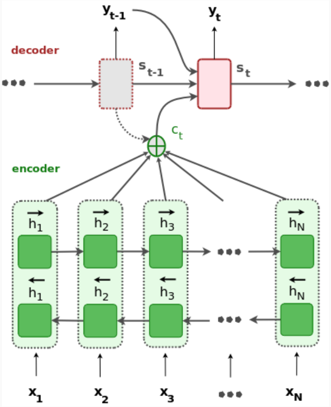


(b) Illustration of a residual connection (He et al., 2015).

(a) `https://goo.gl/jqQEvg`

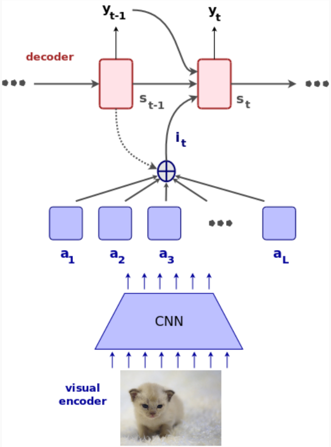
NMT and IDG Architectures

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



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.



Multi-modal MT Shared Task(s)

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

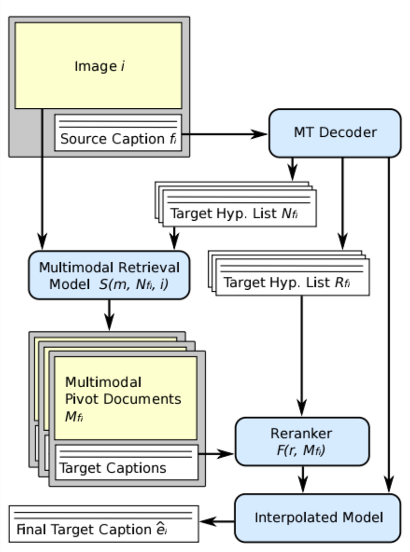
http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

Multimodal MT Shared Tasks – overall ideas

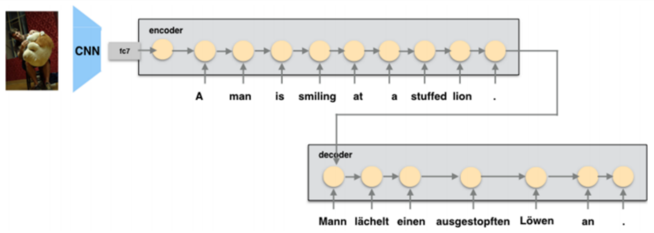
- 3 types of submissions:
  - Two attention mechanisms: compute context vectors over the source language hidden states and location-preserving image features;
  - Encoder and/or decoder initialisation: initialise encoder and/or decoder RNNs with bottleneck image features;
  - Other alternatives:
    - element-wise multiplication of the target-language embeddings with bottleneck image features;
    - sum source-language word embeddings with bottleneck image features;
    - use visual features in a retrieval framework;
    - visually-ground encoder representations by learning to predict bottleneck image features from the source-language hidden states.

http://mtm2017.unbabel.com/assets/images/slides/lucia_specia.pdf

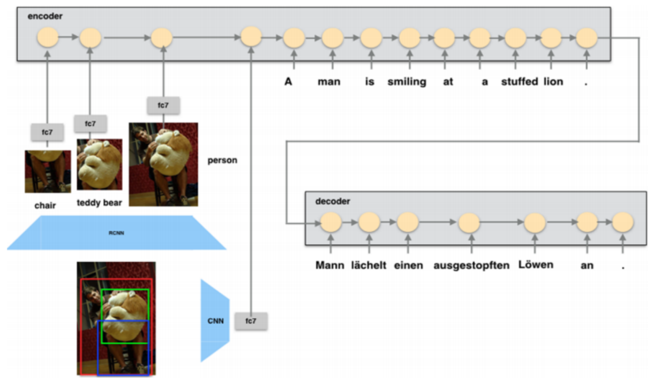
Heidelberg University (Hitschler et al., 2016)



