
Multitask Learning from Multilingual Multimodal Data

Desmond Elliott
University of Copenhagen



6th Workshop on Asian Machine Translation
November 4th 2019

Contextual Language Understanding

- *“You shall know the meaning of a word by the company it keeps”* (Firth, 1957)

*Verb captures 2019
WKC Masters Agility
Grand Champion title*



Cross-lingual Grounded Understanding



Cross-lingual Grounded Understanding

gryde engelsk



All



Images



Maps



Shopping



Videos



More

Settings

Tools

About 378.000 results (0,36 seconds)

gryde - English translation - b

<https://en.bab.la> > dictionary > danish

Translation for 'gryde' in the free Danish-English

You visited this page on 10/22/19.

"GRYDE" ENGLISH TRANSLATION



"gryde" in English



gryde {en}

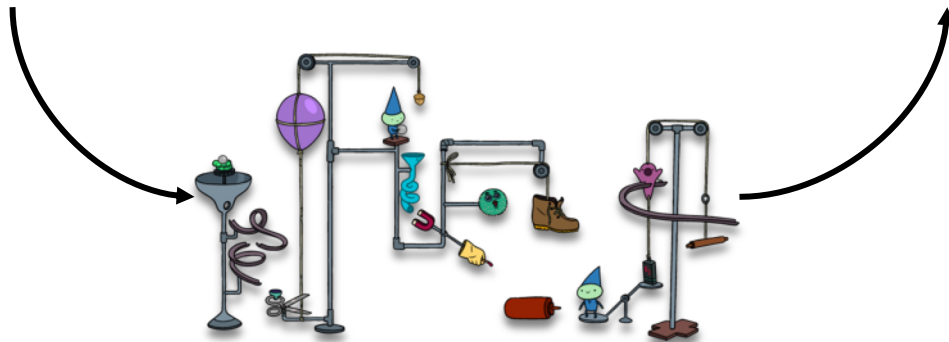


pan · saucepan · cauldron

Machine Translation: an NLP success story ^(mostly)

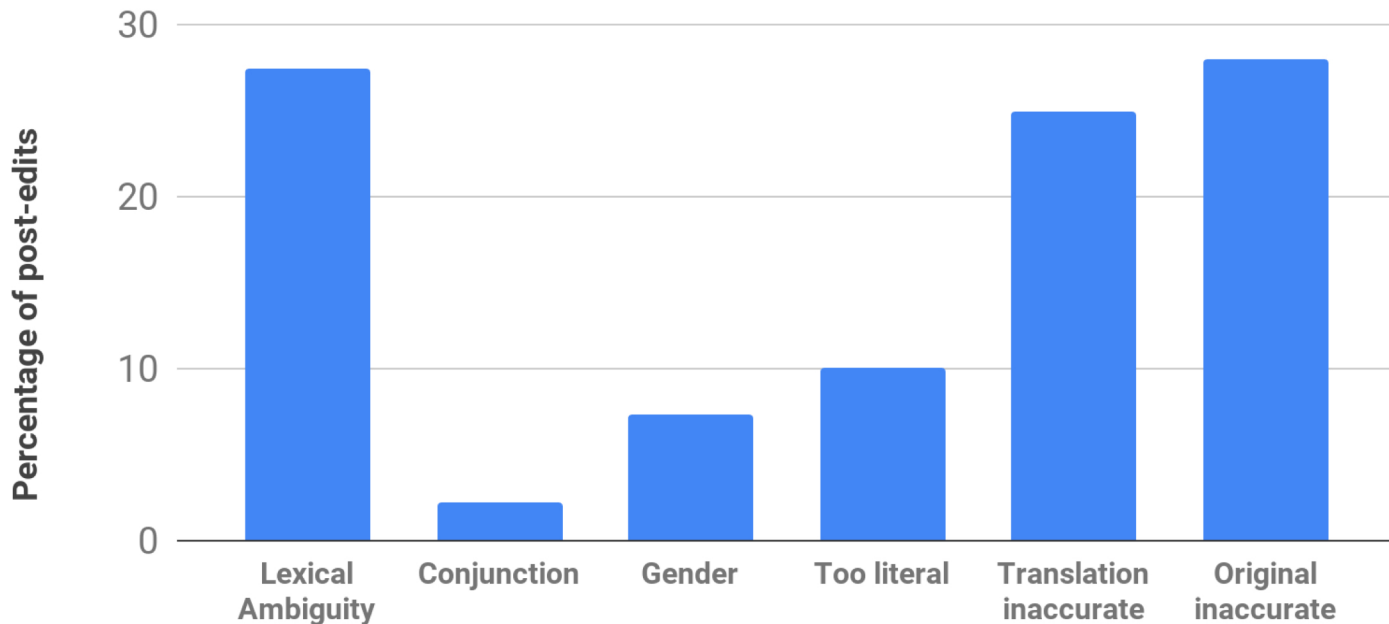
A baseball player in a black shirt just tagged a player in a white shirt.

Ein Baseballspieler in einem schwarzen Shirt fängt einen Spieler in einem weißen Shirt.



The Need for Visual Context in Translation

- Post-editing of translations in Multi30K (Frank et al. JNLE 2018)



Example of noun sense post-editing



En: Three children in football uniforms of two different teams are playing football on a football field.

De: Drei Kinder in Fußballtrikots zweier verschiedener Mannschaften spielen Fußball auf einem Fußballplatz.

Example of noun sense post-editing



En: Three children in football uniforms of two different teams are playing football on a football field.

De: Drei Kinder in Fußballtrikots zweier verschiedener Mannschaften spielen Fußball auf einem Fußballplatz.



PE: Drei Kinder in Footballtrikots zweier verschiedener Mannschaften spielen Football auf einem Footballplatz.

The Need for Multilingual Captions

- Speakers of different languages have different world knowledge



A **strange looking wood trailer** is parked in a street in front of stores.



Een **draaiorgel** in een winkelstraat met voetgangers.

(A **street organ** in a shopping street with pedestrians.)

