Multimodal Machine Translation

Lucia Specia
University of Sheffield, **soon Imperial College London (too)**

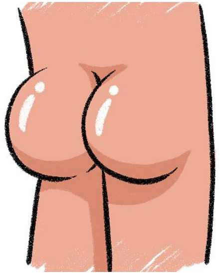
l.specia@sheffield.ac.uk

Overview

1. Motivation and existing approaches
2. Results on WMT16-18 shared tasks
3. On-going work on region-specific multimodal MT

Motivation

Motivation



Example by Desmond Elliott

Motivation

Humans interact with the world in **multimodal** ways.
Language understanding & generation is not an exception

Motivation

- **Multimodality** in computational models
  - Multimodal machine learning
  - Richer context modelling
  - Language grounding
- True for a wide range of NL **tasks**
- In this talk:
  - **Machine translation**
  - Additional modality: visual (**images**)

Motivation in MT: Morphology

- A baseball player in a black shirt just tagged a player in a white shirt.
- Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.
- Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.

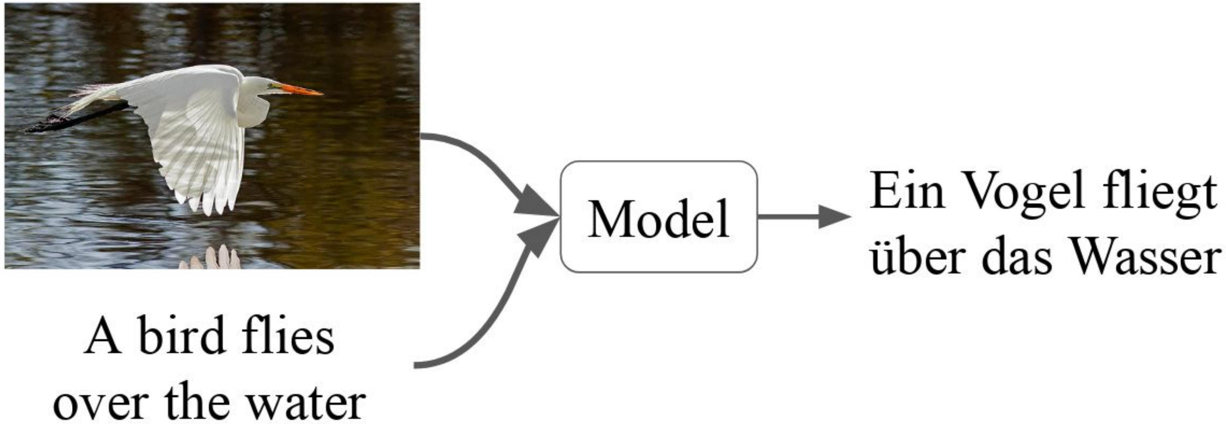
Motivation in MT: Semantics

- A woman sitting on a very large stone smiling at the camera with trees in the background.
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Stein und lächelt in die Kamera.
  - Stein == stone
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Felsen und lächelt in die Kamera.
  - Felsen == rock

Multimodal (Neural) Machine Translation (MMT)

Most slides borrowed from Loïc Barrault and Ozan Caglayan
Le Mans University

Task


Model
Ein Vogel fliegt über das Wasser
A bird flies over the water

Multi30K dataset

- Derived from Flickr30K
- Image captions, few Flickr groups
  - 30K sentences for training
  - 4 test sets (4.5K sentences)
- Used in WMT MMT task (3 editions)

- **EN**: A ballet class of five girls jumping in sequence.
- **DE**: Eine Ballettklasse mit fünf Mädchen, die nacheinander springen.
- **FR**: Une classe de ballet, composée de cinq filles, sautent en cadence.
- **CS**: Baletní třída pěti dívek skákající v řadě.

Research questions

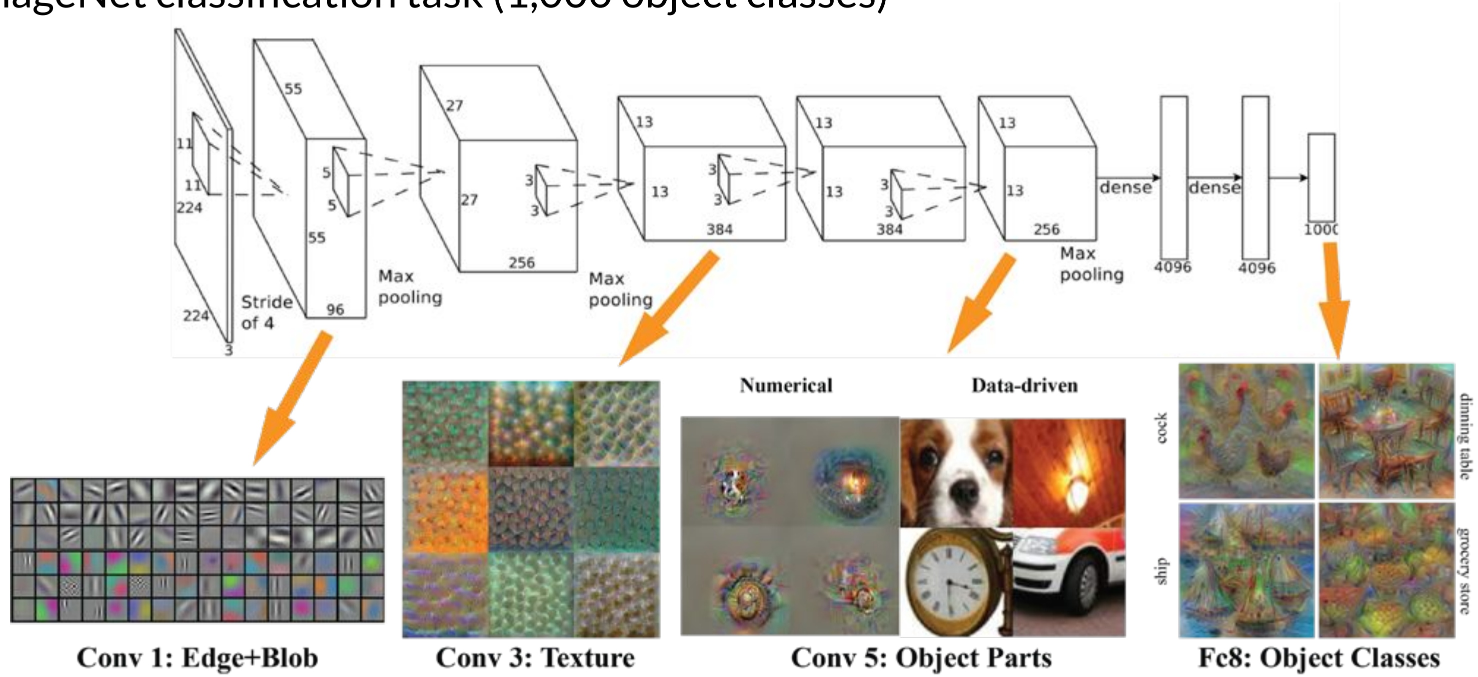
- How to best represent both modalities?
- How/where to integrate them in a model? Which architecture to use?
- Can we really ground language in the visual modality?
- Can we improve the MT system performance with images?

Representing textual input

- As in standard NMT
- **RNN**
  - Bidirectional RNN
  - Can use several layers: more abstract representation?
  - Last state: fixed-size vector representation
  - All states: matrix representation
- Convolutional networks, etc.

Representing images: CNN image networks

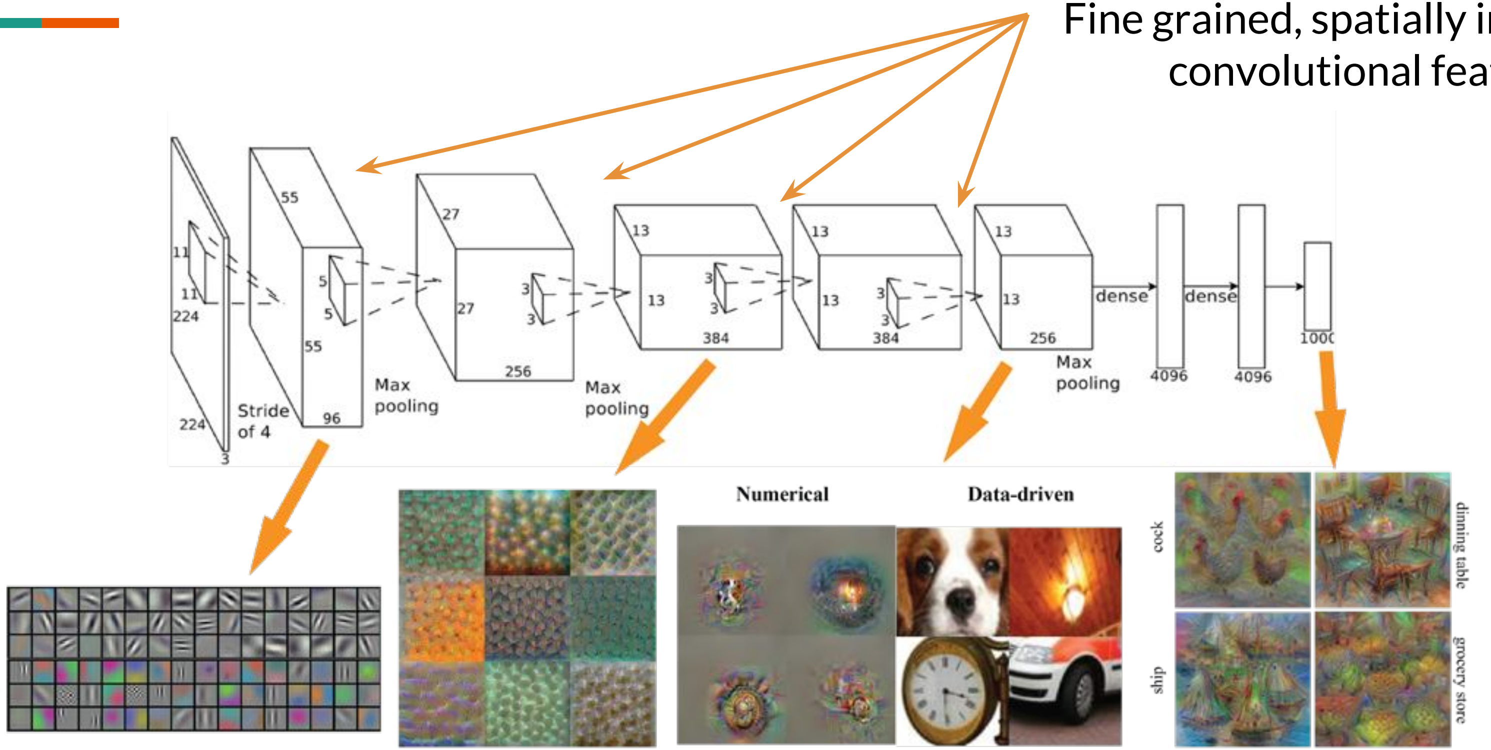
ImageNet classification task (1,000 object classes)