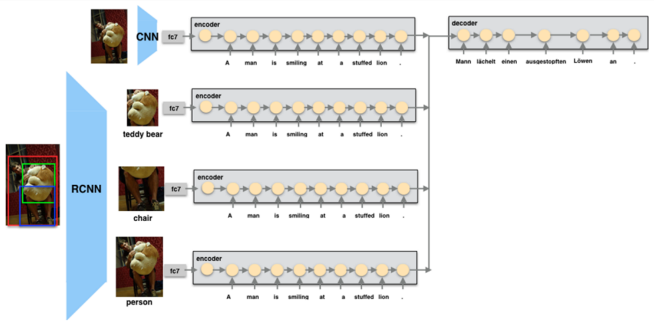
CMU (Huang et al., 2016) [1/3]



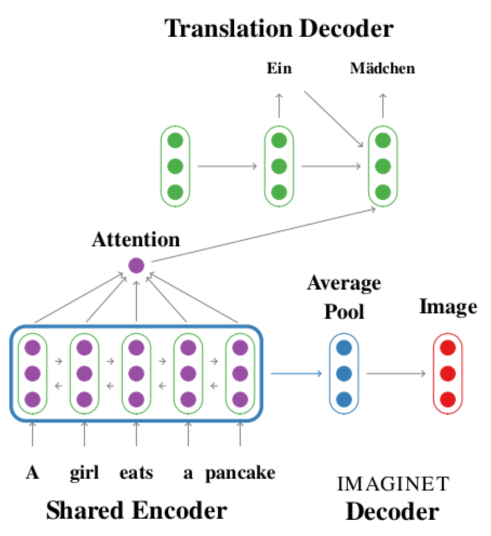
CMU (Huang et al., 2016) [2/3]



CMU (Huang et al., 2016) [3/3]



UvA-TiCC (Elliott and Kádár, 2017)



LIUM-CVC (Caglayan et al., 2017)

- Global visual features, i.e. 2048D `pool5` features from ResNet-50, are multiplicatively interacted with the target word embeddings;
- With 128D embeddings and 256D recurrent layers, their resulting models have $\sim$5M parameters.

(Elliott et al., 2017)

LIUM-CVC (Caglayan et al., 2017)

- Global visual features, i.e. 2048D `pool5` features from ResNet-50, are multiplicatively interacted with the target word embeddings;
- With 128D embeddings and 256D recurrent layers, their resulting models have $\sim$5M parameters.

(Elliott et al., 2017)

Our MMT Models

Doubly-Attentive Multi-Modal NMT – $\text{NMT}_{\text{SRC+IMG}}$

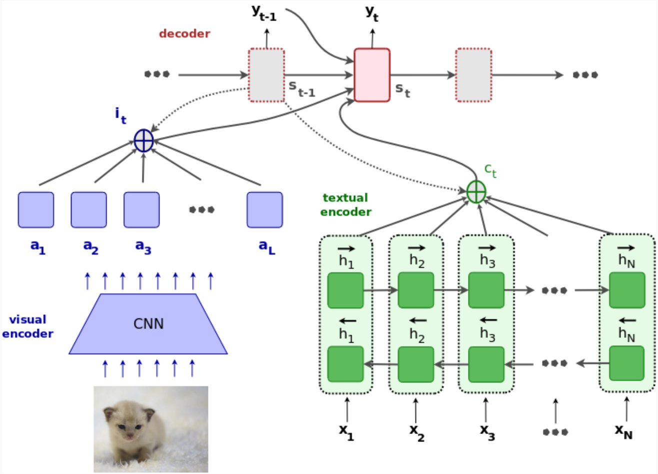


Figure 3: Doubly-Attentive Multi-modal NMT (Calixto et al., 2017a)

image gating

Image as source-language words – IMG_W

- **IMG_W** – **Global visual features** are projected into the **source-language word embeddings space** and used as the **first/last word in the source sequence**.

