

Learning from Multilingual Multimodal Data

Desmond Elliott

Dagstuhl Seminar 19021, 8 January 2019



Machine translation works in practice ...

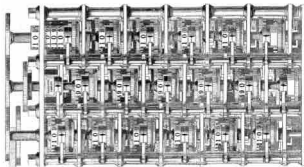
A baseball player in a black shirt just tagged a player in a white shirt.

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



... but multimodality can help to resolve ambiguities

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



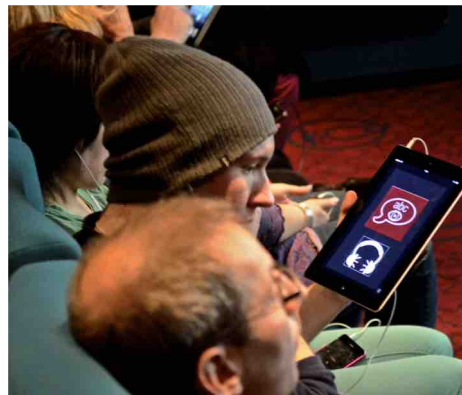
Applications for Multilingual Multimodal Models

- Localised alt-text generation across the Web
- Image search and retrieval
- Audio described movies for more languages



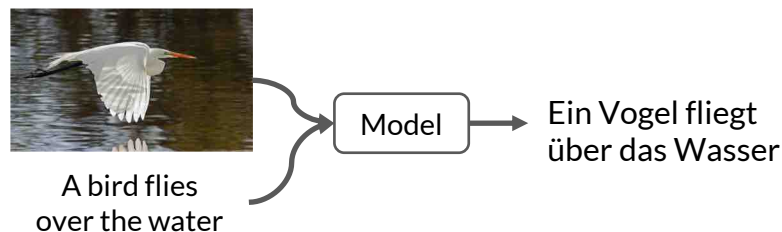
The Danish flag flying against a cloudy sky

Det danske flag vajende mod en blå himmel

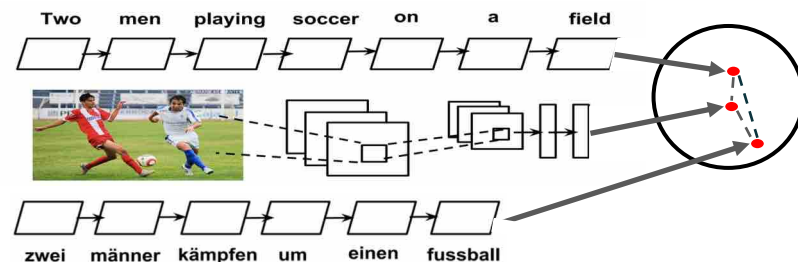


This talk

1. Multimodal machine translation



2. Multilingual image-sentence retrieval



Multimodal machine translation



Elliott and Kádár.
Imagination Improves Multimodal Translation.
IJCNLP 2017

Problem Formulation



Elliott, Frank, Hasler (2015)

- Data $\in (x, y, v)$:
 - x is a description of image v
 - y is a translation of x



A brown dog is running
after the black dog.

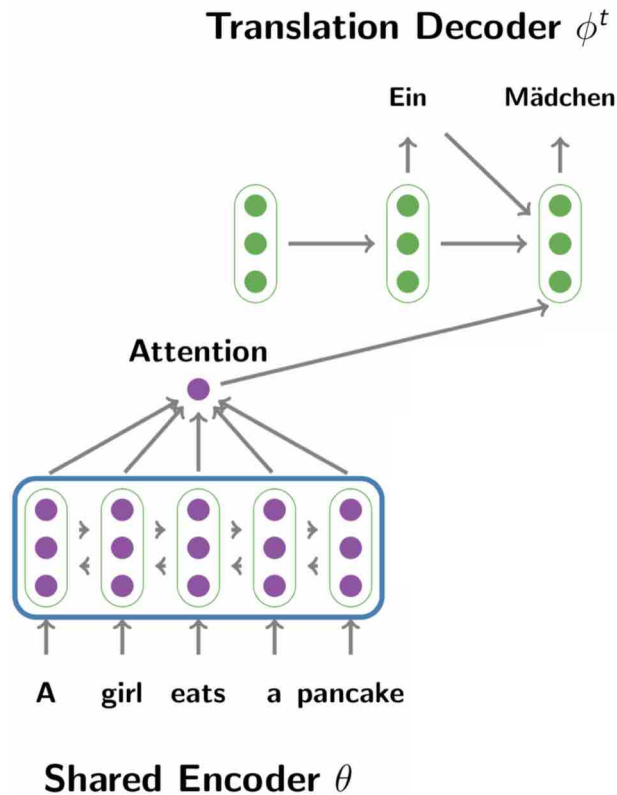
Ein brauner Hund rennt
dem schwarzen Hund
hinterher