Multimodal Machine Translation



A bird flies
over the water

Model

Ein Vogel fliegt
über das Wasser

Use Cases for Multimodal Translation

- Localised alt-text generation across the Web
- Richer e-commerce experiences
- Audio described movies for more languages



The Danish flag flying against a cloudy sky

Det danske flag vajende mod en blå himmel



Overview

1. *Multimodal Learning* for Multilingual NLP

Elliott and Kadar (IJCNLP 2017)

2. Understanding Multimodal Translation

Elliott (EMNLP 2018), Gella et al. (NAACL 2019), Chowdhury and Elliott (LANTERN 2019)

3. *Multilingual Learning* for Multimodal NLP

Kadar et al. (CoNLL 2018)

Multimodal learning for Multilingual NLP

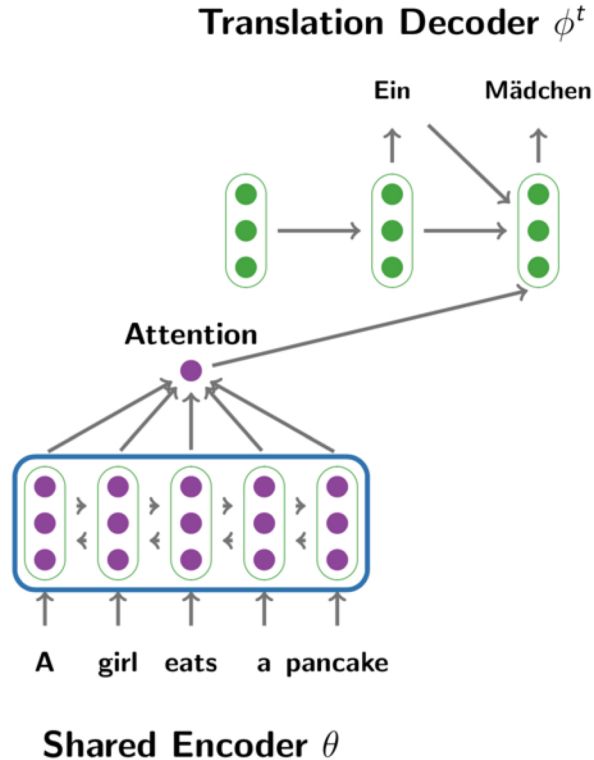


Elliott and Kádár
Imagination improves Multimodal Translation
IJCNLP 2017

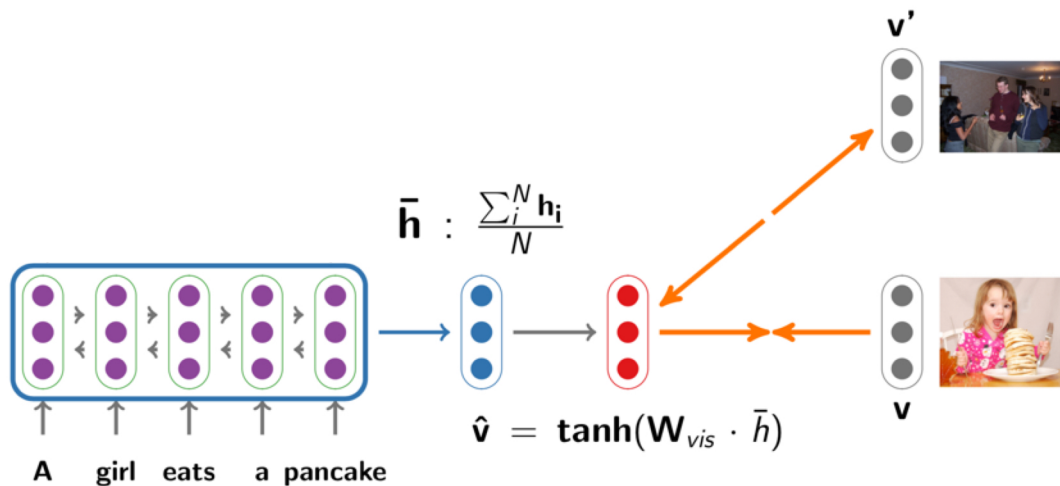
Decomposing Multimodal Translation

- Solve as two separate tasks:
 1. Learning to translate: $J_T(\theta, \phi^t)$
 2. Learning to ground: $J_G(\theta, \phi^g)$
- Multitask learning shared parameters (Caruana, 1997)
 - I. Are images necessary for inference?
 - II. How useful is external data for multimodal translation?

Task 1: Learning to Translate



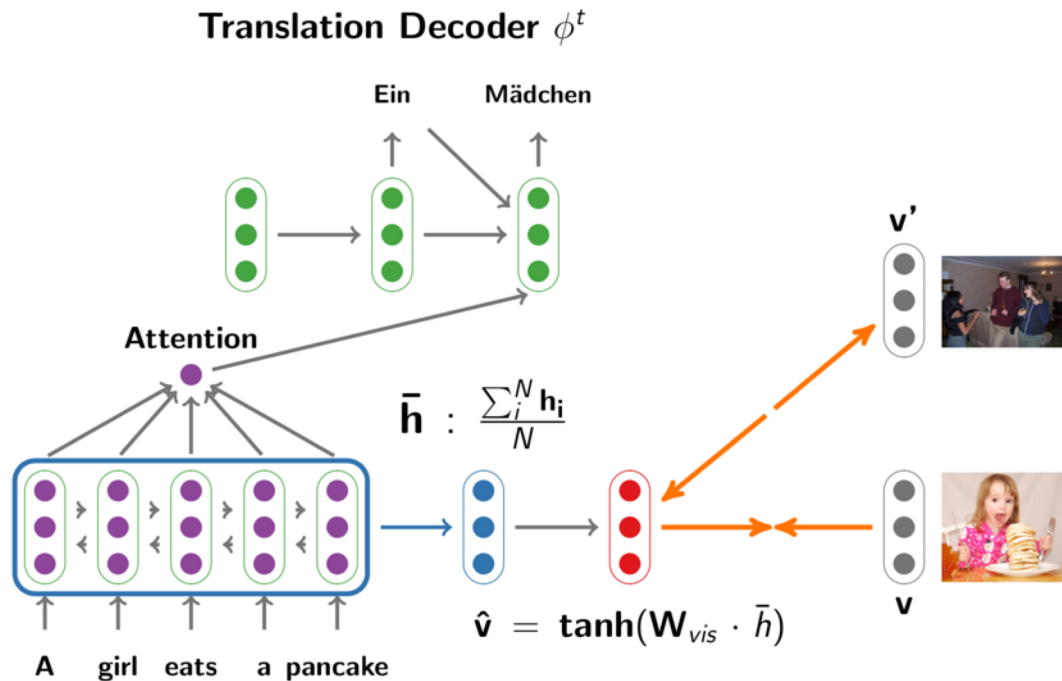
Task 2: Learning to Ground



Shared Encoder θ

IMAGENET

Joint Multitask Learning Model



Shared Encoder θ

IMAGINET

Optimisation

- Translation task:

$$J_T(\theta, \phi^t) = - \sum_j \log p(y_j | y_{<j}, x)$$

- Image prediction task:

$$J_G(\theta, \phi^g) = \sum_{\mathbf{v}' \neq \mathbf{v}} \max\{0, \alpha - \underbrace{\cos(\hat{\mathbf{v}}, \mathbf{v})}_{\text{Maximise similarity between true pair}} + \underbrace{\cos(\hat{\mathbf{v}}, \mathbf{v}')}_{\text{Minimise contrastive pair}}\}$$

- Joint objective:

$$J(\theta, \phi) = w \times J_T(\theta, \phi^t) + (1-w) \times J_G(\theta, \phi^g)$$

Data: Multi30K

- 32K English-captioned images with German, French, and Czech translations

A group of people are eating noodles.