(Calixto et al., 2017b)

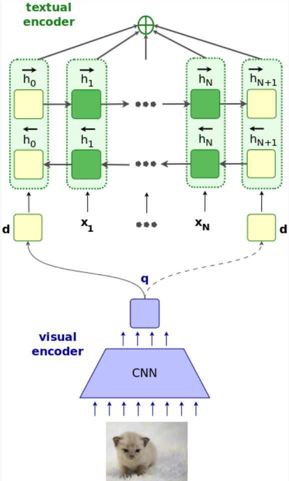


Image for encoder initialisation – IMG_E

- **IMG_E** – **Global visual features** are projected into the **source-language RNN hidden states space** and used to compute the **initial state of the source-language RNN**.

(Calixto et al., 2017b)

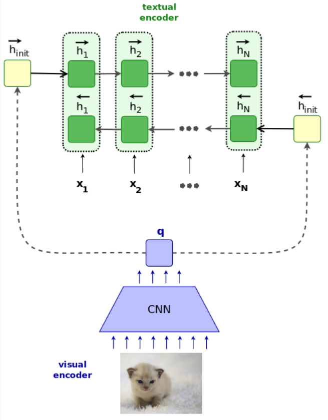
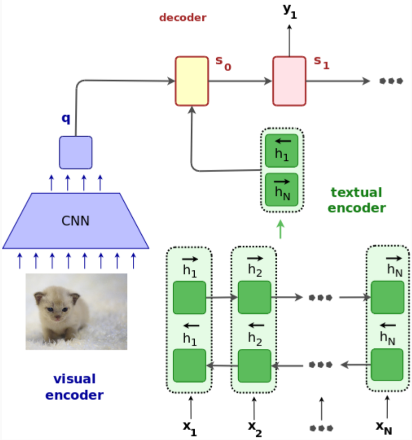


Image for decoder initialisation – IMG_D

- **IMG_D** – **Global visual features** are projected into the **target-language RNN hidden states space** and used as additional data to compute the **initial state of the target-language RNN**. (Calixto et al., 2017b)



Experiments

English–German [1/2]

- Training data: Multi30k data set (Elliott et al., 2016).

Model	**Training data**	**BLEU4↑**	**METEOR↑**	**TER↓**	**chrF3↑**
NMT	M30k$_T$	33.7	52.3	46.7	65.2
PBSMT	M30k$_T$	32.9	54.3†	45.1†	**67.4**
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†‡ (↑ 2.8)	55.0† (↑ 0.9)	43.7†‡ (↓ 1.4)	67.3 (↓ 0.1)
IMG$_W$	M30k$_T$	36.9†‡ (↑ **3.2**)	54.3 ‡ (↑ **0.2**)	**41.9**†‡ (↓ **3.2**)	66.8 (↓ 0.6)
IMG$_E$	M30k$_T$	37.1†‡ (↑ **3.4**)	55.0†‡ (↑ **0.9**)	43.1†‡ (↓ **2.0**)	67.6 (↑ 0.2)
IMG$_D$	M30k$_T$	**37.3**†‡ (↑ **3.6**)	**55.1**†‡ (↑ **1.0**)	42.8†‡ (↓ **2.3**)	**67.7** (↑ 0.3)

English–German [2/2]

- Pre-training on back-translated comparable Multi30k data set (Elliott et al., 2016).

Model	Training data	BLEU4↑	METEOR↑	TER↓	chrF3↑
PBSMT (LM)	M30k$_T$	34.0	**55.0**†	44.7	**68.0**
NMT	M30k$_T$	35.5‡	53.4	43.3‡	65.2
NMT$_{SRC+IMG}$	M30k$_T$	37.1†‡ (↑ 1.6)	54.5† (↓ 0.5)	42.8†‡ (↓ 0.5)	66.6 (↓ 1.4)
IMG$_W$	M30k$_T$	36.7†‡ (↑ 1.2)	54.6 ‡ (↓ 0.4)	42.0†‡ (↓ 1.3)	66.8 (↓ 1.2)
IMG$_E$	M30k$_T$	**38.5**†‡ (↑ 3.0)	55.7†‡ (↑ 0.9)	**41.4**†‡ (↓ 1.9)	68.3 (↑ 0.3)
IMG$_D$	M30k$_T$	**38.5**†‡ (↑ 3.0)	**55.9**†‡ (↑ 1.1)	41.6†‡ (↓ 1.7)	**68.4** (↑ 0.4)

German–English [1/2]

- Training data: Multi30k data set (Elliott et al., 2016).