- Task: Generate best \hat{y} , given x and v .
- Evaluation: Meteor (Denkowski and Lavie, 2014)

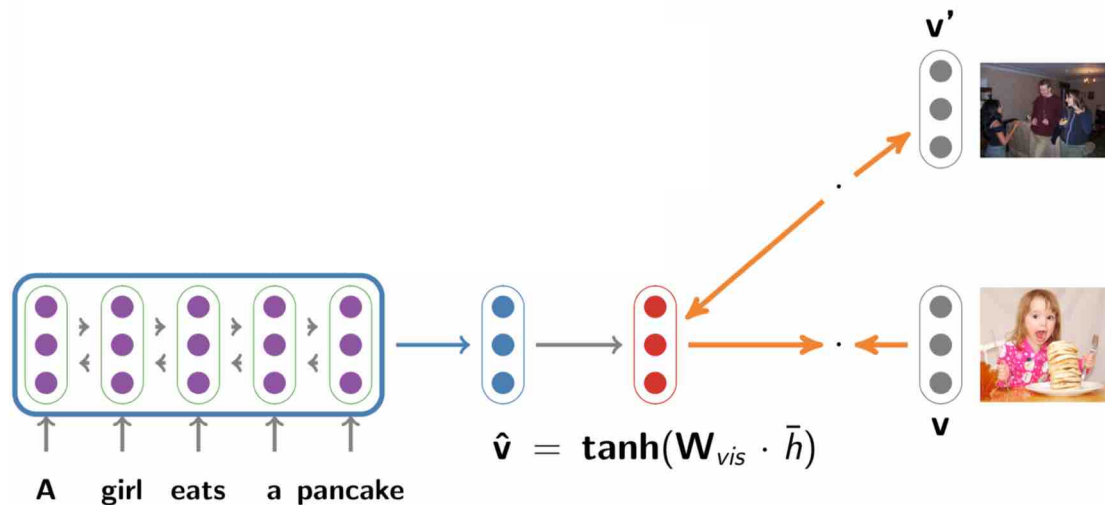
Decomposing Multimodal Translation

- Decompose the problem into two steps:
 1. Learning to translate: $J_T(\theta, \phi^t)$
 2. Learning to ground: $J_G(\theta, \phi^g)$
 - Use external resources for each problem
- Multitask learning shared parameters θ (Caruana, 1997)

Model: Learning to Translate: $J_T(\theta, \phi^t)$



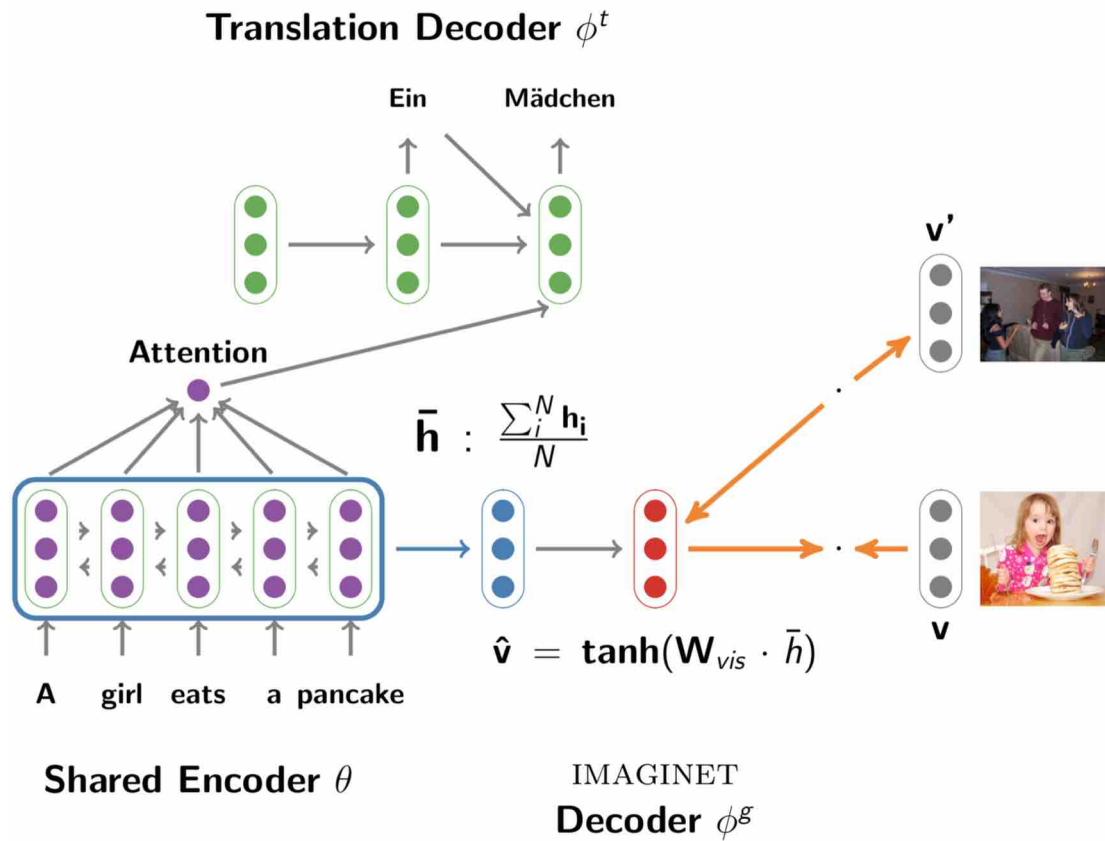
Model: Learning to ground: $J_G(\theta, \phi^g)$