Visualization of AlexNet:: http://vision03.csail.mit.edu/cnn_art/index.html

Representing images: CNN image networks

Fine grained, spatially informative convolutional features



Conv 1: Edge+Blob

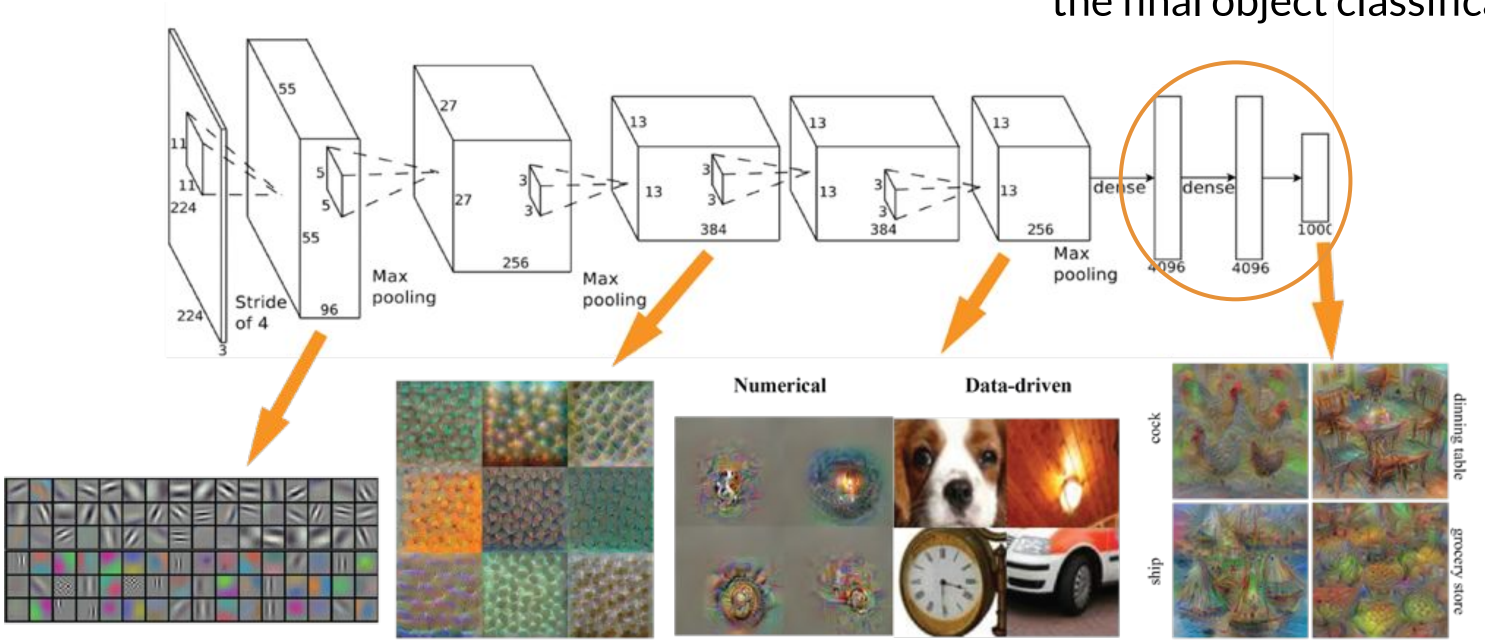
Conv 3: Texture

Conv 5: Object Parts

Fc8: Object Classes

Representing images: CNN image networks

Global features guided towards the final object classification task



Conv 1: Edge+Blob

Conv 3: Texture

Conv 5: Object Parts

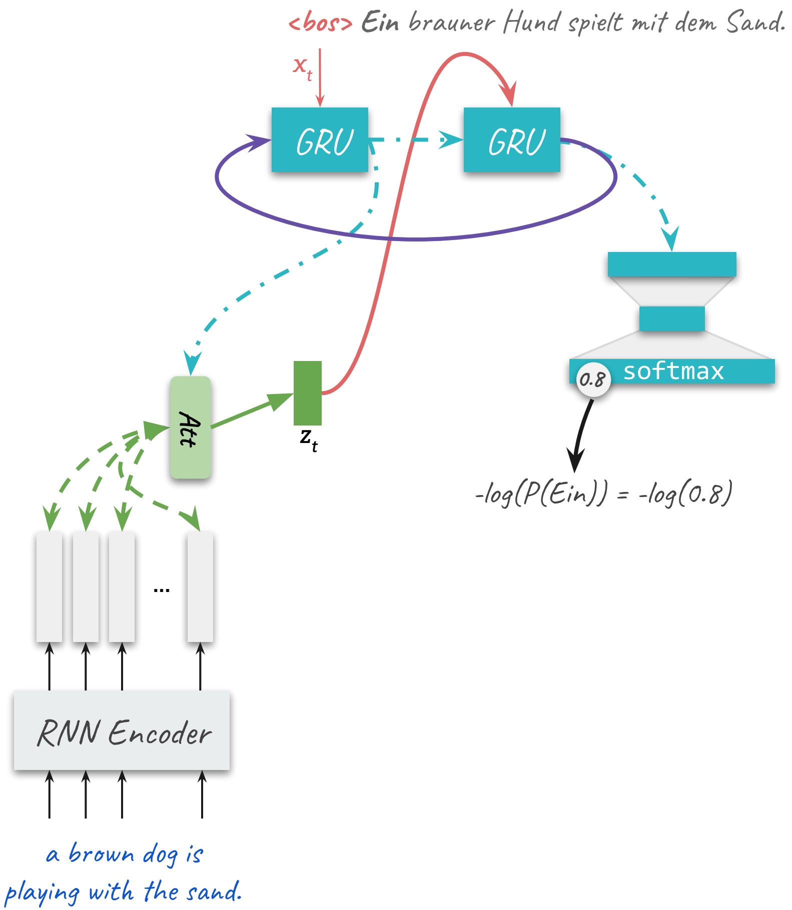
Fc8: Object Classes

Representing images: CNN image networks

- Any network - this is a pre-processing step (feature extraction)
- Common networks:
  - VGG (19 layers)
  - ResNet-101
  - ResNet-152
  - ResNeXt-101 (3D CNN)
- Networks can be pre-trained for different tasks
  - Object classification (1,000 objects)
  - Action recognition (400 actions)
  - Place recognition (365 places)
- Different layers of the CNN can be used as features

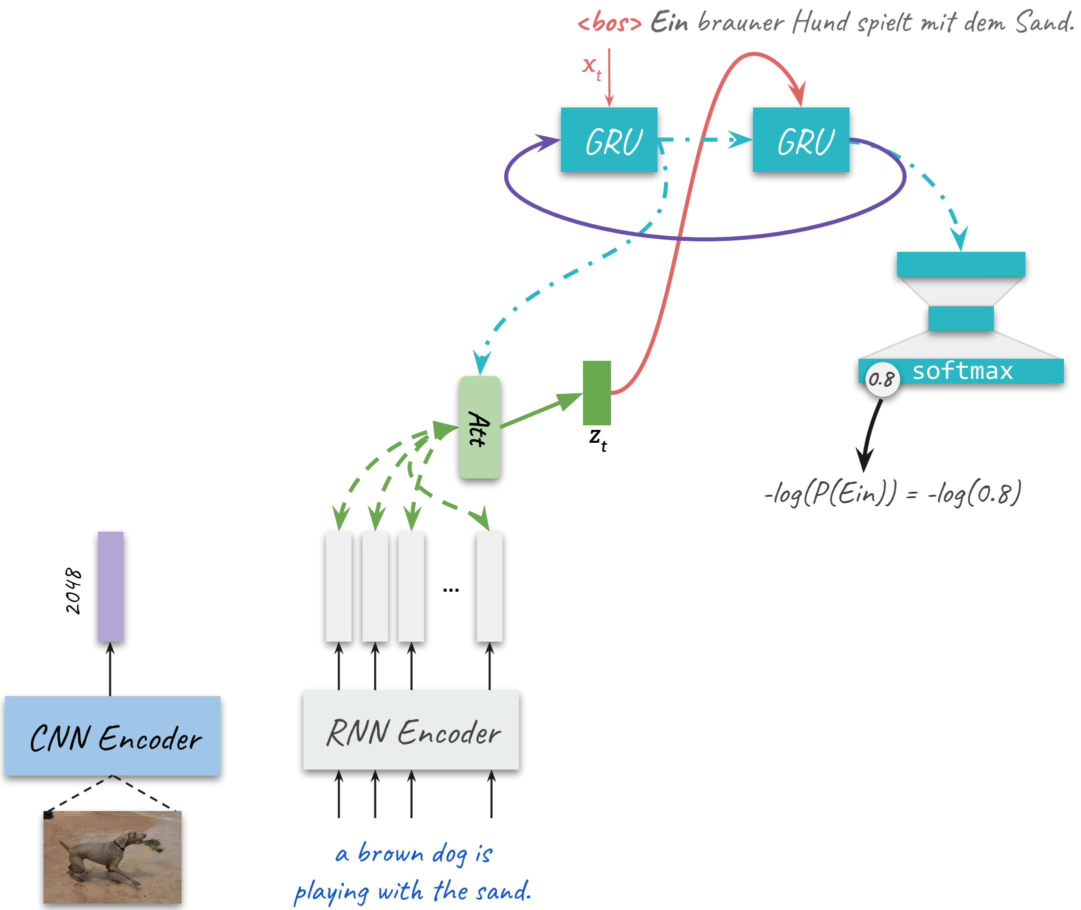
Integration of visual information

Simple Multimodal NMT



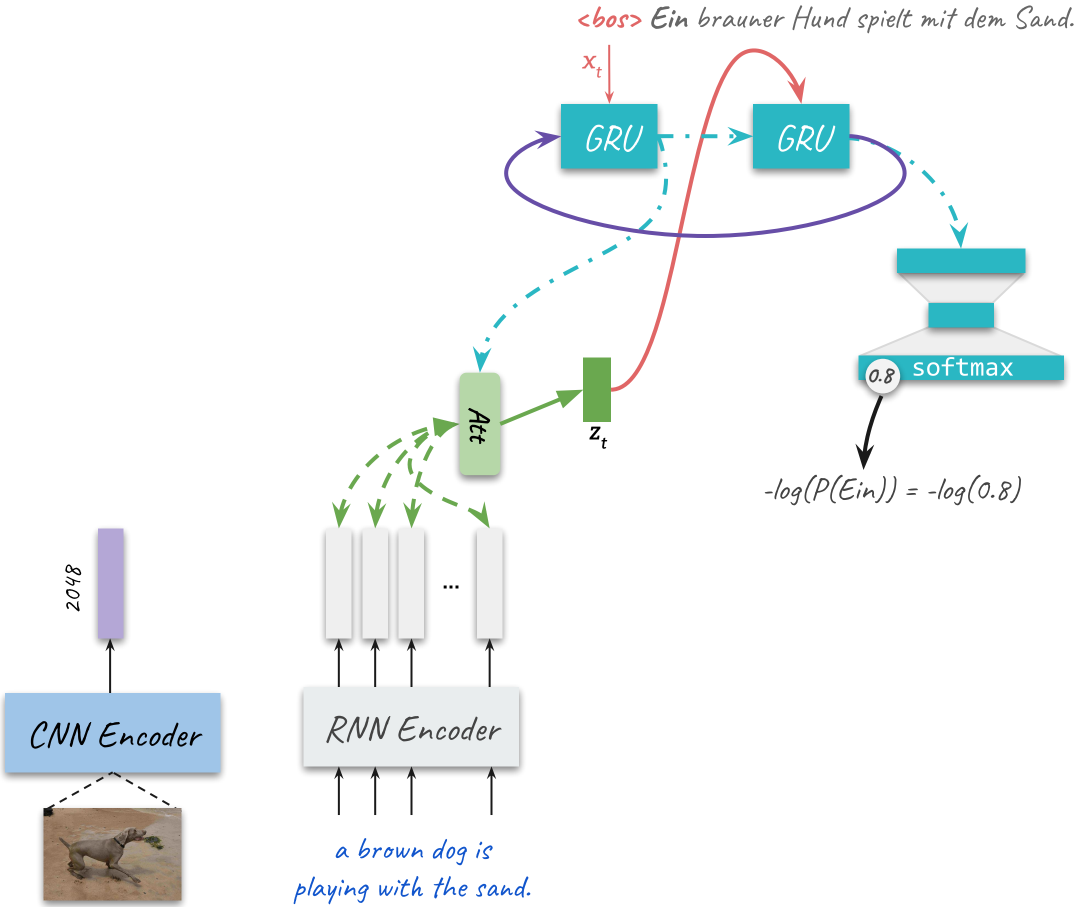
Simple Multimodal NMT

- Extract a single global feature vector from some layer of CNN.



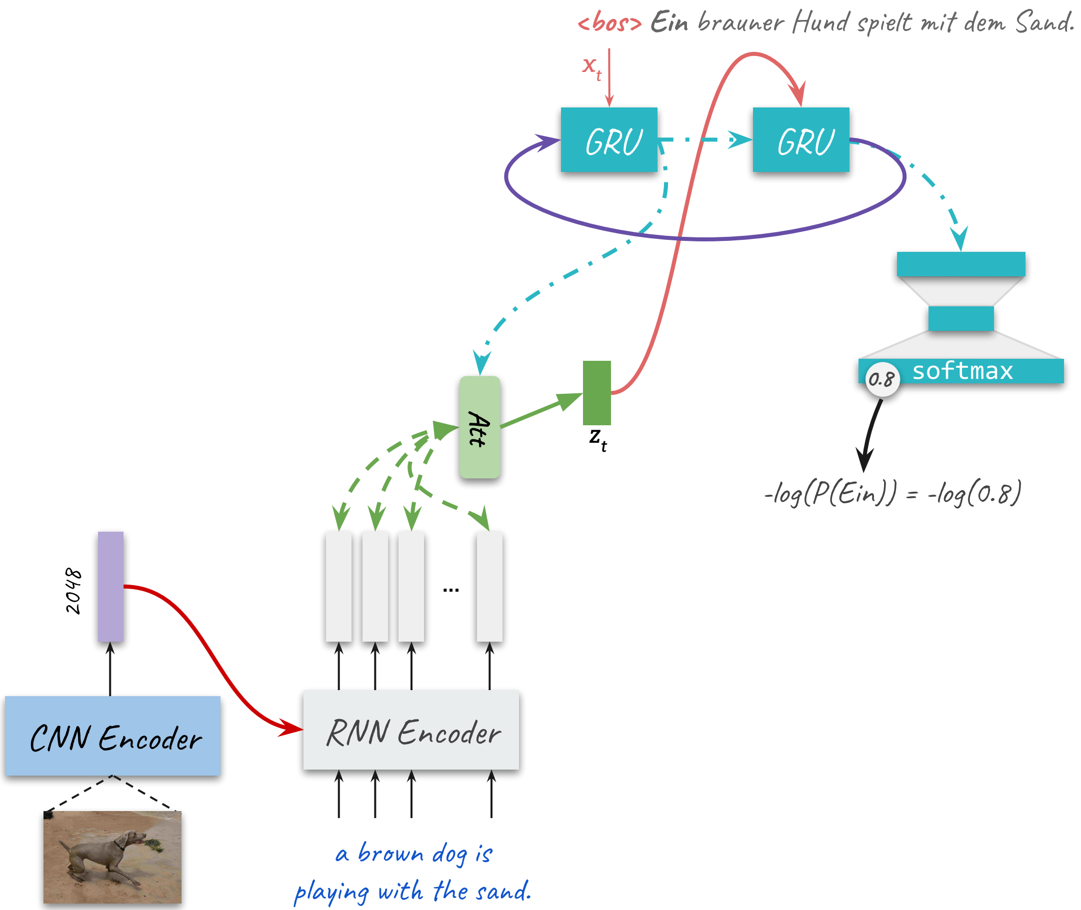
Simple Multimodal NMT

- Extract a single global feature vector from some layer of CNN.
- This vector will be used throughout the network to contextualize language representations.