Eine Gruppe von Leuten isst Nudeln.

Un groupe de gens mangent des nouilles.

Skupina lidí jedí nudle.



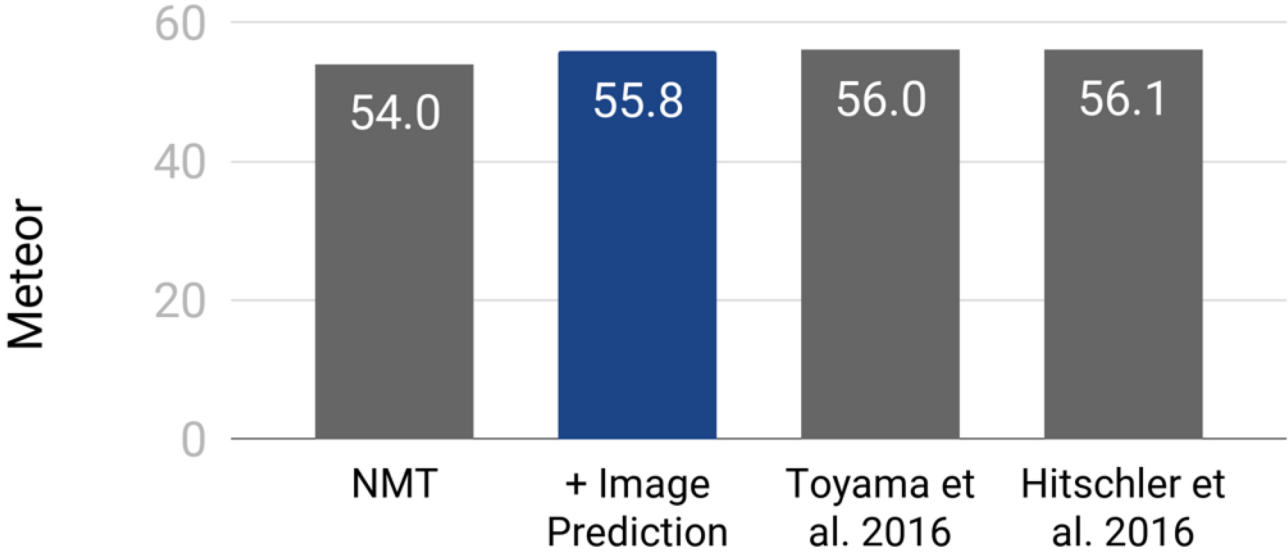
Translation: 32K Image-Sentence-Translation



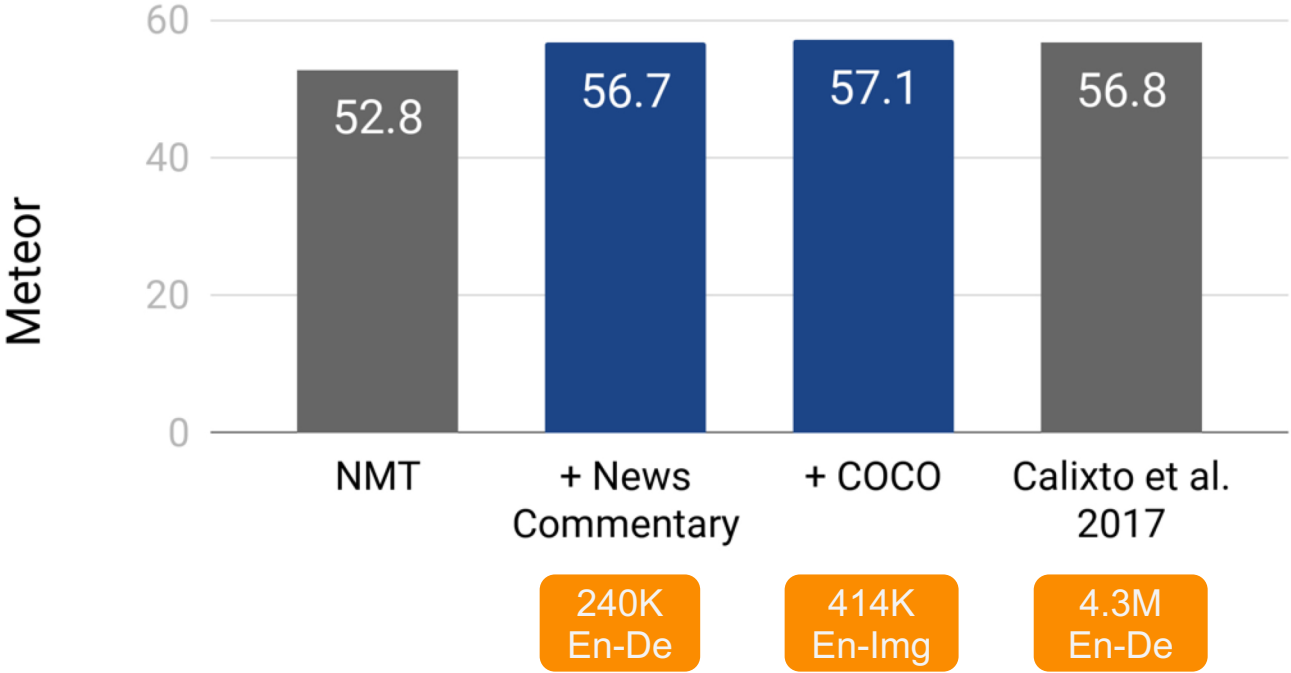
Comparable: 155K Independent Image-Sentence