Shared Encoder θ

IMAGINET
Decoder ϕ^g

Multitask Learning Model



Objectives

Translation model:

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

Image prediction model:

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

Data: Multi30K



Elliott et al. (V&L 2016, WMT 2017)
Barrault et al. WMT 2018

- 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 jí nudle.



Related Work

Models

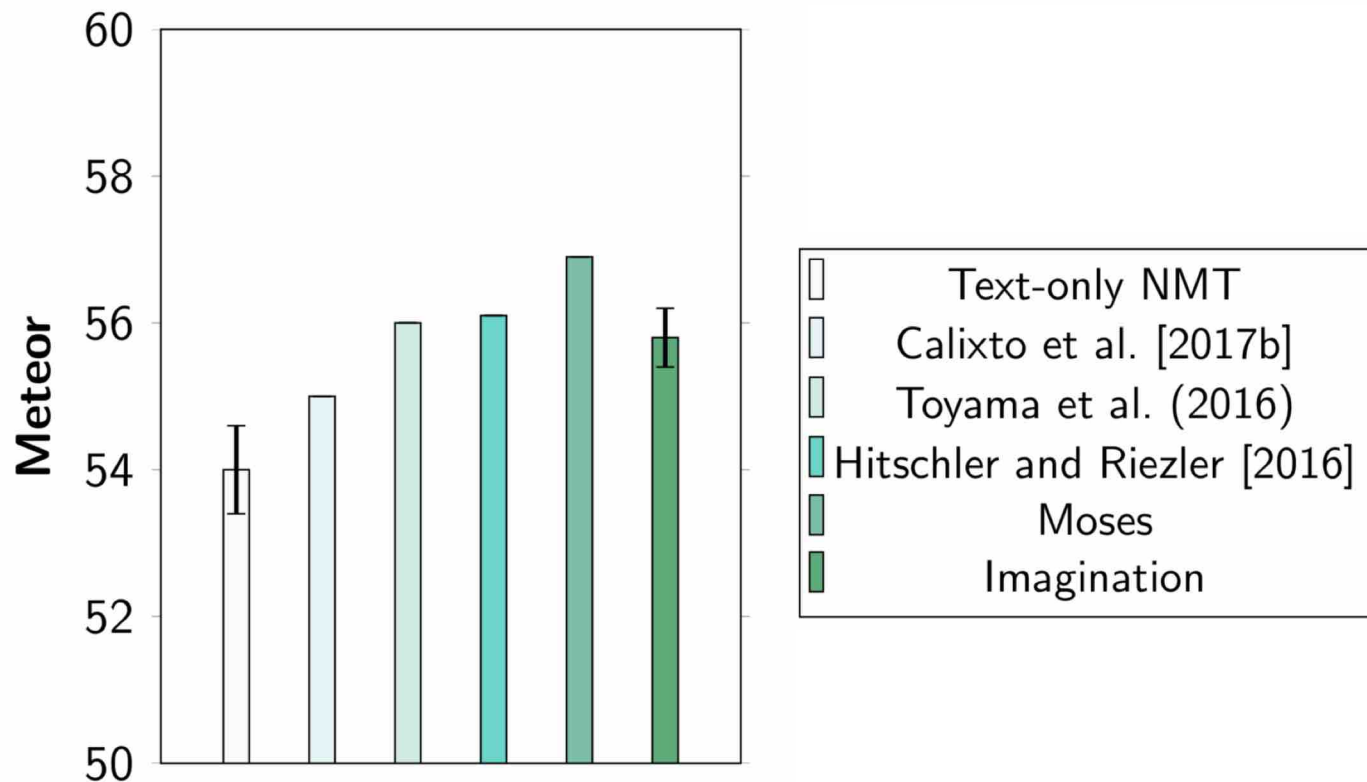
- Sentence-to-image prediction for word similarity and image retrieval (Chrupala et al. ACL 2015)
- Word-to-image prediction for word similarity and zero-shot image retrieval (Collell et al. AACL 2017)
- Video description with video prediction and lexical entailment (Pasunuru and Bansal, EMNLP 2017)
- Related caption prediction and image prediction (Kielar et al. NAACL 2018)