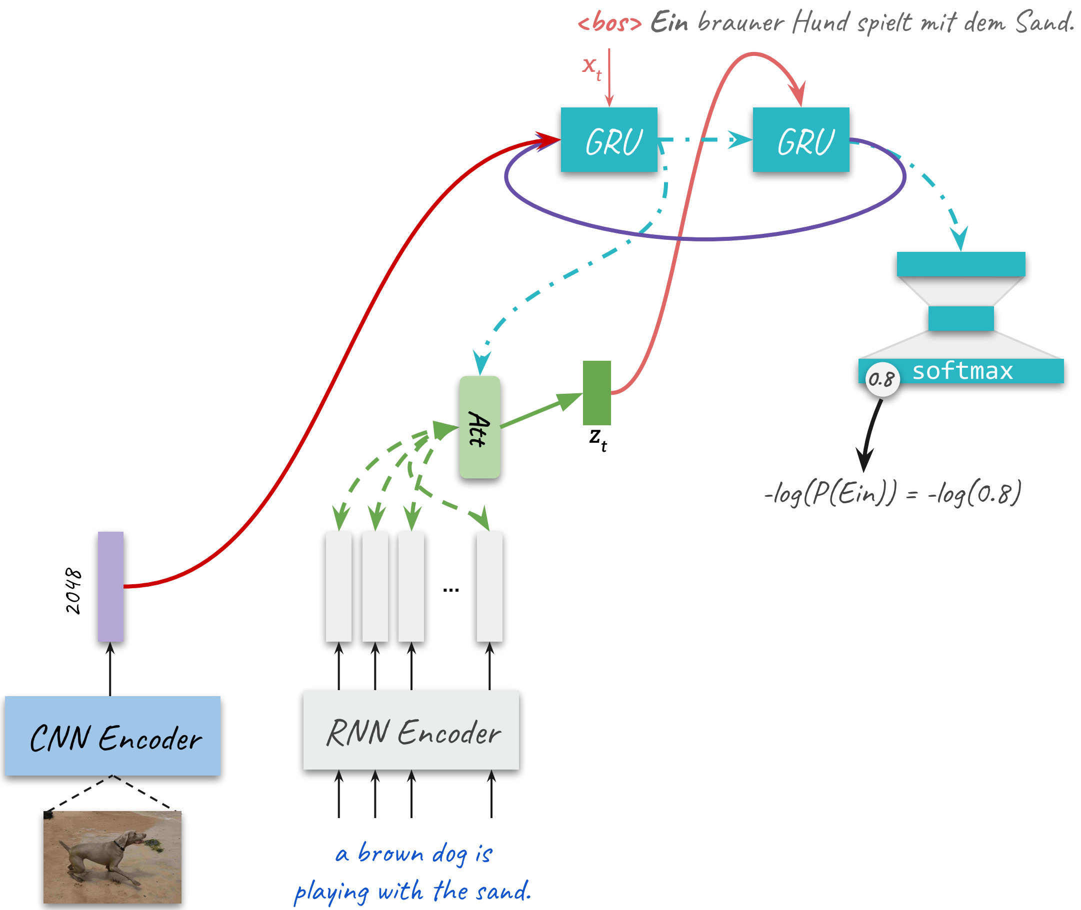
Simple Multimodal NMT

1. Initialize the source sentence encoder.



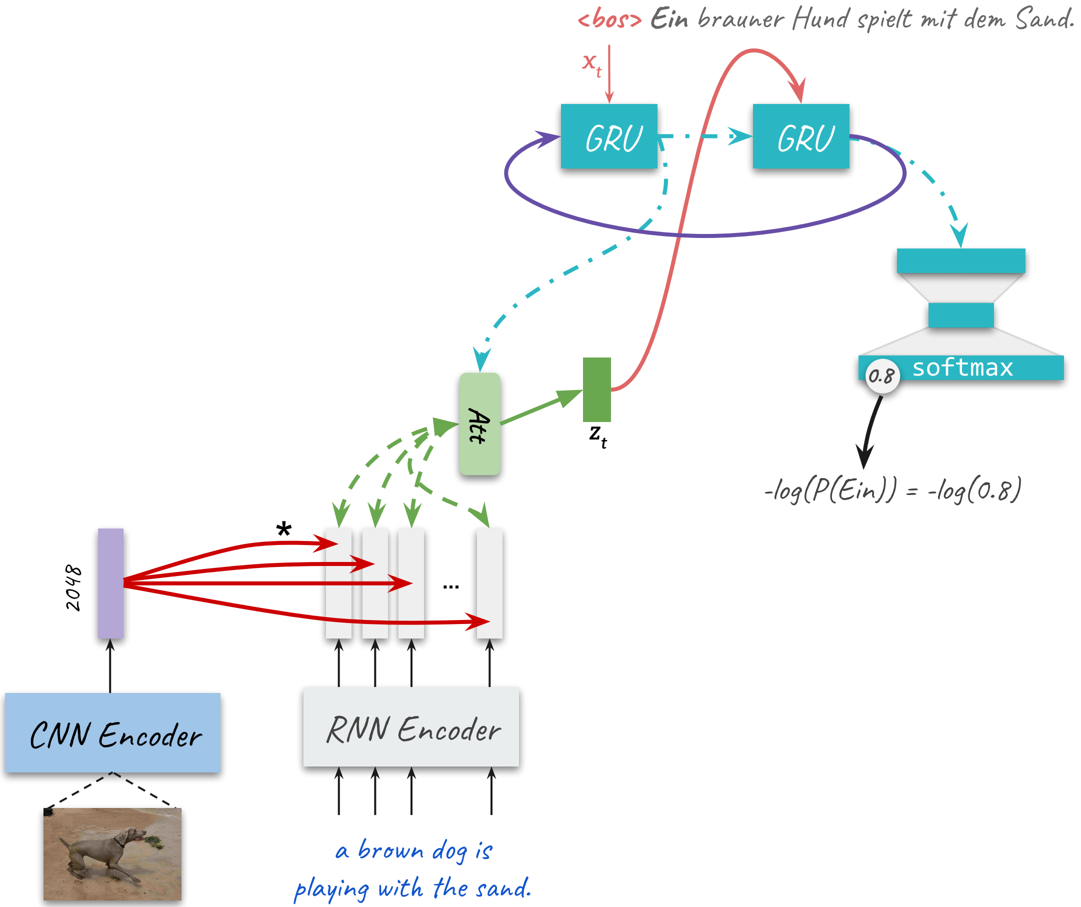
Simple Multimodal NMT

1. Initialize the source sentence encoder
2. Initialize the decoder



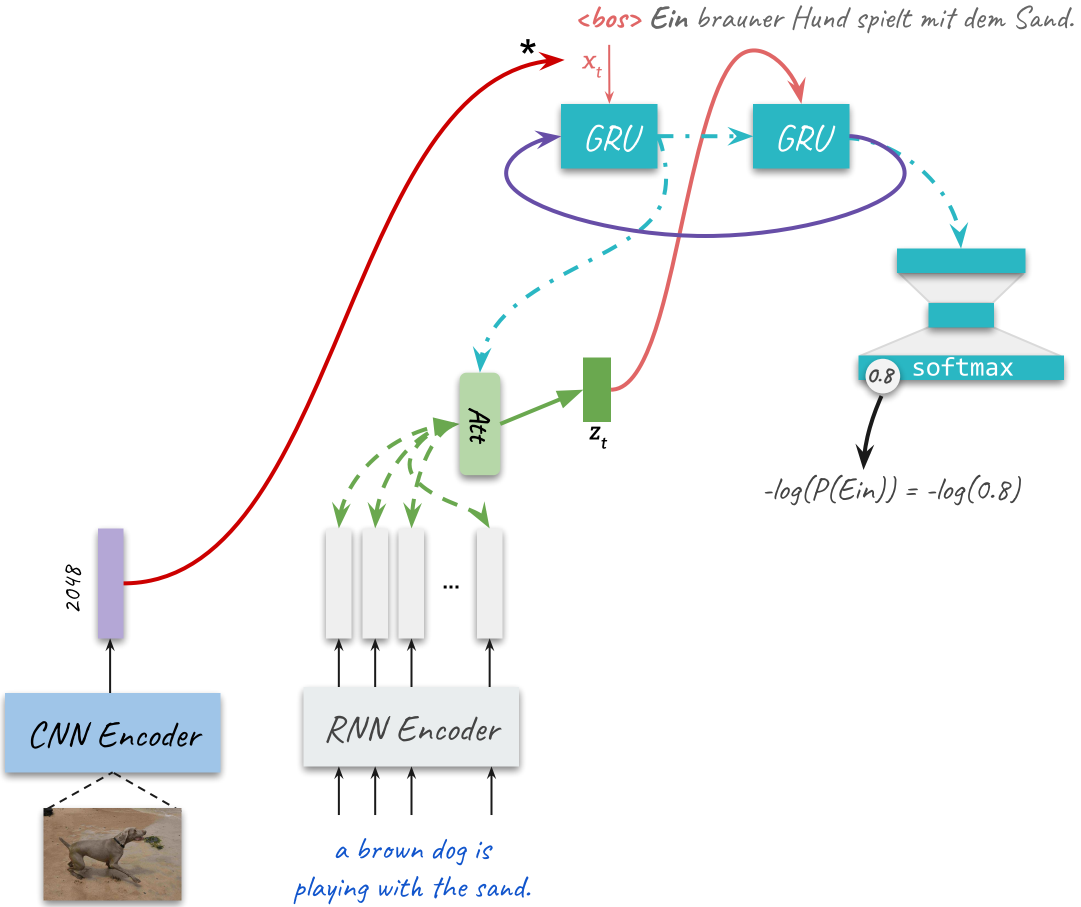
Simple Multimodal NMT

1. Initialize the source sentence encoder
2. Initialize the decoder
3. Element-wise multiplicative interaction with source annotations.



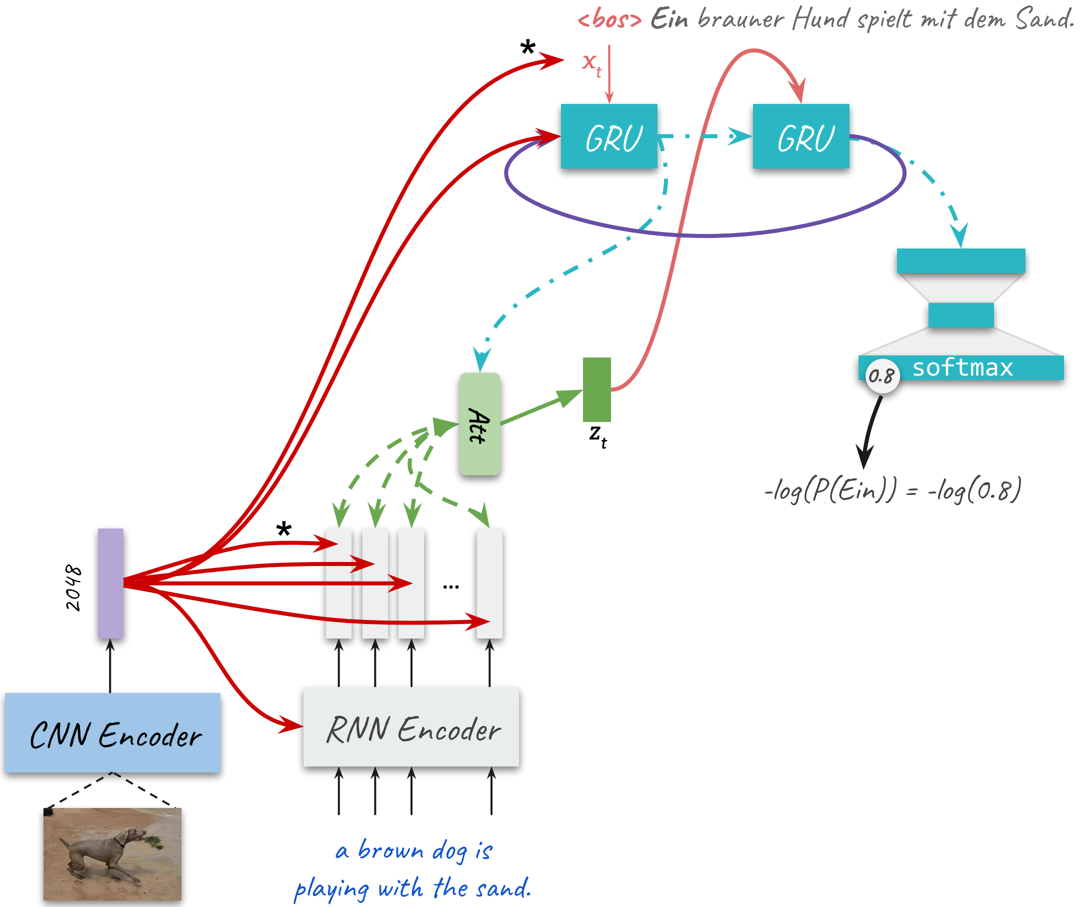
Simple Multimodal NMT

1. Initialize the source sentence encoder
2. Initialize the decoder
3. Element-wise multiplicative interaction with source annotations.
4. Element-wise multiplicative interaction with target embeddings.



Simple Multimodal NMT

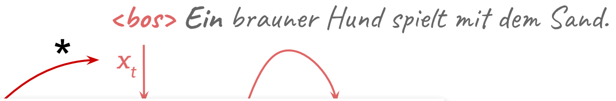
- Initialize the source sentence encoder
- Initialize the decoder
- Element-wise multiplicative interaction with source annotations.
- Element-wise multiplicative interaction with target embeddings.

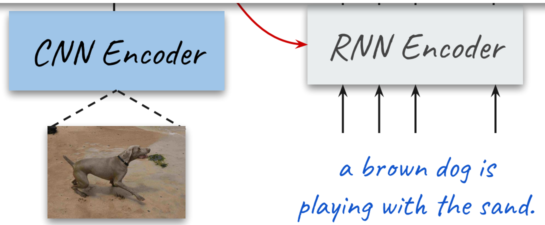


Simple Multimodal NMT

- Initia… enco…
- Initia…
- Elem… inter… anno…
- Elem… inter… emb…

- 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.
- Calixto, I., Elliott, D., and Frank, S. (2016). DCU-UVA multimodal mt system report.
- Madhyastha, P. S., Wang, J., and Specia, L. (2017). Sheffield multimt: Using object posterior predictions for multimodal machine translation.
- Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation.





Summary

- Encode image as a single vector
- Explore different strategies to mix image and text features
  - ➢ Initialize RNN, concatenate, prepend, multiply (element-wise)
- What about grounding?
  - Hard to visualize...

Summary

- Ray Mooney (U. Texas)

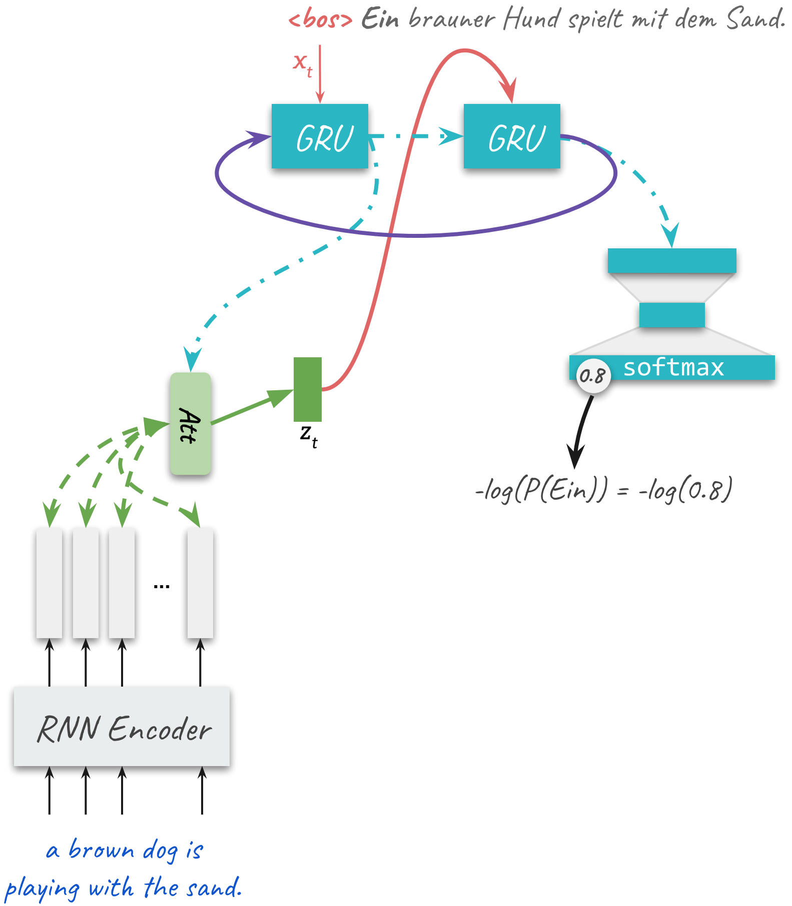
You can't cram the meaning of a whole *$#*! sentence into a single *$#*! vector!