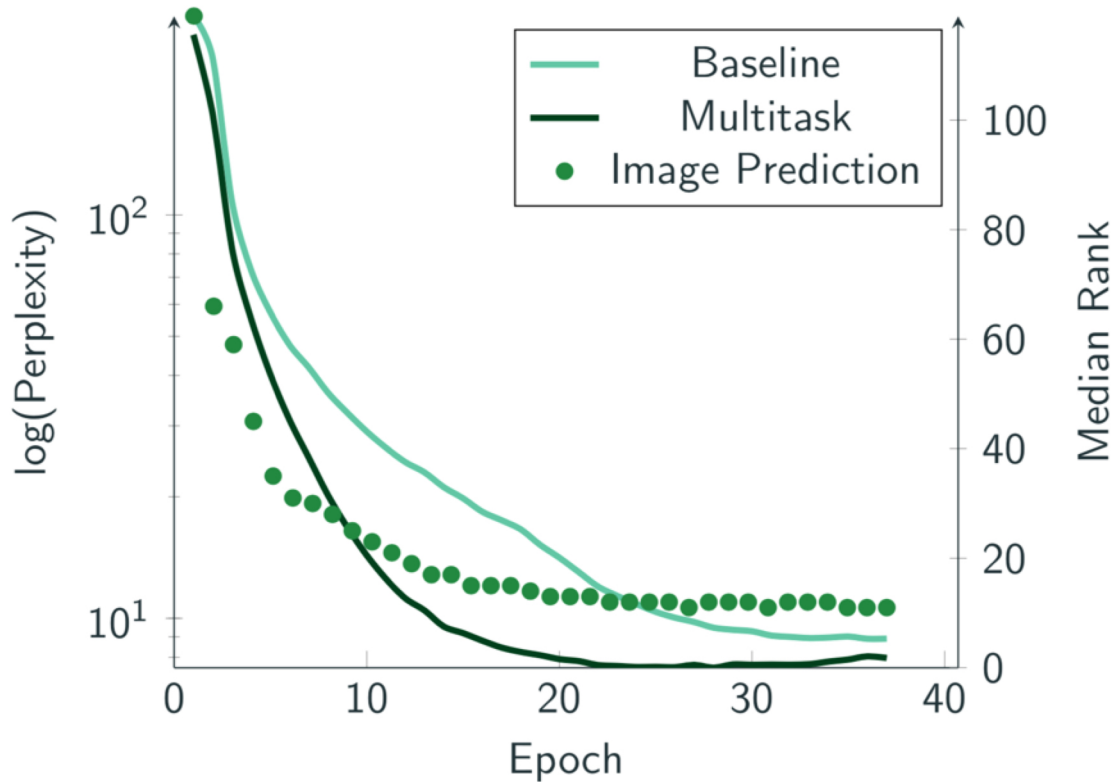
Image Prediction improves Translation



Further Improvements with External Data



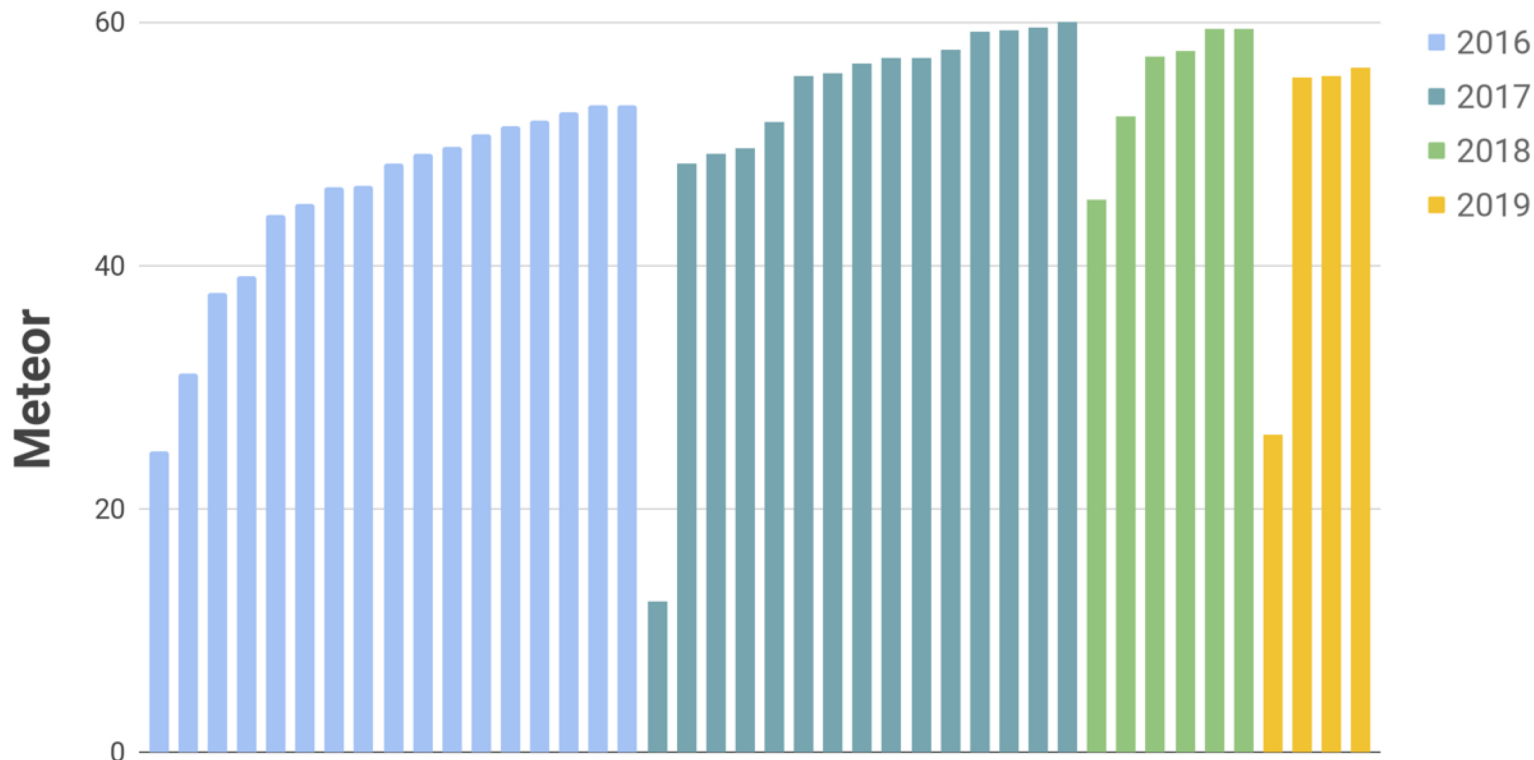
Why does MTL help?