Data

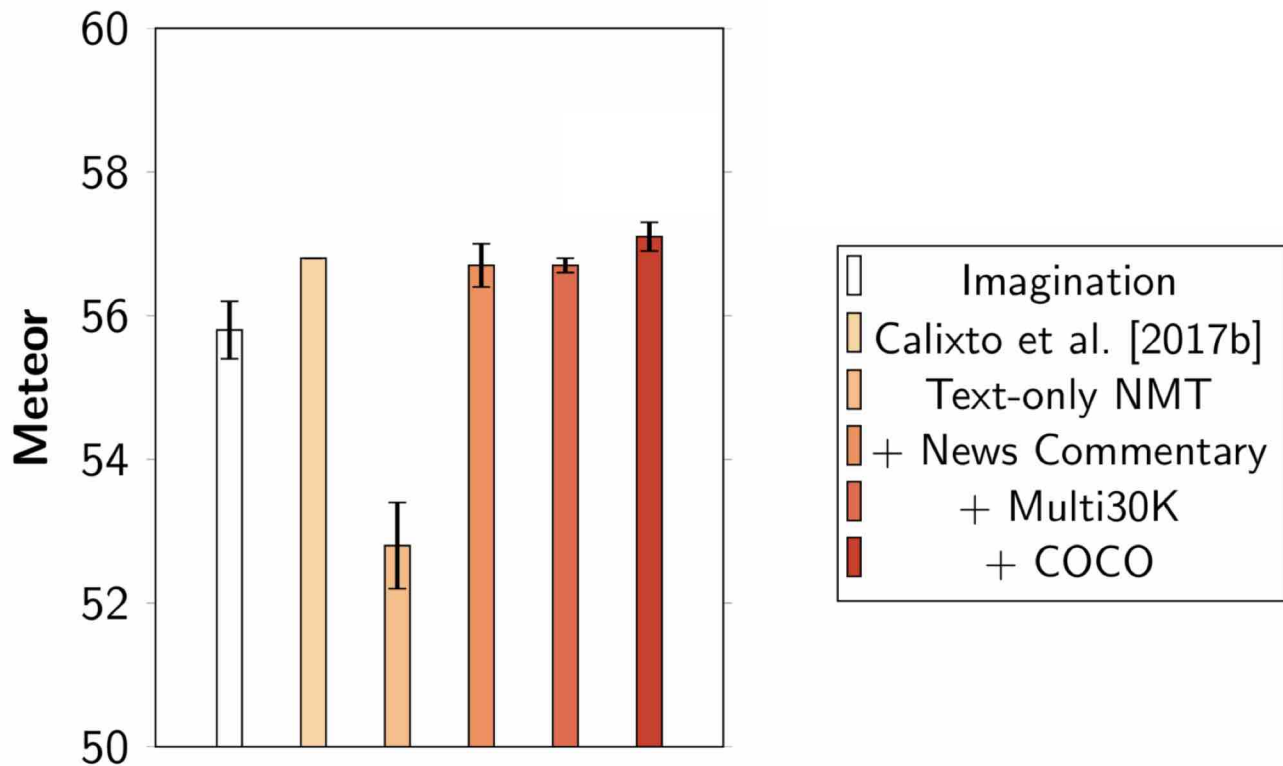
- Turkish Flickr8K (Unal et al. SIU 2016)
- Chinese Flickr8K (Li et al. MM 2016)
- Japanese extension of COCO (Yoshikawa et al. ACL 2017)
- How2: 300 hours of instructional videos with Portuguese translations (Sanabria et al. NeurIPS ViGIL 2018)

See Frank et al. (NLE 2018)
for a more comprehensive
overview of related datasets.

Image Prediction Improves Translation



Further Improvements with External Resources



Conclusions

- Image representation prediction helps multimodal translation
- Easy to train with external data
 - Improvements with out-of-domain
 - Newswire parallel text
 - Crowdsourced described images