- Can we summarise the whole image using a single vector?
  - Probably not for MMT...

- From **coarse** to **fine** visual information

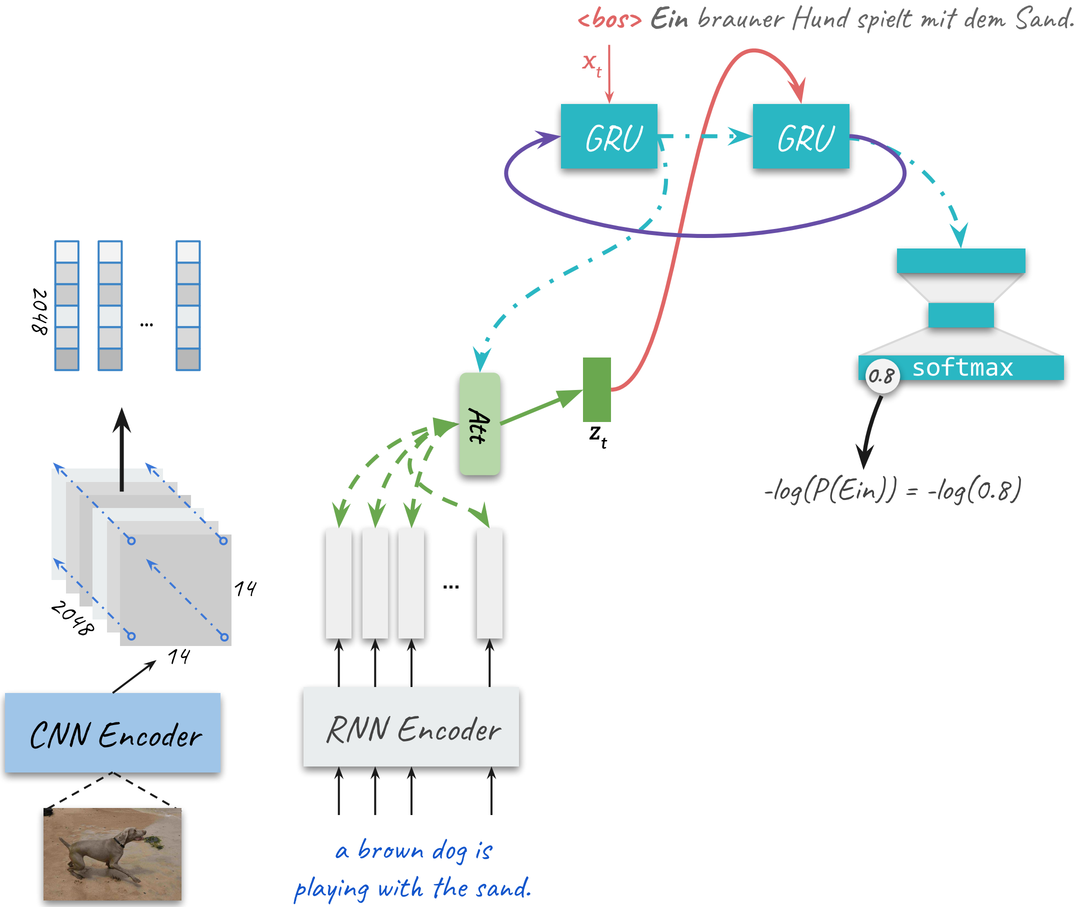
- **Idea**:
  - Use only **relevant parts** of the image, **when needed**
  - E.g. objects related to the input words
  - (Karpathy and Fei-Fei, 2015) for IC

Attentive Multimodal NMT



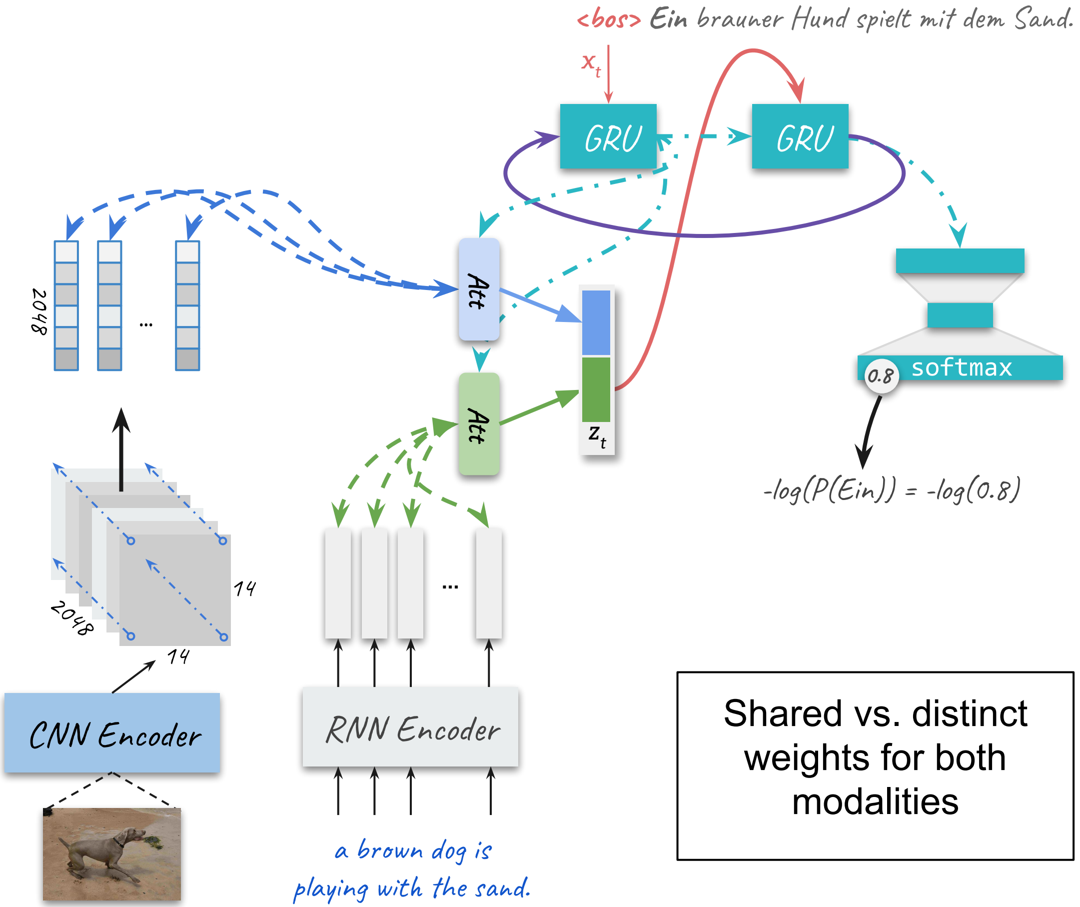
Attentive Multimodal NMT

- Use a CNN to extract **convolutional features** from the image.
  - Preserve spatial correspondence with the input image.



Attentive Multimodal NMT

- Use a CNN to extract **convolutional features** from the image
  - Preserve spatial correspondence with the input image
- A new attention block for the visual annotations
- z_t becomes the fusion of both contexts (e.g. concat).

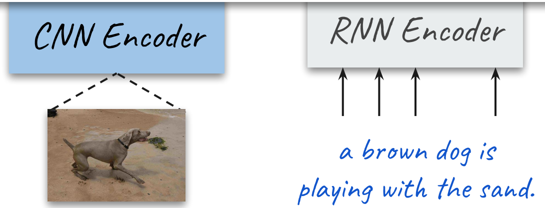


Shared vs. distinct weights for both modalities

Attentive Multimodal NMT

- Use **conv**
  the i
  -
- A ne
  visu
- z_t be
  both

- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention
- Caglayan, O., Barrault, L., and Bougares, F. (2016b). Multimodal attention for neural machine translation
- Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning.
- Calixto, I., Liu, Q., & Campbell, N. (2017). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation.

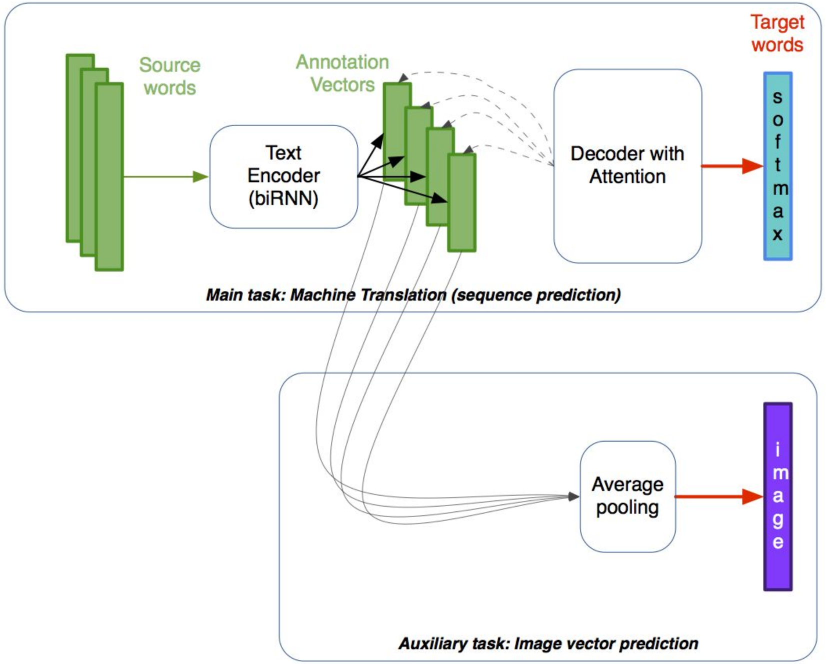


Shared vs. distinct weights for both modalities

Integration: multitask learning -- Imagination

- Predict image vector from source sentence during training only

➢ Gradient flow from image vector impact the source text encoder and embeddings
  - Elliott and Kádár (2017)



Some Results

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	213M	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Average of 3 runs
vs
Ensemble

Caglayan et al., 2017

Some Results

Attentive MNMT with **shared** / **separate** visual attention

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	213M	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Caglayan et al., 2017

Some Results

Simple MNMT variants

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	213M	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Caglayan et al., 2017

Some Results

Multiplicative interaction with target embeddings

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	213M	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Caglayan et al., 2017

Some Results

Huge models overfit and are slow.

Small dimensionalities are better for small datasets (no need for a strong regularization)

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	$\boxed{62.0M}$	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	$\boxed{213M}$	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Caglayan et al., 2017

Some Results

Models are early-stopped w.r.t METEOR

Best METEOR does not guarantee best BLEU

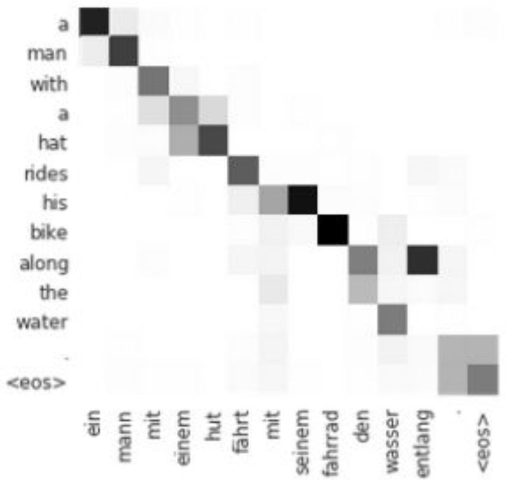
En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5
Huang et al. (2016)	-	36.5	54.1
Calixto et al. (2017a)	213M	36.5	55.0
Calixto et al. (2017b)	-	37.3	55.1
Elliott and Kádár (2017)	-	36.8	55.8
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**

Caglayan et al., 2017

What about grounding?

Attention mechanism

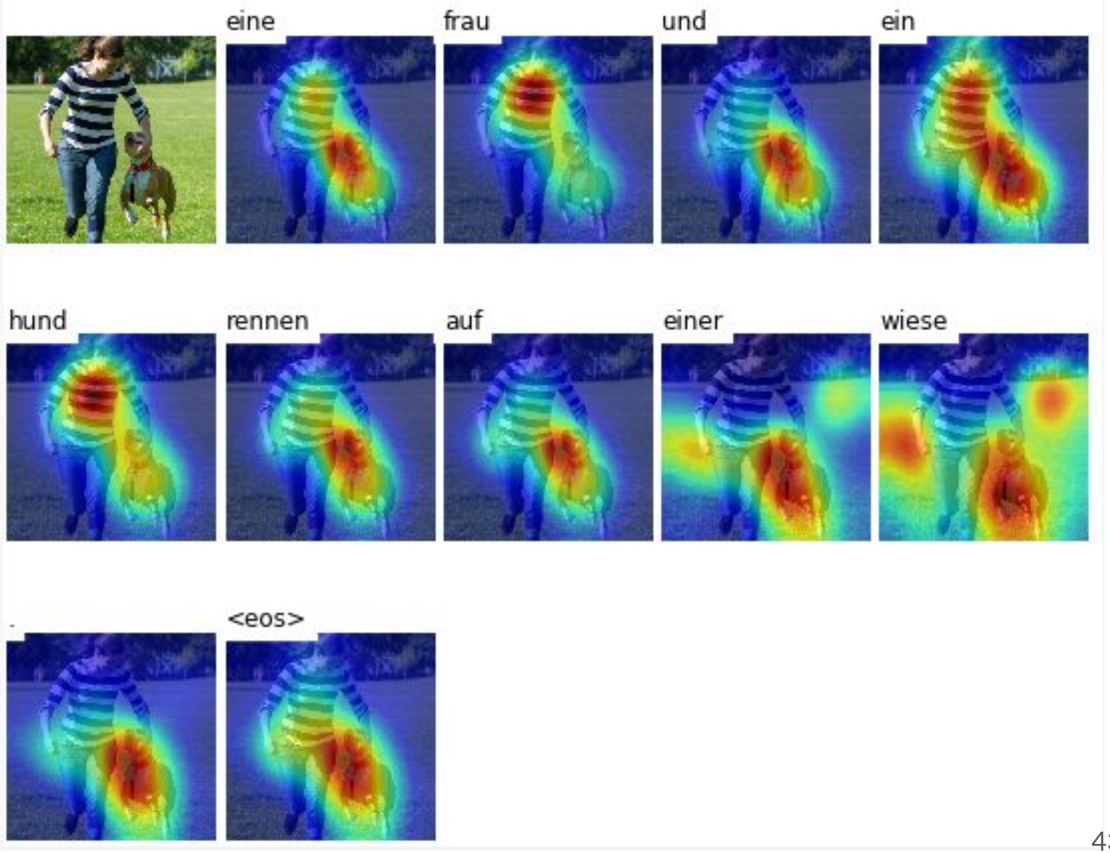
- Attention weights can be thought of as **link** between modalities
  - Alignment (?)