Take-away messages

- Predicting the visual features during training improves multimodal translation
- Framework makes it easy to train with external parallel text or monolingual described images

Empirical progress on Mult30K



(Incomplete graph: not every paper reports performance on Test 2016)

Understanding Multimodal Translation

Elliott EMNLP 2018

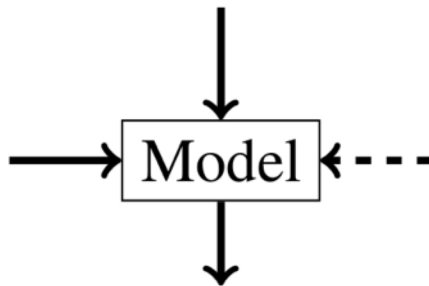
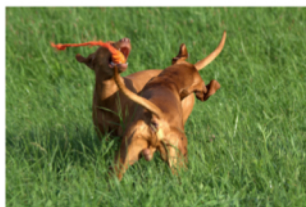
Gella et al. NAACL 2019

Chowdhury and Elliott LANTERN 2019



Adversarial Evaluation

Two dogs play with an orange toy in tall grass.



Zwei Hunde spielen im hohen Gras mit einem orangen Spielzeug.

Measuring Image Awareness

x Source language

y Target language

v Congruent image

\bar{v} Incongruent

\mathcal{E} Evaluation measure

$$a_{\mathcal{M}}(x, y, v, \bar{v}) = \underbrace{\mathcal{E}(x, v, y)} - \underbrace{\mathcal{E}(x, \bar{v}, y)}$$


Congruent:
should be better

Incongruent:
should be worse


- Train on standard Multi30K training set
- Evaluate on 5 random shuffles of En-De- $\overline{\text{Img}}$

Models

1. Visual modulation of the the target embeddings (Caglayan et al. 2017)

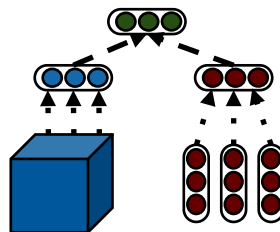
$$y_j = y_j \odot \tanh(\mathbf{W}_{\text{img}} \cdot V)$$


2. Initialise decoder with image features (Elliott et al. 2015; Caglayan et al. 2017; *inter-alia*)

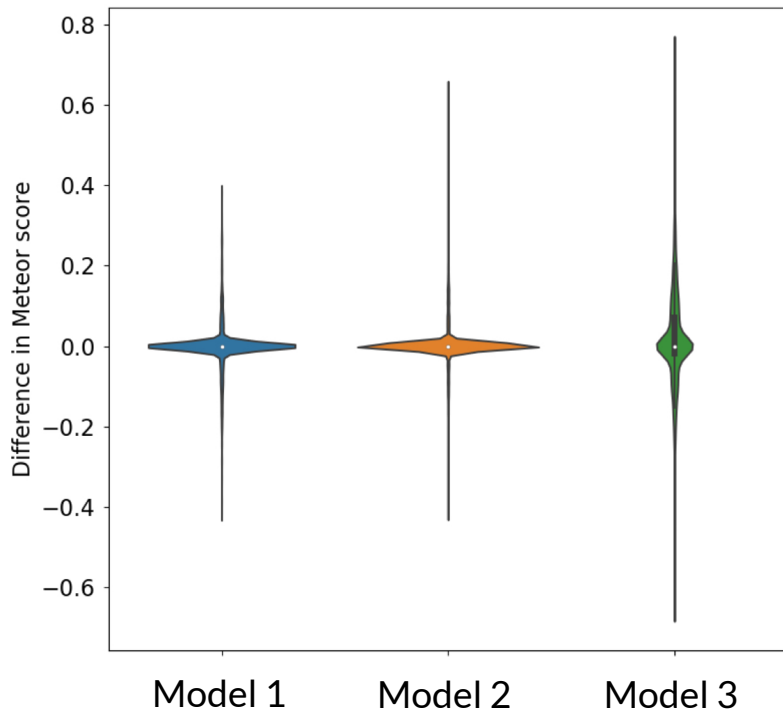
$$h_0 = \tanh(\mathbf{W}_{\text{img}} \cdot V)$$


3. Source and image hierarchical attention (Libovický and Helcl 2017; *inter-alia*)

$$c_i = \sum_{k=1}^N \beta_i^{(k)} U_c^{(k)} c_i^{(k)}$$



Results $a_{\mathcal{M}} := \text{Meteor}$

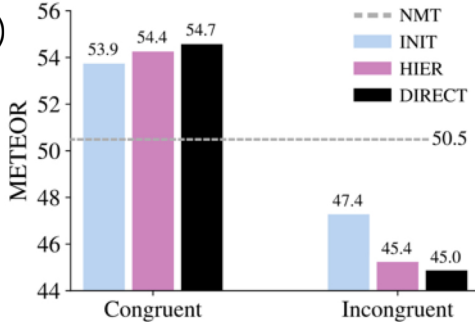


- Models 1 and 2 are **not affected** by the incongruent visual data.
- Model 3 is **more affected by incongruent visual data**. This may be because it calculates independent context vectors.

Understanding the Roles of the Modalities

- Train with entity and colour masking (Caglayan et al. NAACL19)

\mathcal{D}	a	lady	in	a	blue	dress	singing
\mathcal{D}_C	a	lady	in	a	[v]	dress	singing
\mathcal{D}_N	a	[v]	in	a	blue	[v]	singing



- Pre-trained models are sensitive to textual perturbations

(Chowdhury and Elliott LANTERN 2019; **Best Poster Award**)



A group of young people dressed up for halloween.
Eine Gruppe jünger Menschen verkleidet.