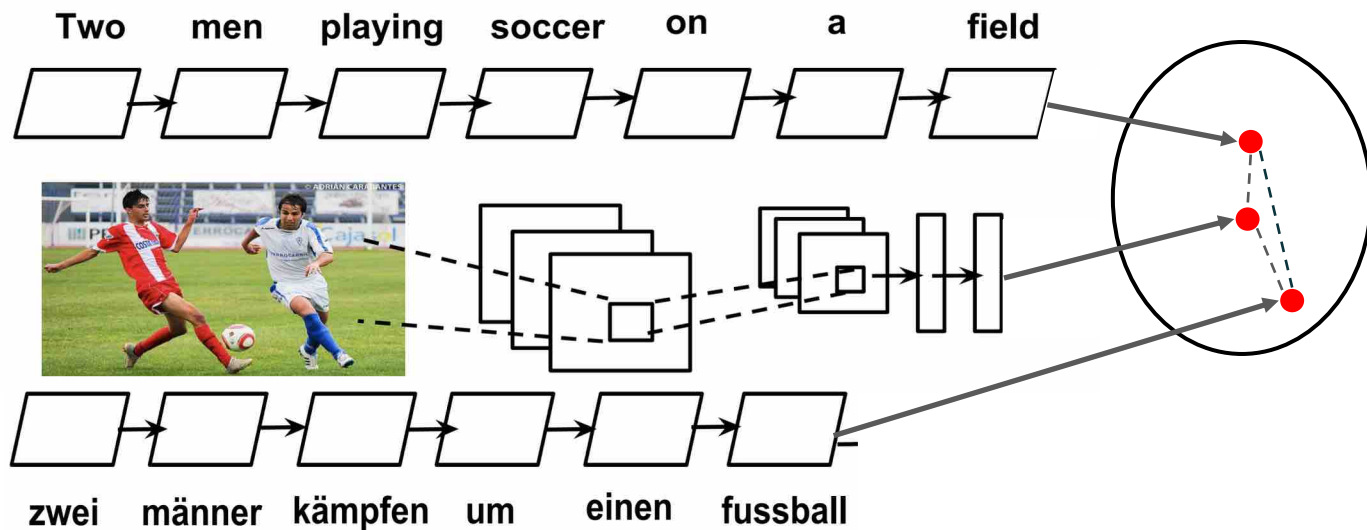
Multilingual image - sentence retrieval



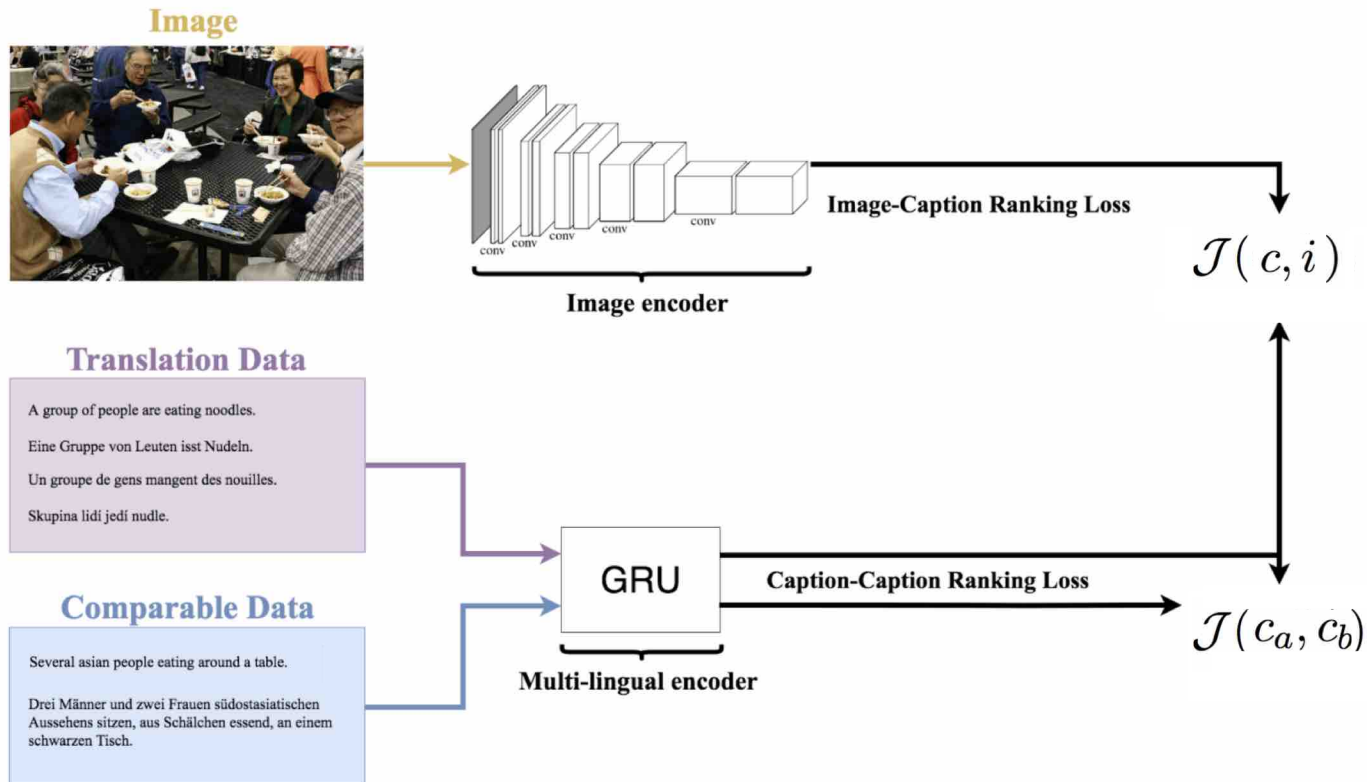
Kádár, Elliott, Côté, Chrupała, Alishahi.
Lessons learned in multilingual grounded language learning.
CoNLL 2018