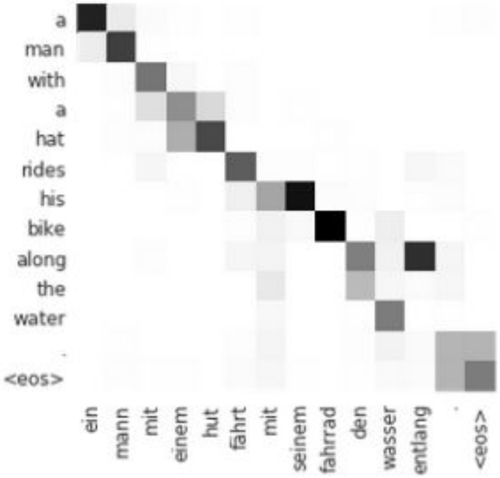
Attentive Multimodal NMT

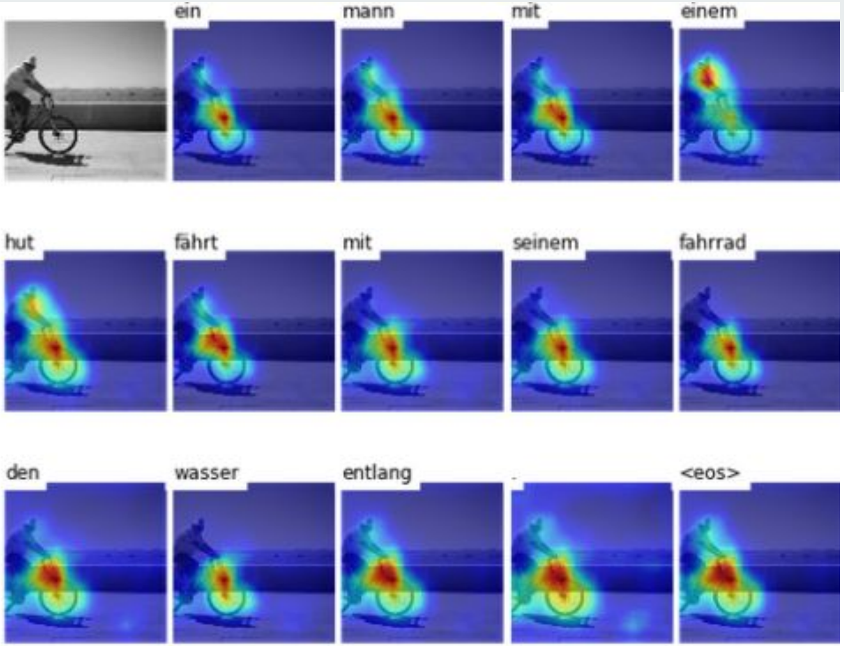
- Attention over spatial regions while translating from English → German

A woman and a dog run on a meadow .



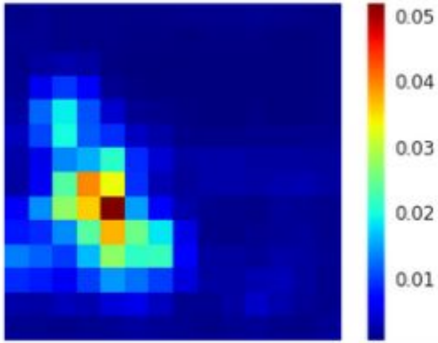
Textual Attention





Sequential spatial attention

Average spatial attention

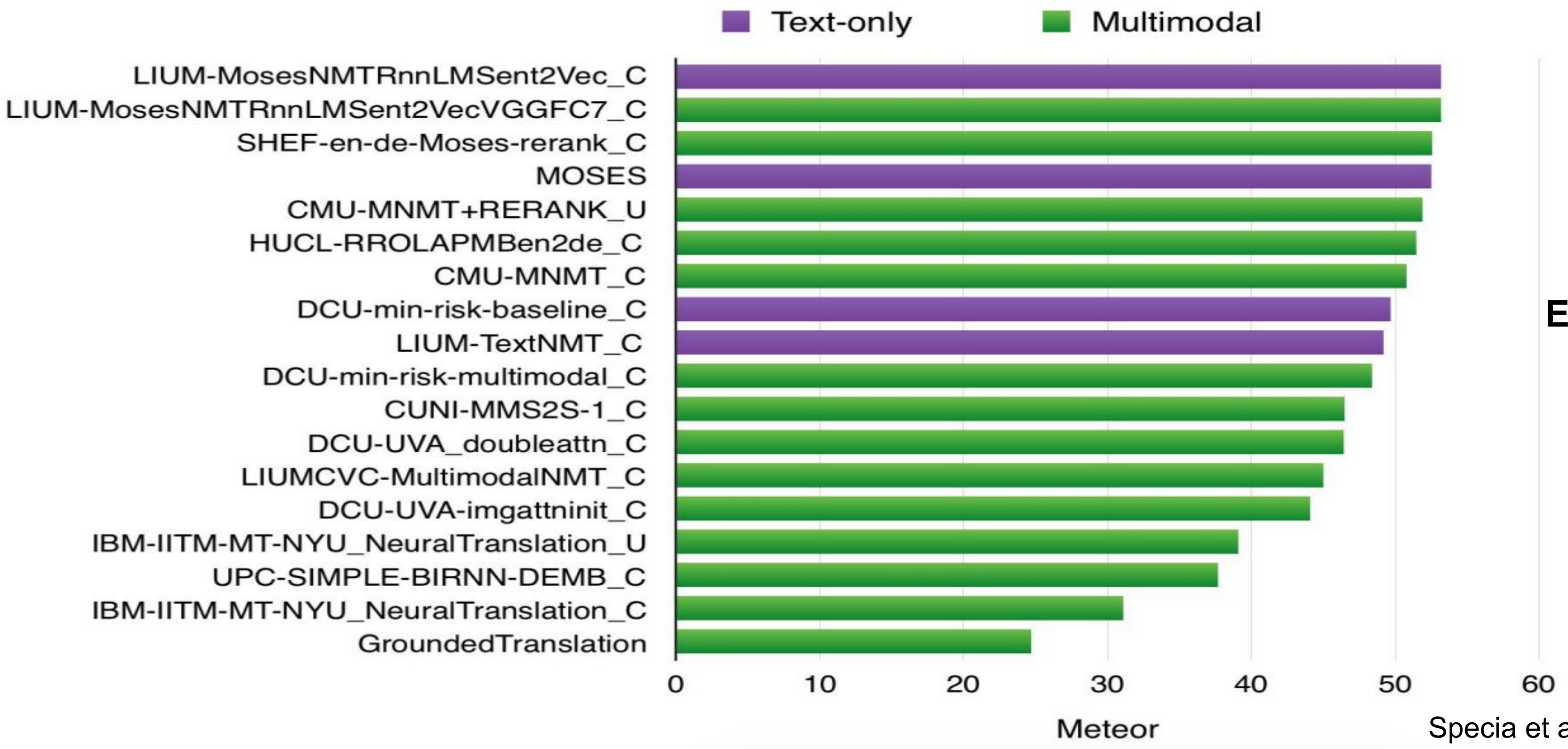


A man with a hat is riding his bike along the water .

Does MMT improve translation quality?

Blind evaluations

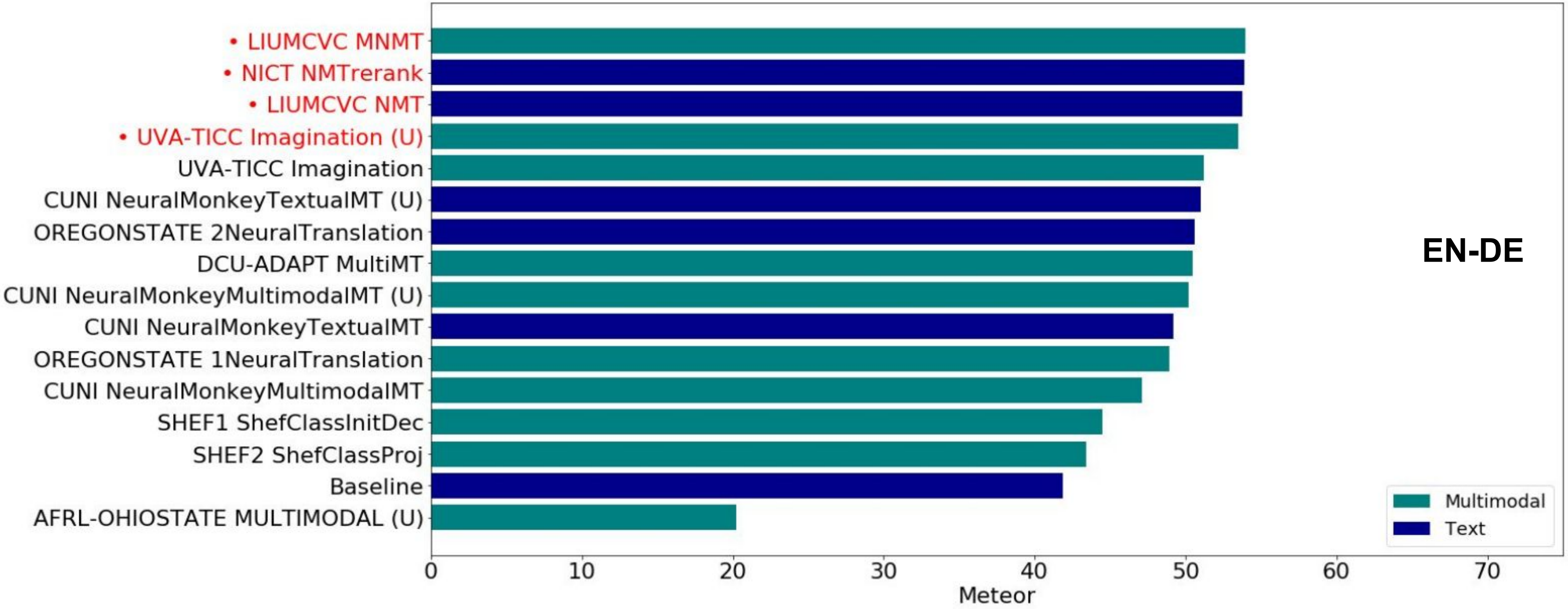
Results from WMT shared task - 2016



EN-DE

Specia et al., 2016

Results from WMT shared task - 2017



Elliott et al., 2017

Results from WMT shared task - 2017

#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

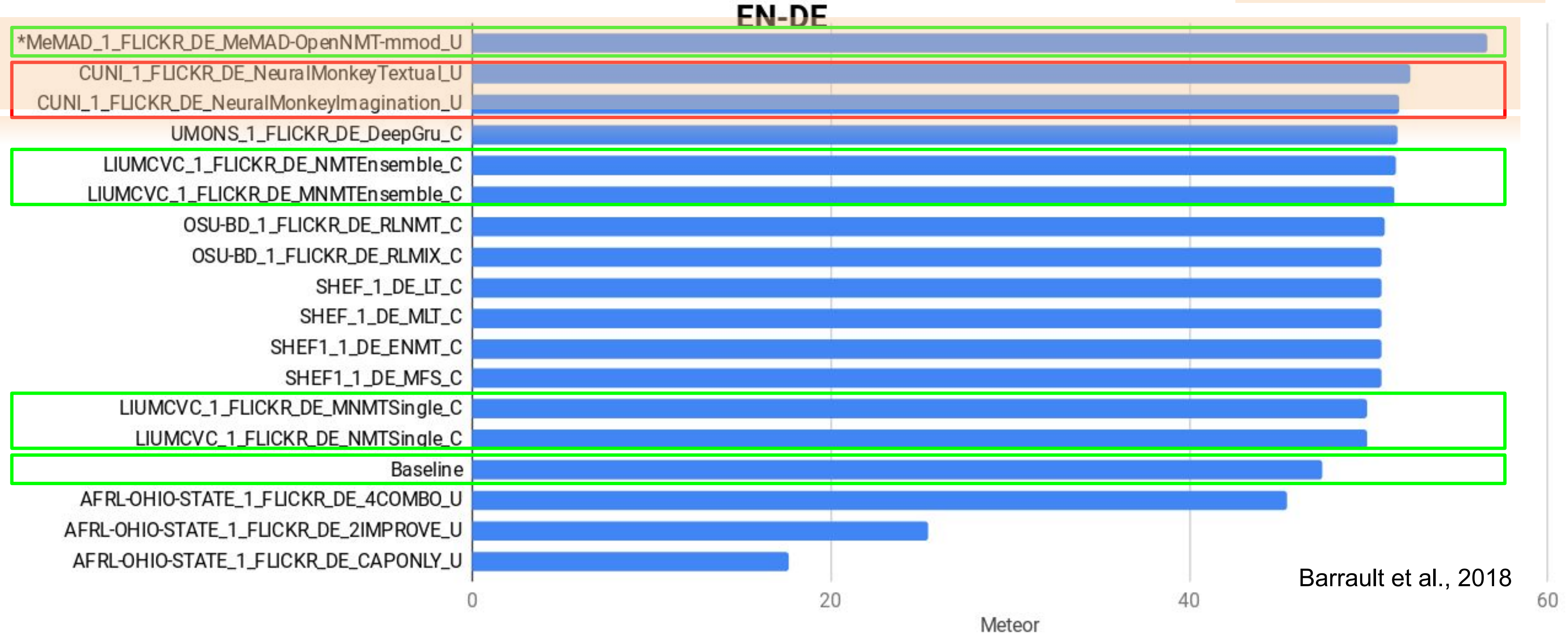
Human evaluation
EN-DE

Multimodal
Text

Elliott et al., 2017

Results from WMT shared task - 2018

Transformer architecture



Barrault et al., 2018

Results from WMT shared task - 2018

#	Ave %	Ave z	English→French System
1	90.3	0.487	gold_FR_1
2	86.8	0.349	MeMAD_MeMAD-OpenNMT-mmod_U
3	78.5	0.047	CUNI_NeuralMonkeyImagination_U
	77.3	-0.005	UMONS_DeepGru_C
	74.9	-0.05	LIUMCVC_NMTEnsemble_C
	74.9	-0.075	SHEF1_1_FR_MFS_C
	74.5	-0.088	SHEF_1_FR_MLT_C
	73.0	-0.11	LIUMCVC_MNMTEnsemble_C
	74.4	-0.12	OSU-BD_RLNMT_C
	66.0	-0.376	baseline_FR