Two groups of young people dressed up for halloween.
*Zwei Gruppen von jungen Menschen **in Japan**.*



MultiSense: Towards Targeted Evaluations

- 995 images with ambiguous verb senses (Gella et al. NAACL 2019)
 - Measure verb sense accuracy and translation quality

*A large herd of sheep is
blocking the road.*



*Eine große Herde Schafe
blockiert die Straße.*

abdecken



verdecken



abblocken



blockieren

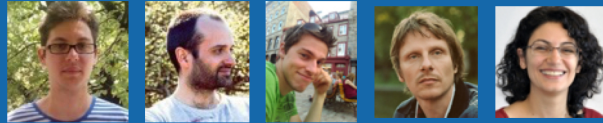


Senses for
“blocking” in
MultiSense

Take-away messages

- Awareness estimates the contribution of additional context in MMT models
- Textual and visual adversaries offer useful **hints** about the strengths of our models
- More effort in creating and evaluating models on challenging datasets

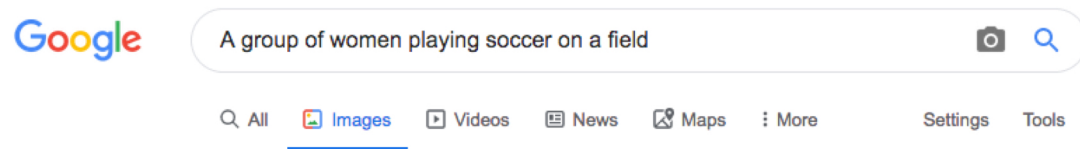
Multilingual learning for Multimodal NLP



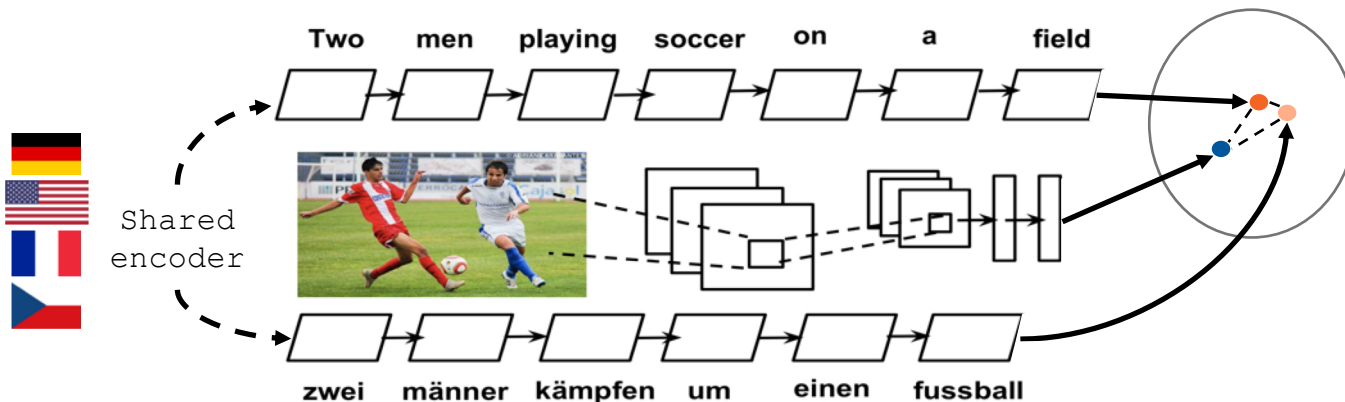
Kádár, Elliott, Côté, Chrupała, Alishahi.
Lessons learned in multilingual grounded language learning.
CoNLL 2018

Cross-modal retrieval

- Given a sentence, retrieve that it describes (and vice-versa)



Multi-task Multimodal Multilingual Model

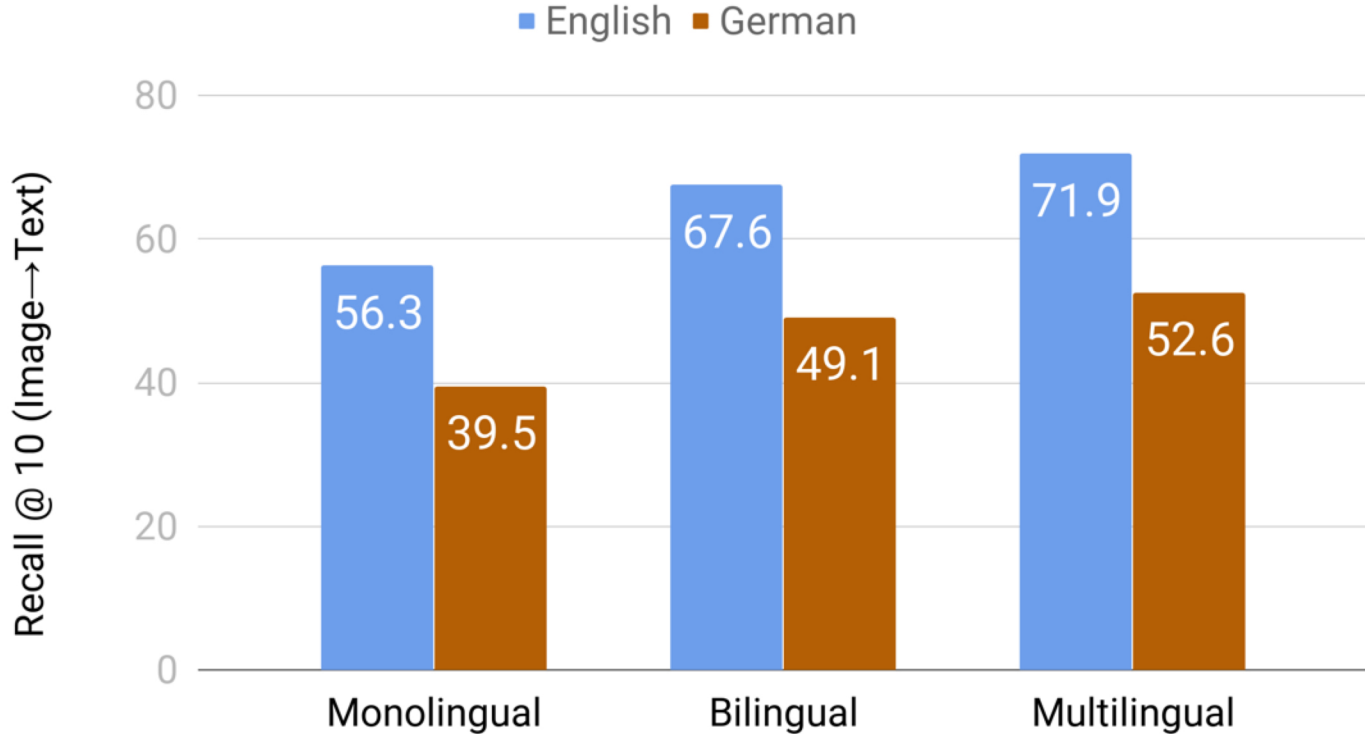


- Loss: $\mathcal{J}(a, b) = \max_{\langle \hat{a}, b \rangle} [\max(0, \alpha - s(a, b) + s(\hat{a}, b))] + \max_{\langle a, \hat{b} \rangle} [\max(0, \alpha - s(a, b) + s(a, \hat{b}))]$