Problem Formulation

- Given an image, retrieve its sentence from a shared space (and vice-versa)
- Evaluation: Recall@K, Median Rank



Model (following Gella et al. 2017)



Training

while not stopping criterion **do**

$T \sim \text{Bern}(p)$

if $T = 1$ **then**

$D_n \sim \mathcal{D}_{c2i}$

$\langle c, i \rangle \sim D_n$

$\mathbf{a} \leftarrow \phi(c, \theta_\phi)$

$\mathbf{b} \leftarrow \psi(i, \theta_\psi)$

else

$\langle c_a, c_b \rangle \sim \mathcal{D}_{c2c}$

$\mathbf{a} \leftarrow \phi(c_a, \theta_\phi)$

$\mathbf{b} \leftarrow \phi(c_b, \theta_\phi)$

end if

$[\theta_\phi; \theta_\psi] \leftarrow \text{SGD}(\nabla_{[\theta_\phi; \theta_\psi]} \mathcal{J}(\mathbf{a}, \mathbf{b}))$

end while

- Choose a task T
- \mathcal{D}_{c2i} : image--caption datasets
- $\phi(c, \theta_\phi)$: sentence encoder
- $\psi(i, \theta_\psi)$: image encoder
- \mathcal{D}_{c2c} : caption--caption datasets

Gella et al. (EMNLP 2017)

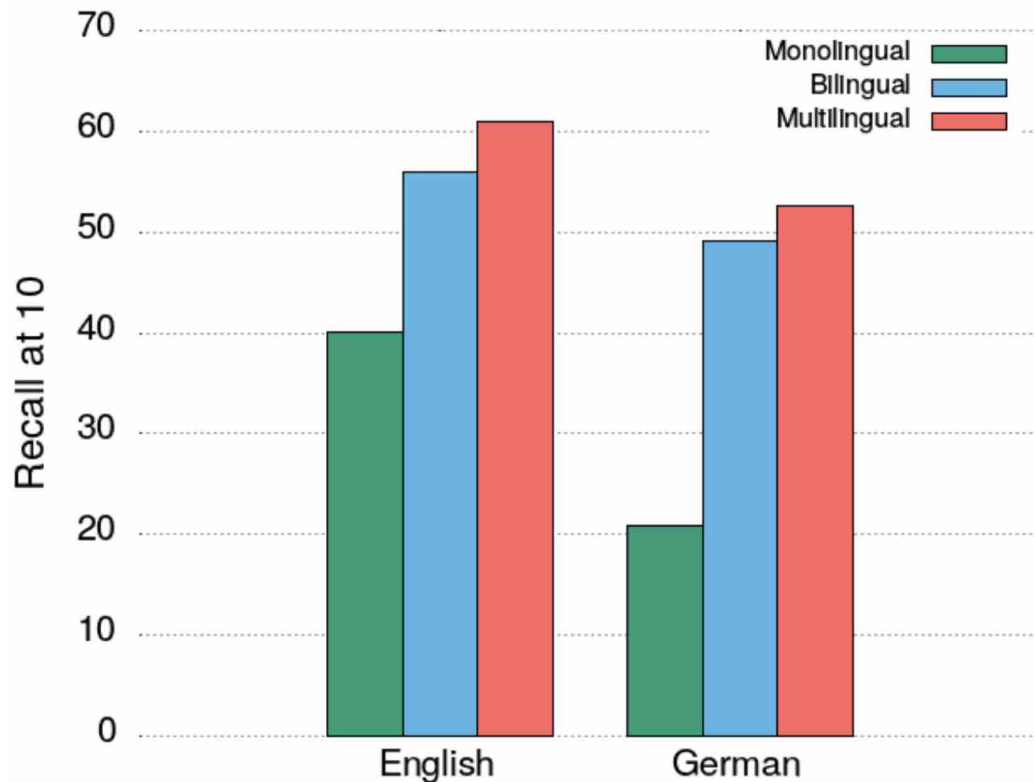
$$\mathcal{J}(\mathbf{a}, \mathbf{b}) = \max_{c'} [\alpha + s(i, c') - s(i, c)]_+ + \max_{i'} [\alpha + s(i', c) - s(i, c)]_+$$

Faghri et al. (BMVC 2018)