Human evaluation EN-FR

Barrault et al., 2018

Results from WMT shared task - 2018

#	Ave %	Ave z	English→Czech System
1	93.2	0.866	gold_CS_1
2	70.2	0.097	CUNI_NeuralMonkeyImagination_U.txt
	62.4	-0.162	SHEF_1_CS_MLT_C
	60.6	-0.225	SHEF1_1_CS_MFS_C
	59.1	-0.248	OSU-BD_RLNMT_C
3	57.8	-0.337	baseline_CS

Human evaluation EN-CZ

Barrault et al., 2018

Conclusions

- Various ways of integrating textual and visual features
- Check WMT18 papers - out soon
- Results in terms of METEOR are only slightly impacted
- Manual evaluation shows clear trend
  - Multimodal systems are perceived as better by humans

- Dataset is not ideal...
  - Multi30k is simplistic and repetitive - predictable
  - Not all sentences need visual information to produce a good translation

Grounding over regions

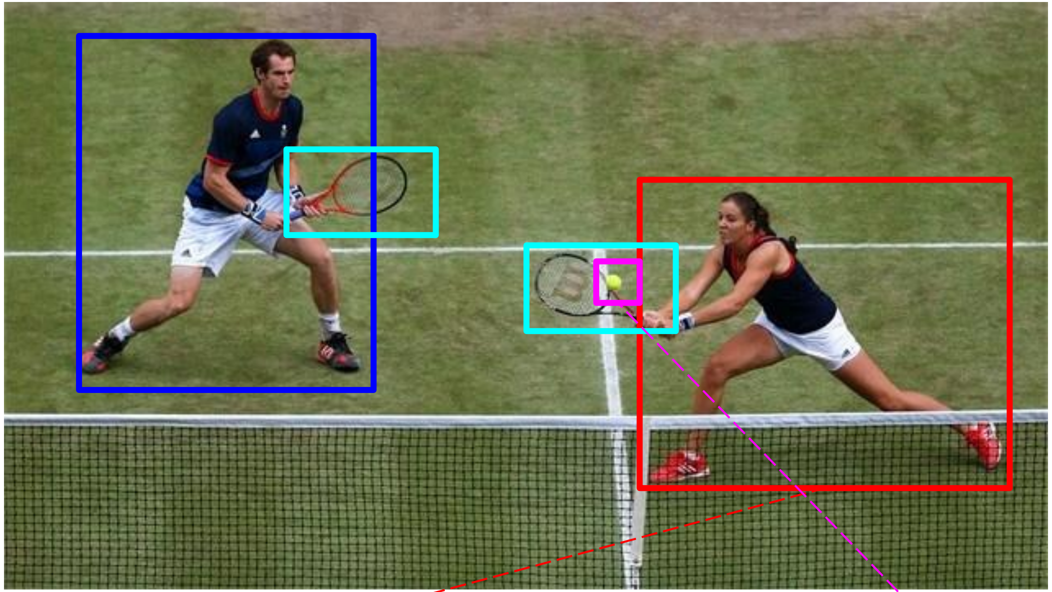
Joint work with Josiah Wang, Jasmine Lee, Alissa Ostapenko and Pranava Madhyastha

Image regions

The player on the right has just hit the ball

~~O jogador~~ à direita acaba de acertar a bola

Image regions



The player on the right has just hit the ball

🌳

A jogadora à direita acaba de acertar a bola

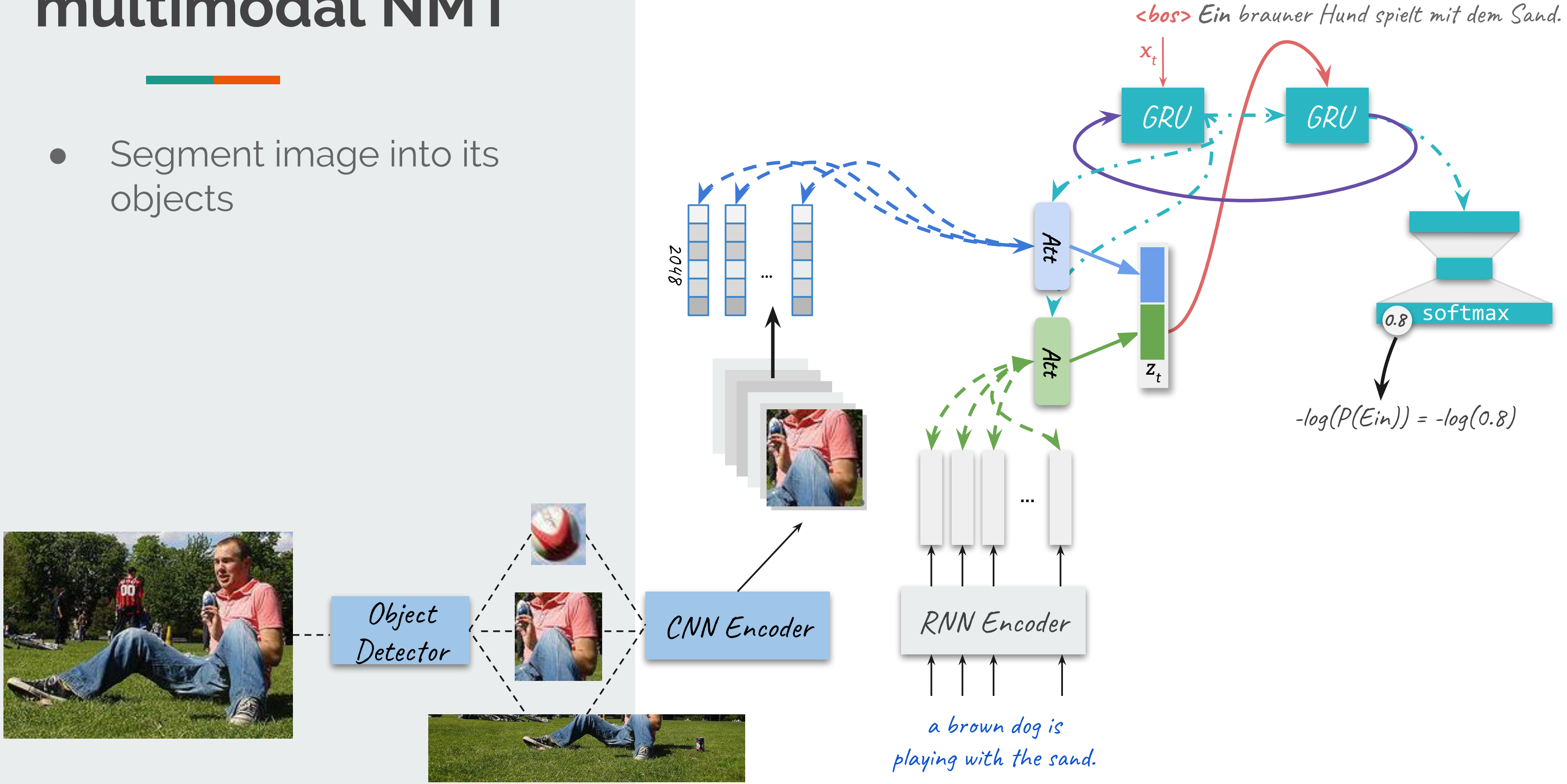
Image regions

- **Idea**: alignment between regions in image and words
- Beyond attention: ‘trusted’ alignments
- First detect objects, then guide model to translate certain words based on certain objects
- Two approaches:
    - Implicit alignment (different forms of attention - but over regions)
    - Explicit alignment (pre-grounding)