Model	**BLEU4**↑	**METEOR**↑	**TER**↓	**chrF3**↑
PBSMT	32.8	34.8	43.9	61.8
NMT	38.2	35.8	40.2	62.8
$\text{NMT}_{\text{SRC+IMG}}$	$40.6^{\dagger\ddagger}$ **(↑ 2.4)**	$37.5^{\dagger\ddagger}$ **(↑ 1.7)**	$37.7^{\dagger\ddagger}$ **(↓ 2.5)**	65.2 **(↑ 2.4)**
IMG_{W}	$39.5^{\ddagger}$ **(↑ 1.3)**	**37.1**†‡ **(↑ 1.3)**	$37.1^{\dagger\ddagger}$ **(↓ 3.1)**	63.8 **(↑ 1.0)**
IMG_{E}	$41.1^{\dagger\ddagger}$ **(↑ 2.9)**	$37.7^{\dagger\ddagger}$ **(↑ 1.9)**	$37.9^{\dagger\ddagger}$ **(↓ 2.3)**	**65.7** **(↑ 2.9)**
IMG_{D}	**41.3**†‡ **(↑ 3.1)**	**37.8**†‡ **(↑ 2.0)**	$37.9^{\dagger\ddagger}$ **(↓ 2.3)**	**65.7** **(↑ 2.9)**

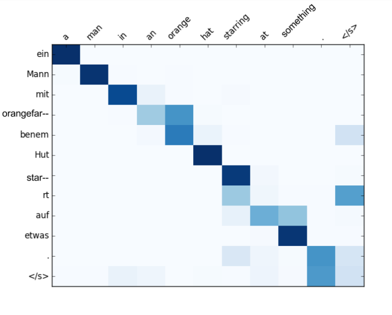
German–English [2/2]

- Pre-training on back-translated comparable Multi30k data set (Elliott et al., 2016).

Model	**BLEU4**↑	**METEOR**↑	**TER**↓	**chrF3**↑
PBSMT	36.8	36.4	40.8	64.5
NMT	<u>42.6</u>	<u>38.9</u>	<u>36.1</u>	<u>67.6</u>
$\text{NMT}_{\text{SRC+IMG}}$	43.2‡ (↑ 0.6)	39.0‡ (↑ 0.1)	35.5‡ (↓ 0.6)	67.7 (↑ 0.1)
IMG_{2W}	42.4 ‡ (↓ 0.2)	39.0 ‡ (↑ 0.1)	**34.7**†‡ **(↓ 1.4)**	67.6 (↑ 0.0)
IMG_{E}	**43.9**†‡ **(↑ 1.3)**	**39.7**†‡ **(↑ 0.8)**	34.8†‡ **(↓ 1.3)**	**68.6 (↑ 1.0)**
IMG_{D}	43.4 ‡ (↑ 0.8)	39.3 ‡ (↑ 0.4)	35.2 ‡ (↓ 0.9)	67.8 (↑ 0.2)

$NMT_{SRC+IMG}$ — Visualisation of attention states

(a) Image–target word alignments.



(b) Source–target word alignments.

References I

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.

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.

Calixto, I., Liu, Q., and Campbell, N. (2017a). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In Proceedings of the 55th Conference of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1913–1924, Vancouver, Canada.

Calixto, I. and Liu, Q. (2017b). Incorporating Global Visual Features into Attention-based Neural Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1003–1014, Copenhagen, Denmark.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. In Proceedings of the 5th Workshop on Vision and Language, VL@ACL 2016, Berlin, Germany.

Elliott, D., Kádár, Á. (2017). Imagination improves Multimodal Translation. arXiv preprint arXiv:1705.04350.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385.

Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2399–2409, Berlin, Germany.

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.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

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!

Questions?