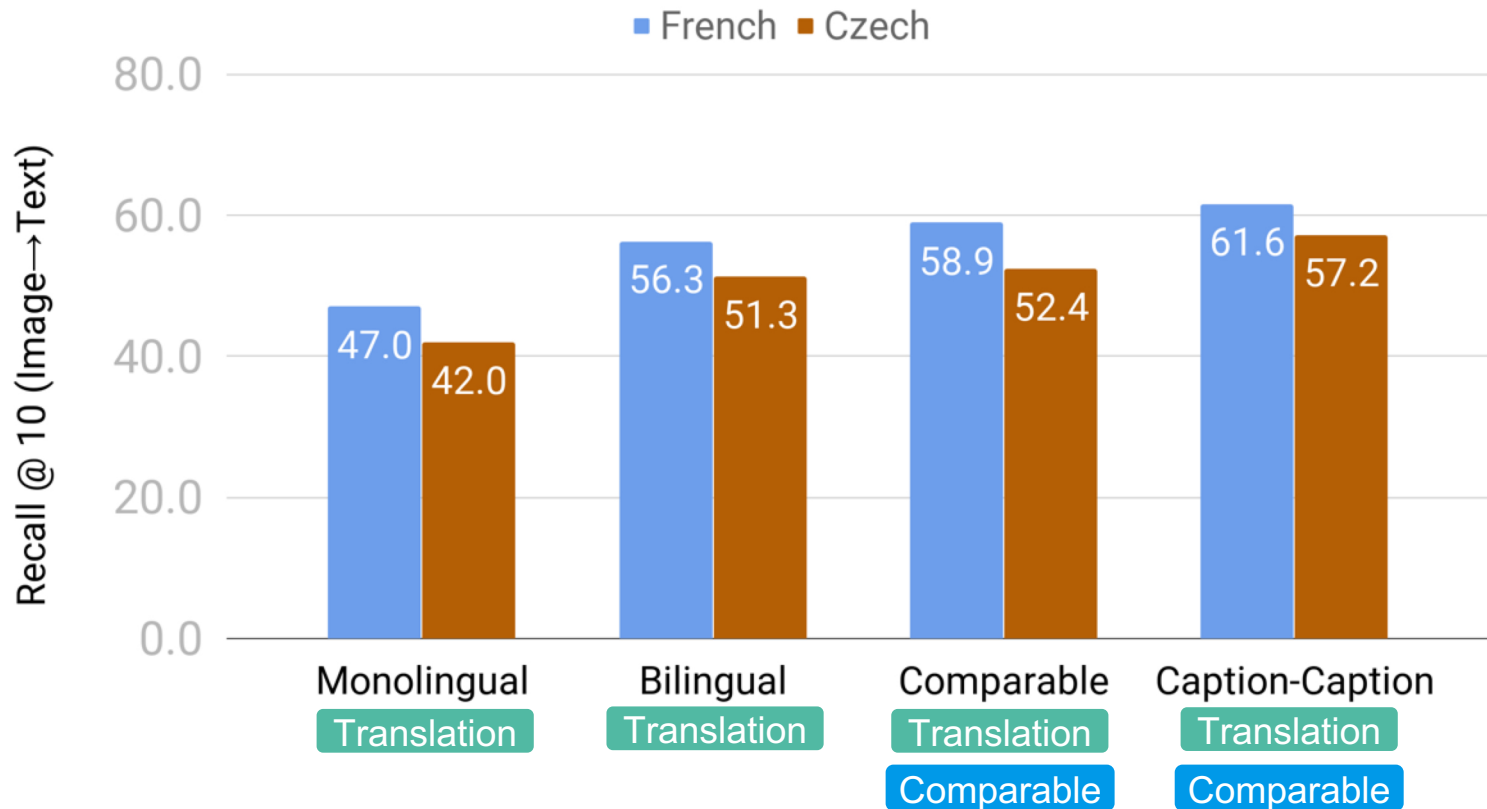
$\langle a, b \rangle$ Image - English **or** Image - German **or** German - English

When is multilingual data useful for cross-modal retrieval?

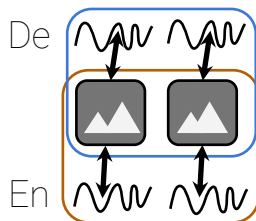
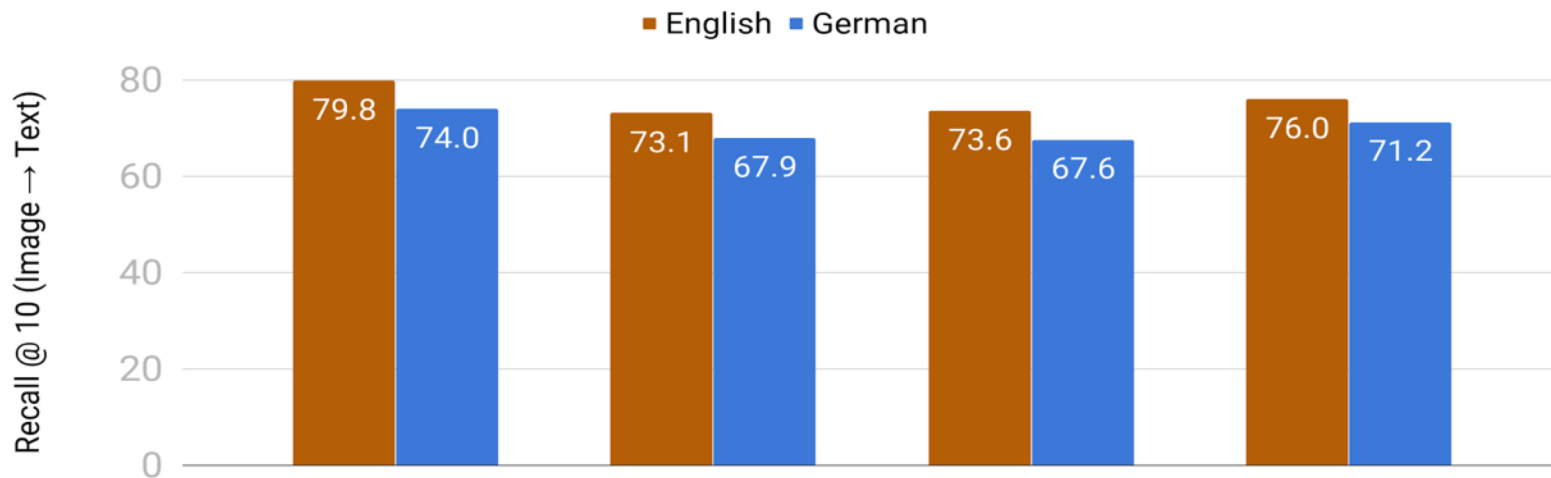
Multilingual data improves retrieval