Implicit alignments

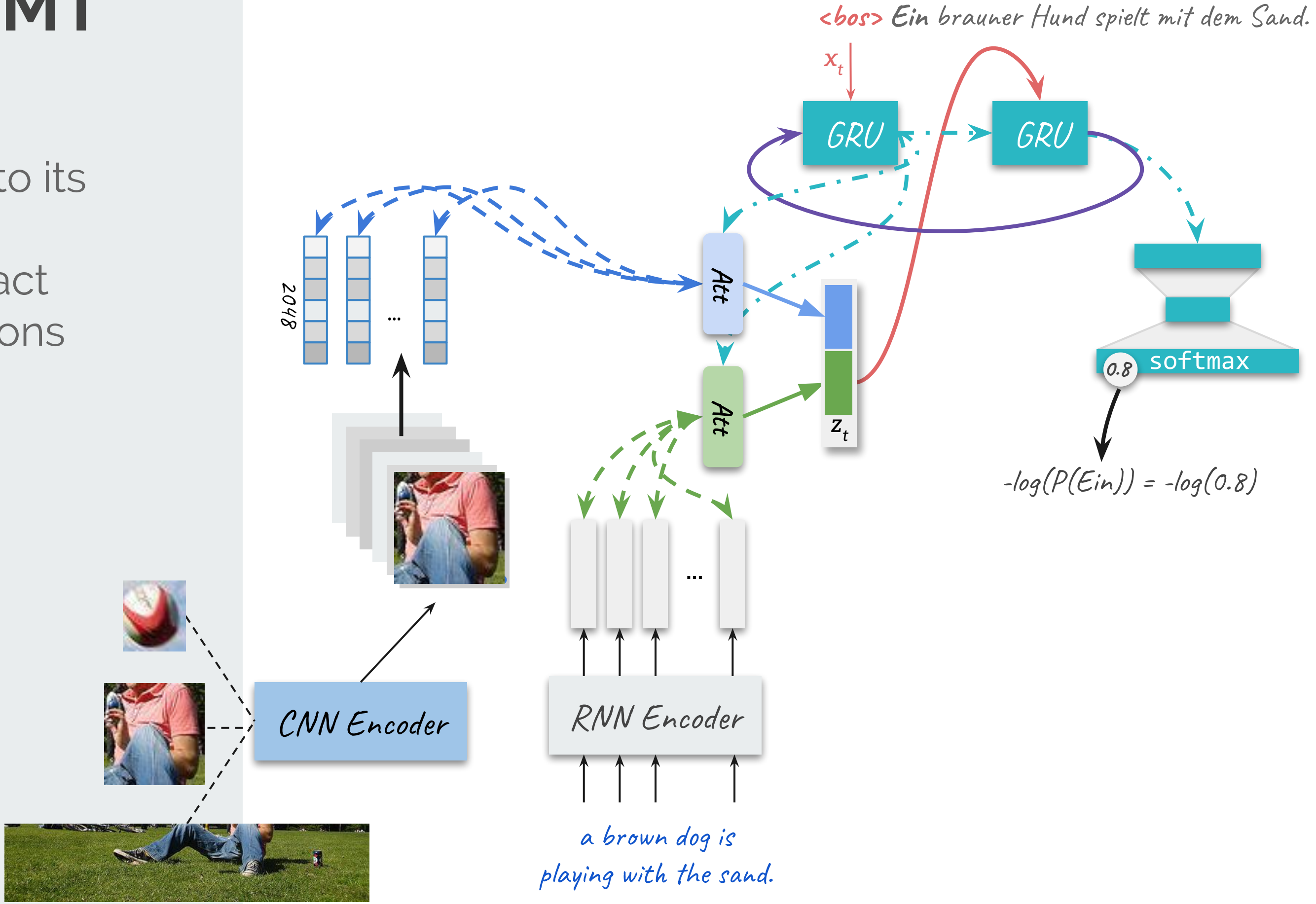
Region-attentive multimodal NMT

- Segment image into its objects



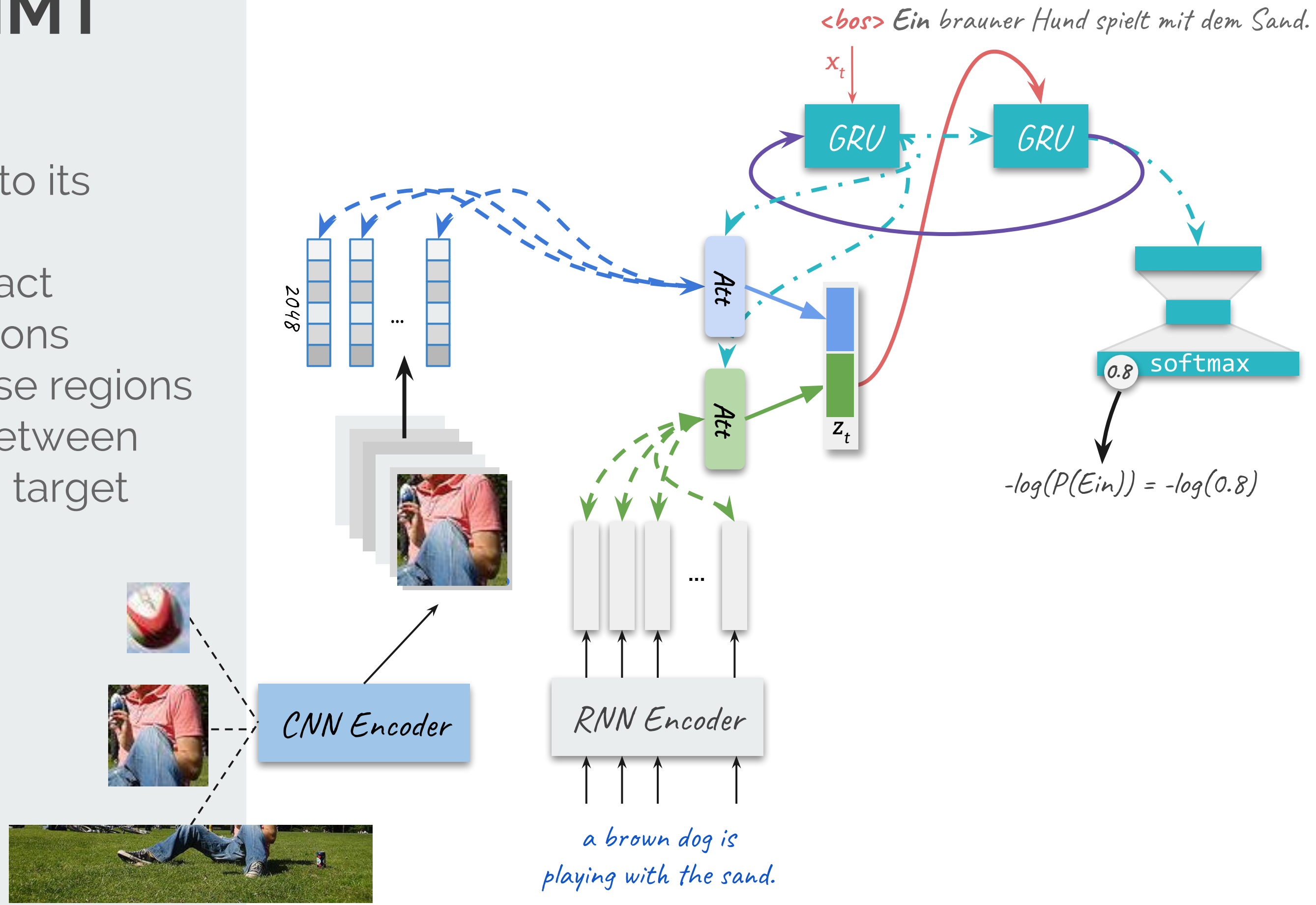
Region-attentive multimodal NMT

- Segment image into its objects
- Use a CNN to extract **features** from regions



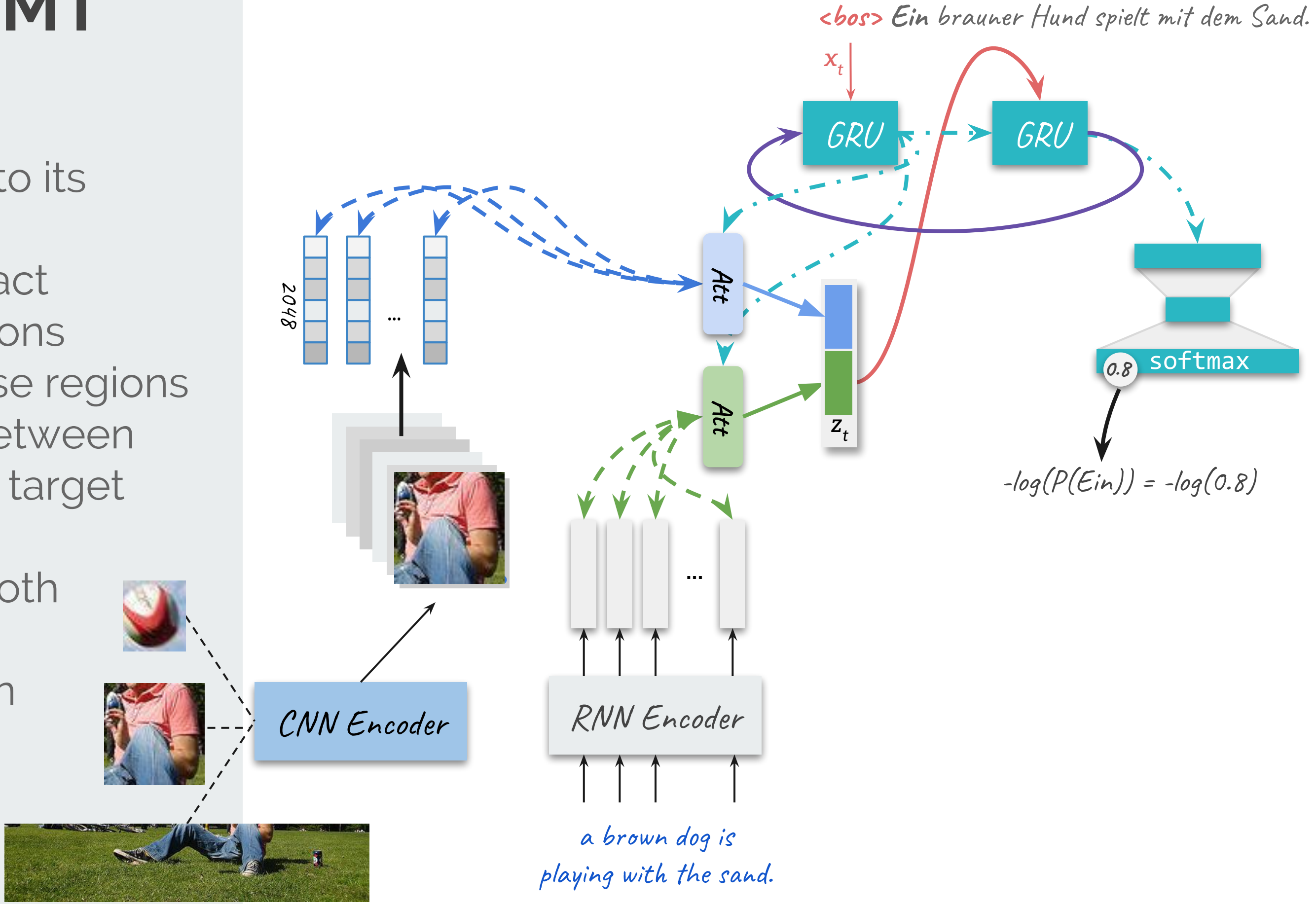
Region-attentive multimodal NMT

- Segment image into its objects
- Use a CNN to extract **features** from regions
- Attention over these regions
- **Idea**: **alignment** between regions & words in target language



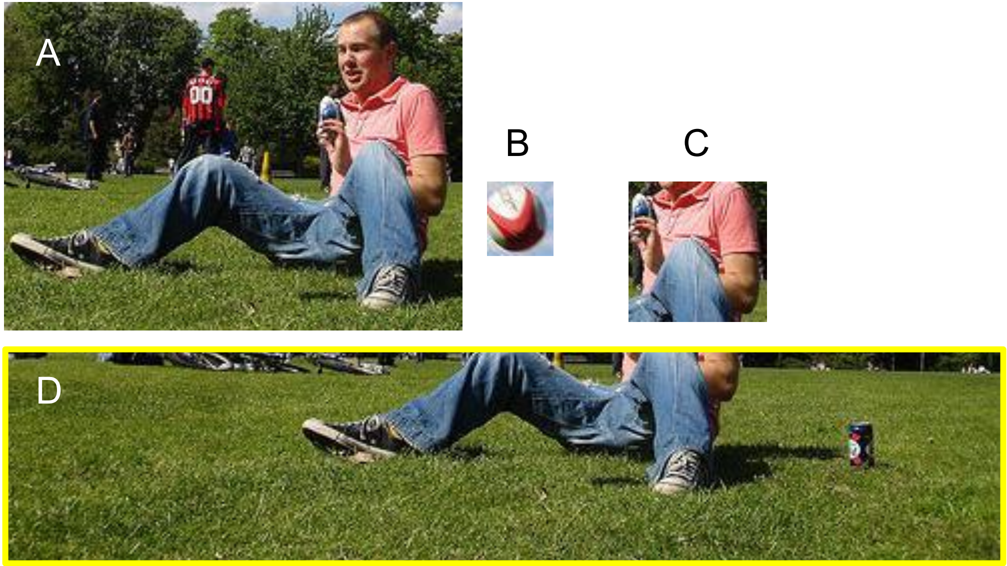
Region-attentive multimodal NMT

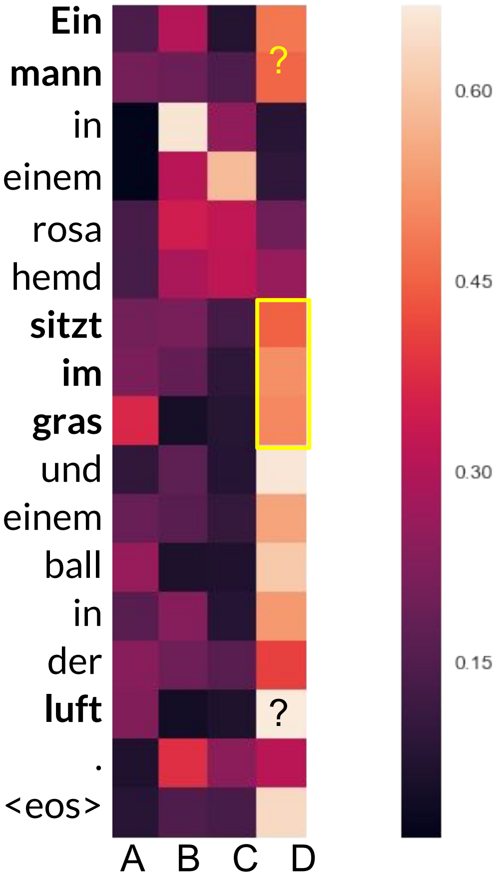
- Segment image into its objects
- Use a CNN to extract **features** from regions
- Attention over these regions
- **Idea**: **alignment** between regions & words in target language
- z_t is the fusion of both contexts
  - Concatenation
  - Sum
  - Hierarchical



Attend to image regions - concat

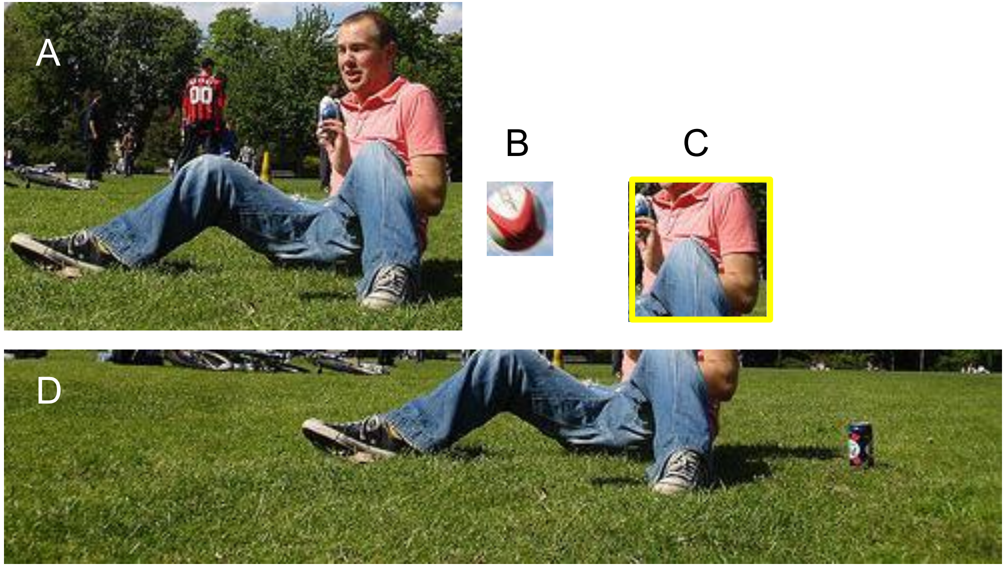
S: A man in a pink shirt is sitting in the grass and a ball is in the air.