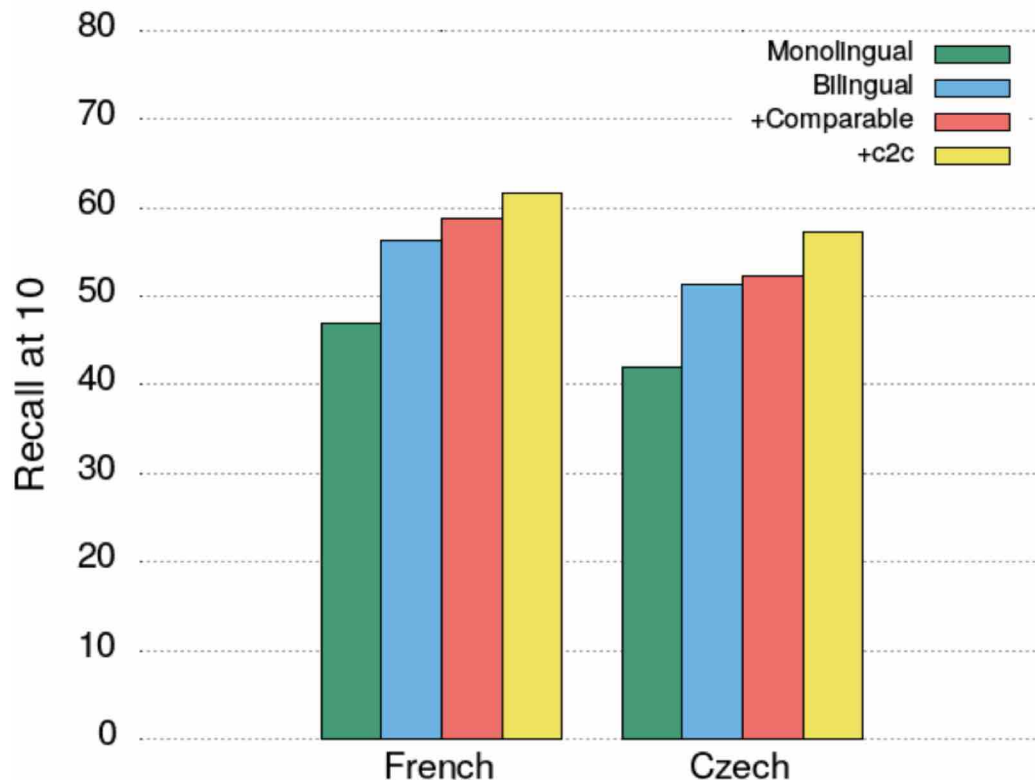
Related Work

- Image—sentence ranking with KCCA (Hodosh et al. JAIR 2013)
- Ranking with dependency tree recursive neural nets (Socher et al. TACL 2014)
- Order-embeddings for ranking (Kiros et al. ICLR 2015)
- Bilingual ranking with caption—caption objective (Gella et al. EMNLP 2017)
- Max-of-hinges training for ranking models (Faghri et al. BMVC 2018)

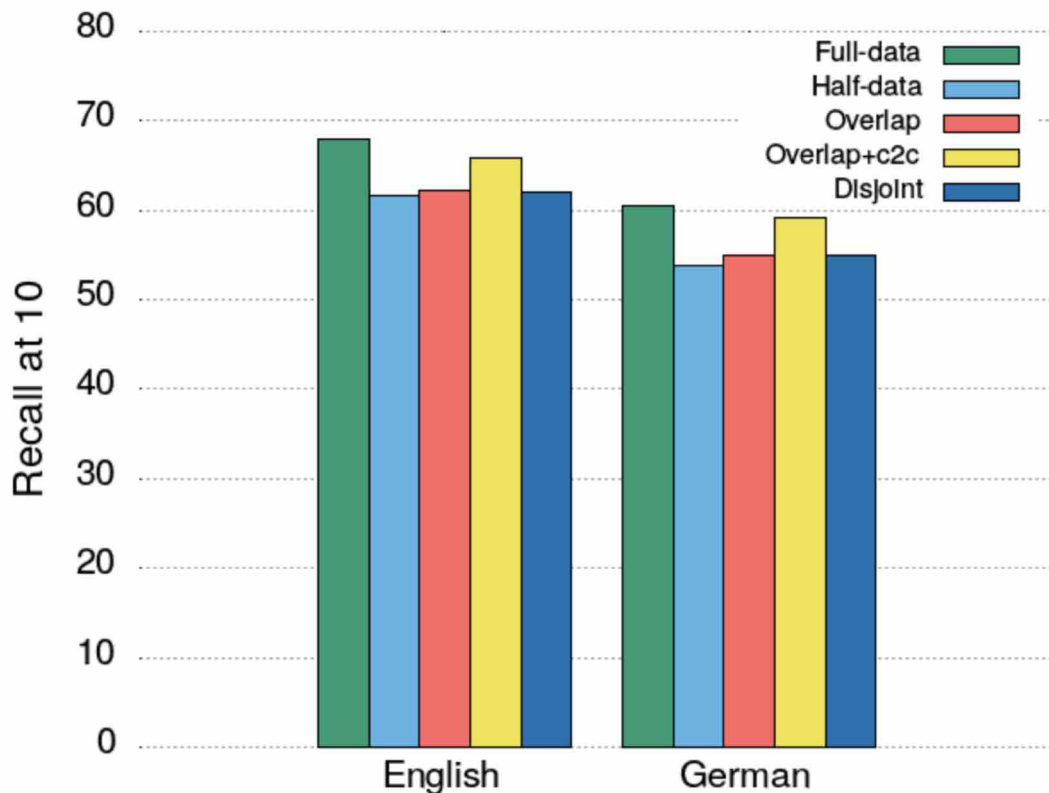
Multilingual data improves image retrieval