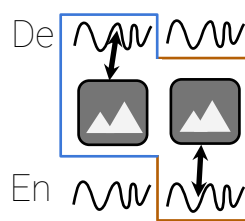
High-to-low multilingual resource transfer



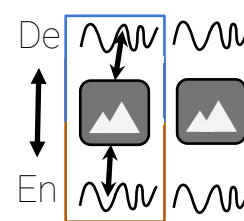
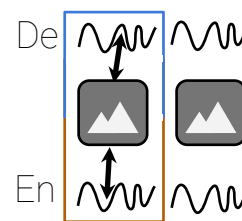
Controlling for 155K Training Data Points



Single-task
monolingual



Multitask and multilingual



Take-away messages

- Multilingual learning improves image- to-sentence and sentence-to-image retrieval
- Improvements hold in “low-resource” settings
- Caption-to-caption retrieval is an powerful additional objective: more data doesn't entirely explain the improvements

Some Open Problems

- Larger multilingual multimodal datasets (Sanabria et al. 2018, Wang et al. 2019)
- Naturally-occurring multilingual grounded data (e.g. Schamoni et al. 2018)
- Modelling audio, video, and text (e.g. Sanabria et al. 2018, Caglayan et al. 2019)
- Robustness to adversaries (e.g. Elliott 2018; Caglayan et al. 2019)
- Learning from unaligned data (e.g. Su et al. 2019)
- More linguistically diverse language pairs (e.g. Parida et al. 2019)
- Combining translation and ranking (e.g. Nikolaus et al. 2019)
- Multilingual learning with disjoint character sets

Acknowledgements

- Afra Alishahi
- Loic Barrault
- Fethi Bougarres
- Iacer Calixto
- Koel Dutta Chowdhury
- Grzegorz Chrupała
- Marc-Alexandre Côté
- Stella Frank
- Spandana Gella
- Eva Hasler
- Ákos Kádár
- Frank Keller
- Mirella Lapata
- Lucia Specia
- Khalil Sima'an
- Arjen de Vries



Final Conclusions

- Two views on multilingual multimodal data
 - Translation task: *multimodality* is useful
 - Retrieval task: *multilinguality* is useful
- **Multitask learning** was key to success
 - Jointly solve multiple tasks
 - Easily integrate external resources

Our Work

- S. Frank et al. [Assessing Multilingual Multimodal Image Description: Studies of Native Speaker Preferences and Translator Choices](#). JNLE 2018.
- D. Elliott, S. Frank, and E. Hasler. [Multilingual Image Description with Neural Sequence Models](#). arXiv cs.CL 1510.04709.
- L. Specia et al. [A Shared Task on Multimodal Machine Translation and Crosslingual Image Description](#). In WMT 2016.
- D. Elliott et al. [Multi30K: Multilingual English-German Image Descriptions](#). Workshop on Vision and Language.
- D. Elliott et al. [Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description](#). In WMT 2017.
- L. Barrault et al. [Findings of the Third Shared Task on Multimodal Machine Translation](#). In WMT 2018.
- D. Elliott. [Adversarial Evaluation of Multimodal Translation](#). In EMNLP 2018.
- S. Gella, D. Elliott, and F. Keller. [Cross-lingual Visual Verb Sense Disambiguation](#). In NAACL 2019.
- K. D. Chowdhury and D. Elliott. [Understanding the Effect of Textual Adversaries in Multimodal Machine Translation](#). In LANTERN 2019.
- Á. Kádár et al. [Lessons learned in multilingual grounded language learning](#). In CoNLL 2018.

References

- J. Hitschler, S. Schamoni, and S. Riezler. [Multimodal Pivots for Image Caption Translation](#). In ACL 2016.
- R. Caruana. [Multitask learning](#). Machine learning 28.1 (1997): 41-75.
- I. Calixto and Q. Liu. [Incorporating Global Visual Features into Attention-based Neural Machine Translation](#). In EMNLP 2017.
- J. Toyama, et al. [Neural machine translation with latent semantic of image and text](#). arXiv preprint arXiv:1611.08459.
- O. Caglayan, et al. [LIUM-CVC Submissions for WMT17 Multimodal Translation Task](#). In WMT 2017.
- J. Libovický and J. Helcl. [Attention Strategies for Multi-Source Sequence-to-Sequence Learning](#). In ACL 2017.
- Grönroos et al. [The MeMAD Submission to the WMT18 Multimodal Translation Task](#). In WMT 2018.
- O. Caglayan, et al. [Probing the Need for Visual Context in Multimodal Machine Translation](#). In NAACL 2019.
- C. Lala and Lucia Specia. [Multimodal lexical translation](#). In LREC 2018.
- S. Gella et al. [Image Pivoting for Learning Multilingual Multimodal Representations](#). In EMNLP 2017.
- F. Faghri, et al. [VSE++: Improving Visual-Semantic Embeddings with Hard Negatives](#). In BMVC 2018.
- S. Parida et al. [Hindi Visual Genome](#) arXiv:1907.08948.
- M. Nikolaus et al. [Compositional Generalization in Image Captioning](#). In CoNLL 2019.

Figure Credits

- Slides 3 and 4: https://www.pinclipart.com/pindetail/iiwhTw_prank-machine-rube-goldberg-machine-scissor-clipart/
- Slide 7: <https://www.pexels.com/photo/children-playing-soccer-2898317/>
- Slide 10: https://da.wikipedia.org/wiki/Fil:Dannebrog_3.jpg, <https://www.dyslexi.org/term/movietalk>
- Slide 23: travelandleisure.com; modified with the GIMP
- Slides 24 and 24: Emojis by [Jyoti Vyas](#) CC-BY
- Slide 29: https://www.amazon.com/dp/B06X9YW32T/?tag=097-20&ascsubtag=v7_1_4_62m_wd1_4_x01_-srt-,
<https://www.extremetech.com/extreme/135490-why-cut-down-on-carbon-emissions-when-you-can-curb-climate-change-by-simply-blotting-out-the-sun>,
- Slide 32: <https://worldjusticeproject.org/photo-essays/brazil-female-warriors-fight-level-playing-field>
- Slide 33: Modified version of Gella et al. EMNLP 2017