High-to-low resource transfer with multilingual data



Controlling for data exposure



Conclusions

- Multilingual data improves the ranking model
- Improvements also hold for “low-resource” settings
- Mixed results when controlling for data exposure

Summary

- Two ways of looking at multilingual and multimodal data
 - Retrieval task: *multilinguality* is useful
 - Translation task: *multimodality* is useful
- Both models benefited from learning to solve multiple tasks

Why does MTL help for translation?

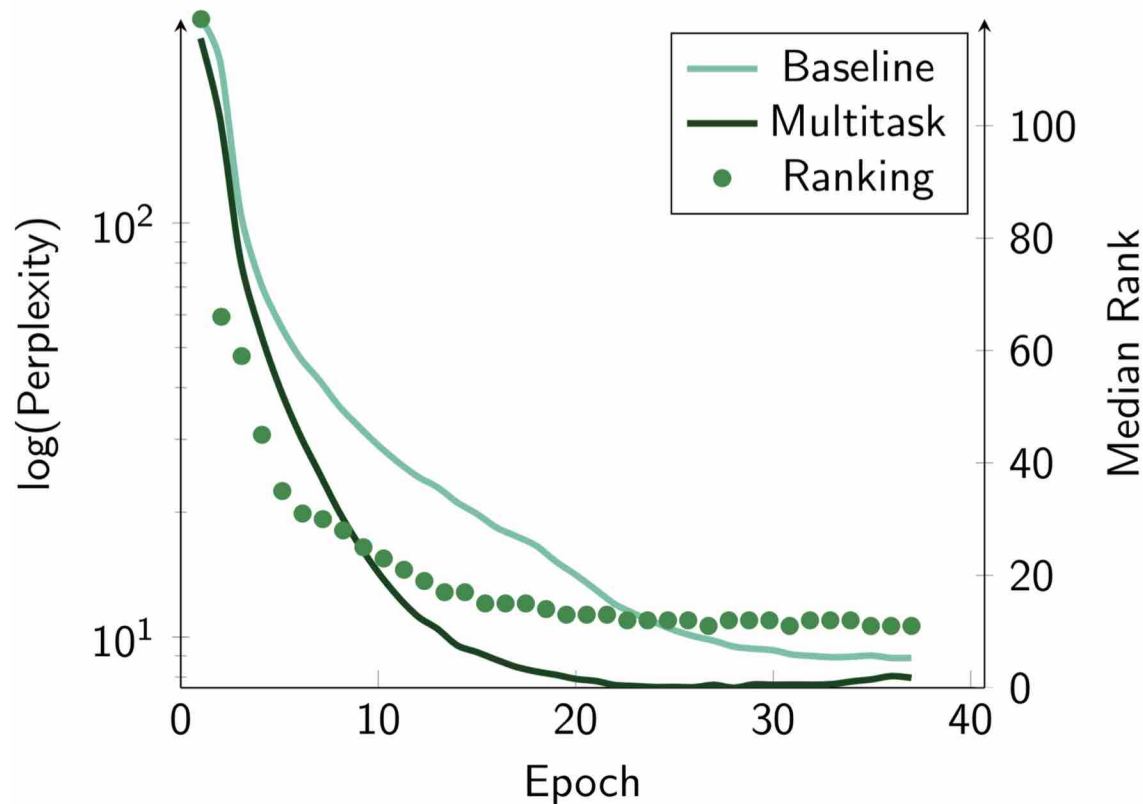
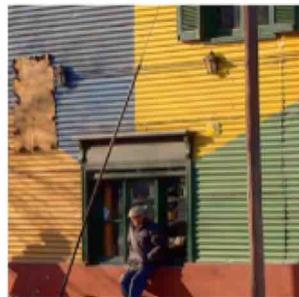


Image Prediction Visualisation

“there is a cafe on the street corner with an oval painting on the side of the building.”



Improved prepositional phrase translation



Two children on their stomachs lay on the ground under a pipe

Zwei Kinder **auf ihren Gesichtern** liegen unter dem Boden auf dem Boden

Zwei Kinder liegen Bäuchlings auf dem Boden unter einer **Schaukel**

Worse preposition selection



A bird flies across the water

Ein Vogel fliegt über das Wasser

Ein Vogel fliegt **durch** das Wasser

Data: translation won't always work



“draaiorgel”

- A yellow truck is standing on a busy street in front of the Swarovski store.
- A strange looking wood trailer is parked in a street in front of stores.
- An unusual looking vehicle parked in front of some stores.