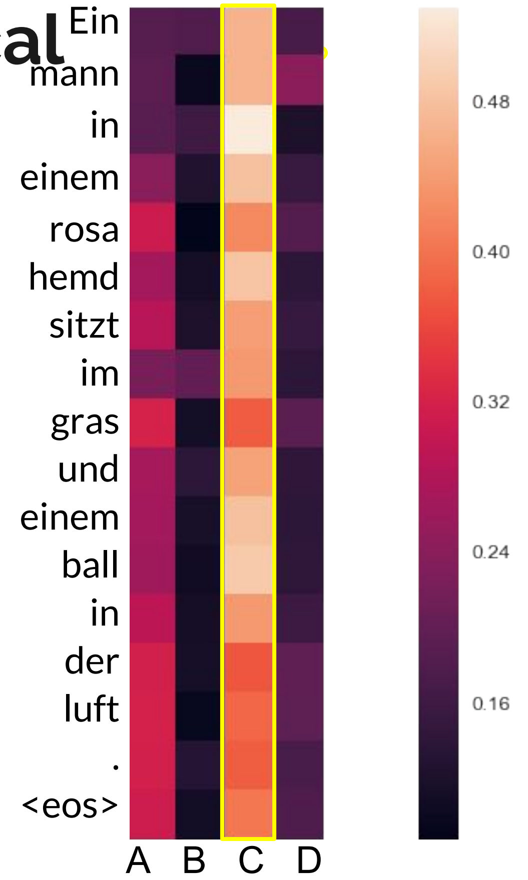




Attend to image regions - hierarchical

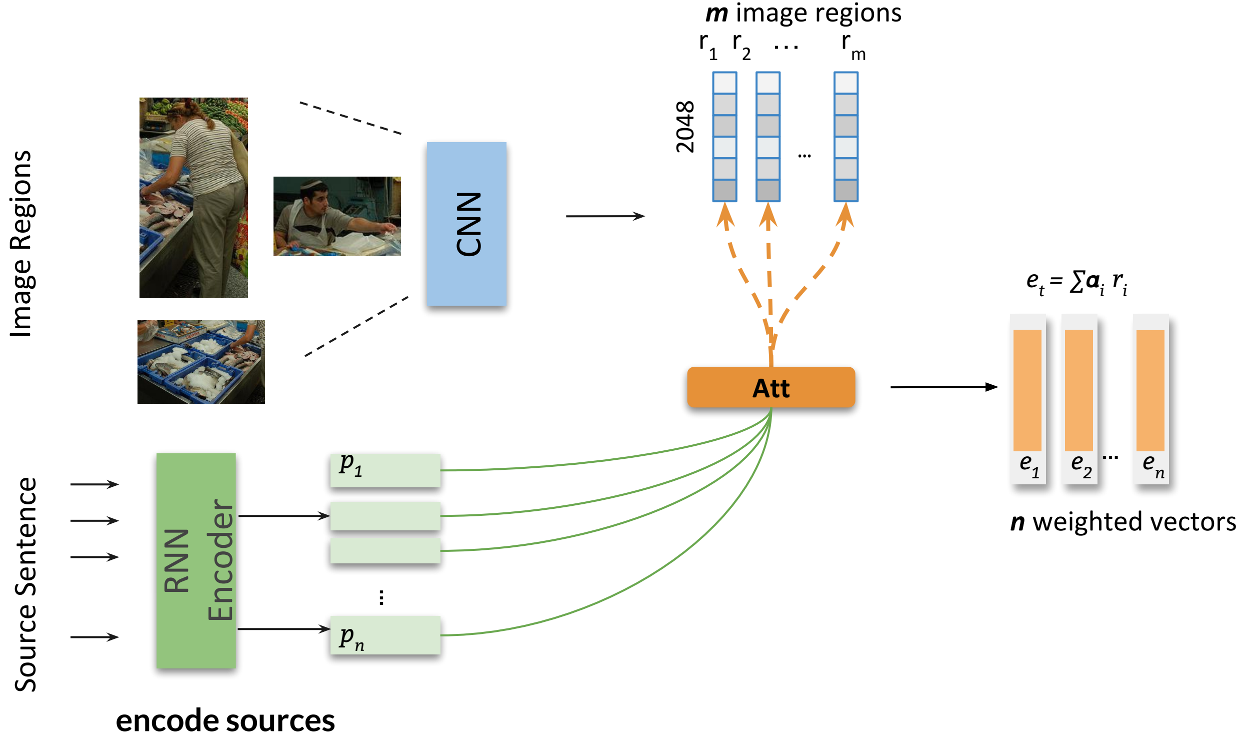
S: A man in a pink shirt is sitting in the grass and a ball is in the air.



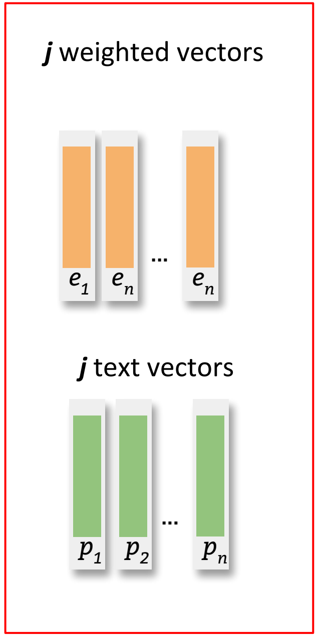


Attention at encoding

Idea: Ground the images in the ***source***

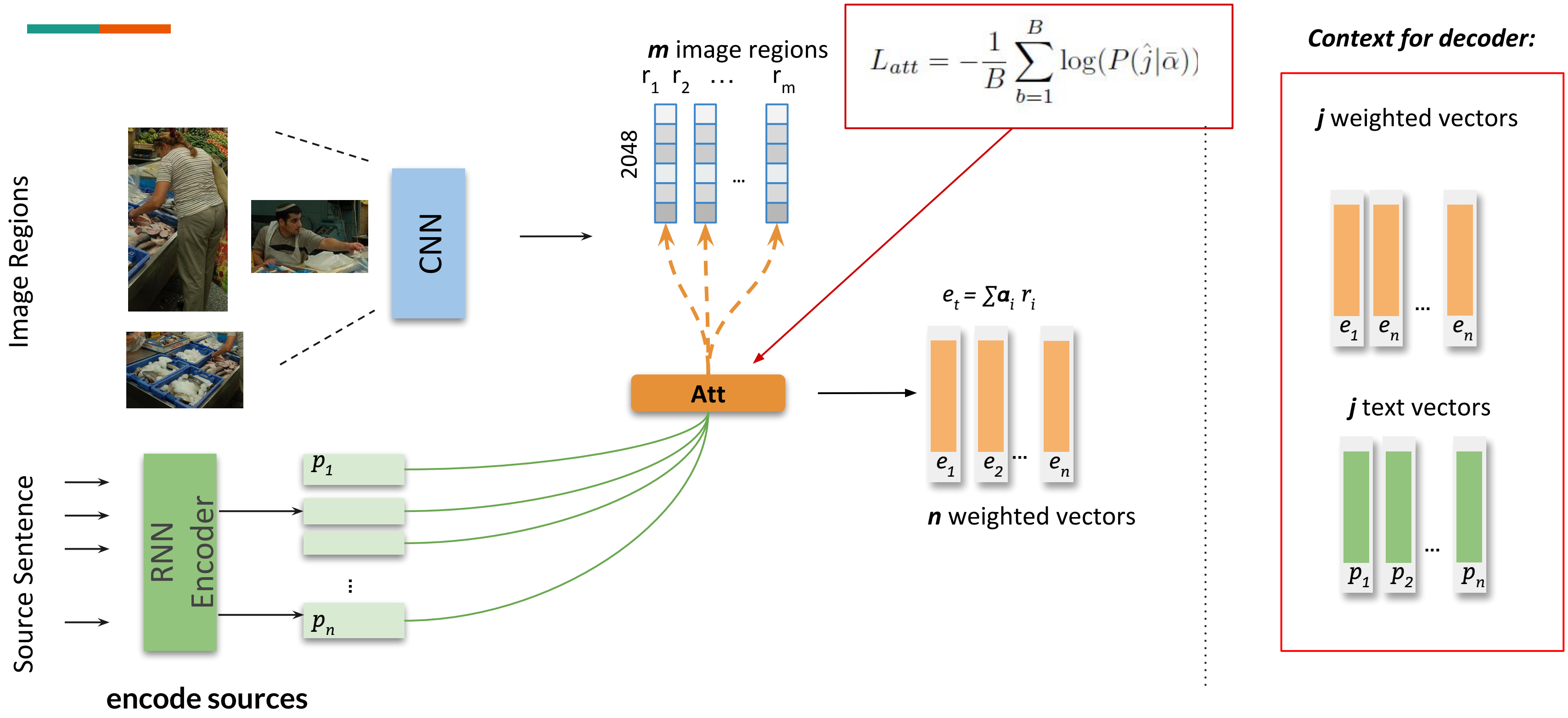


Context for decoder:



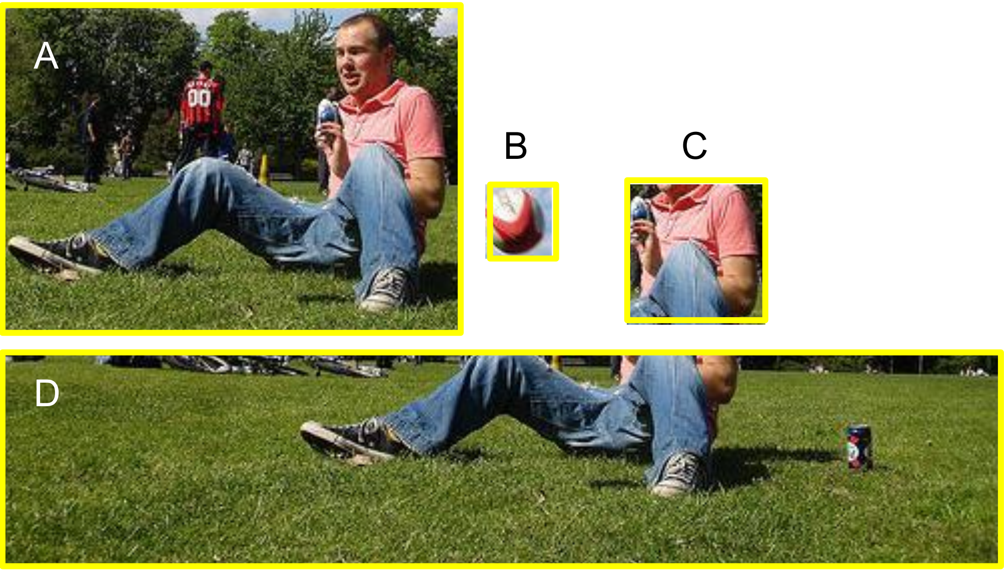
Given gold word-region alignments, add an auxiliary loss to main MT loss

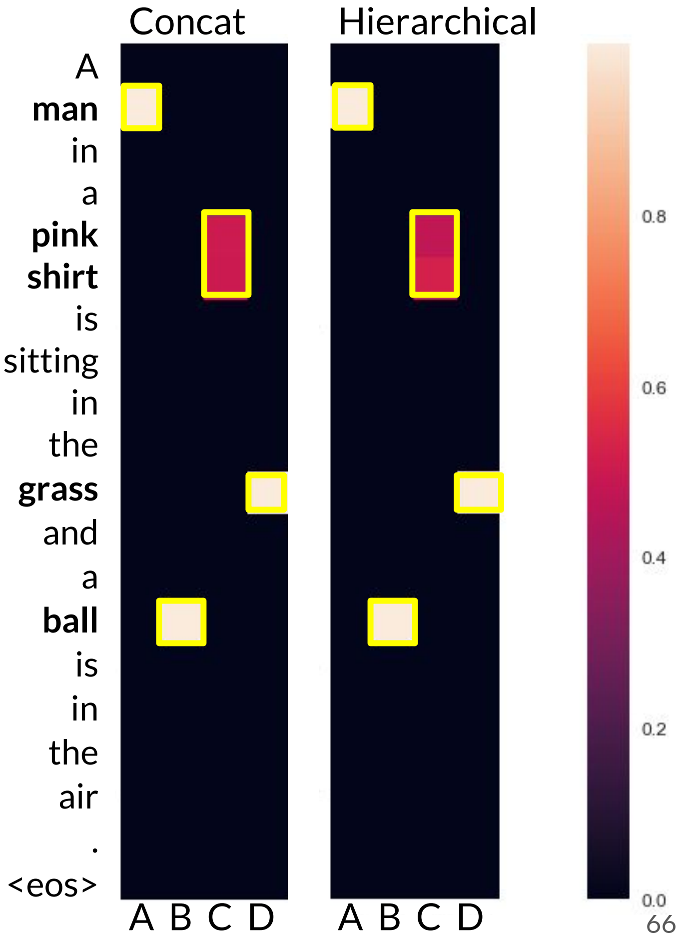
Attention at encoding



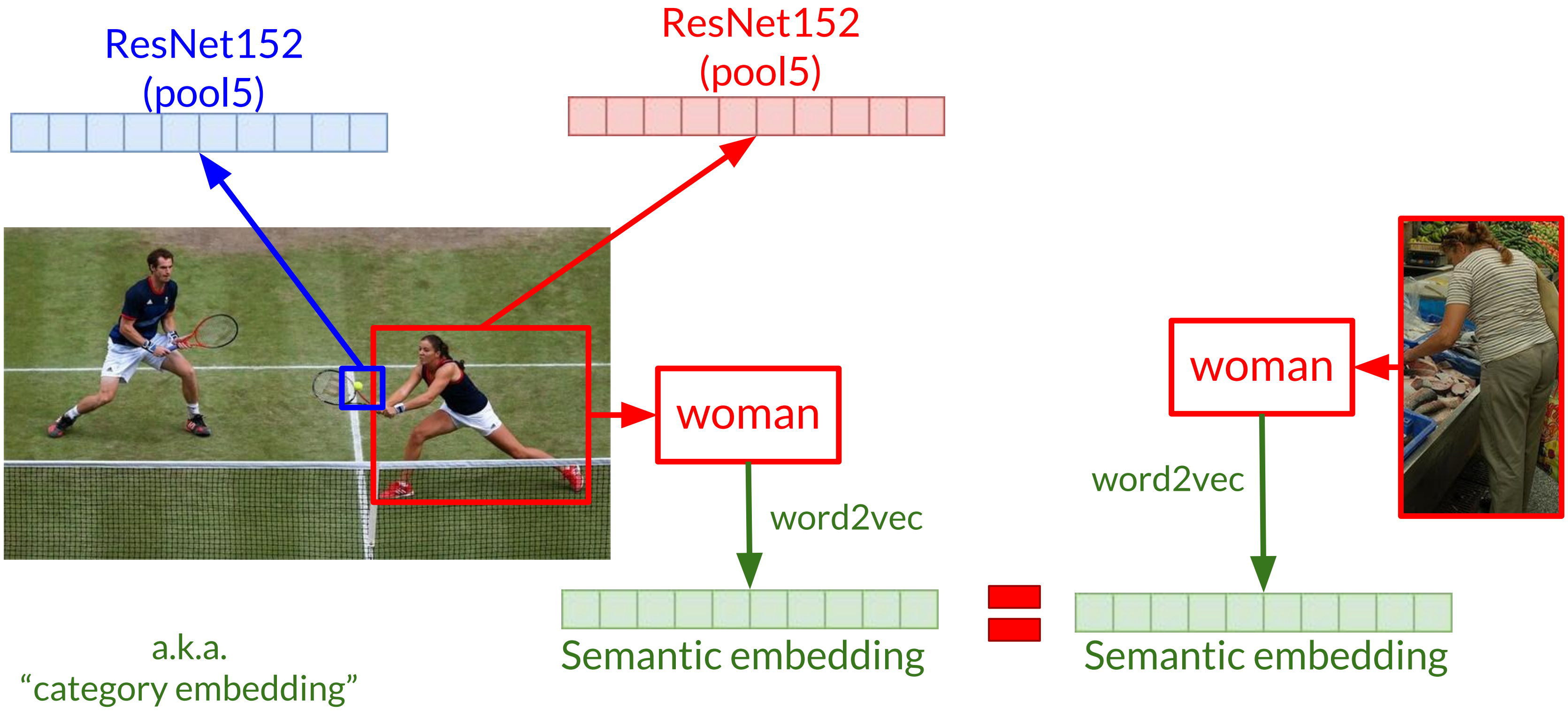
Attention at encoding

S: A man in a pink shirt is sitting in the grass and a ball is in the air.



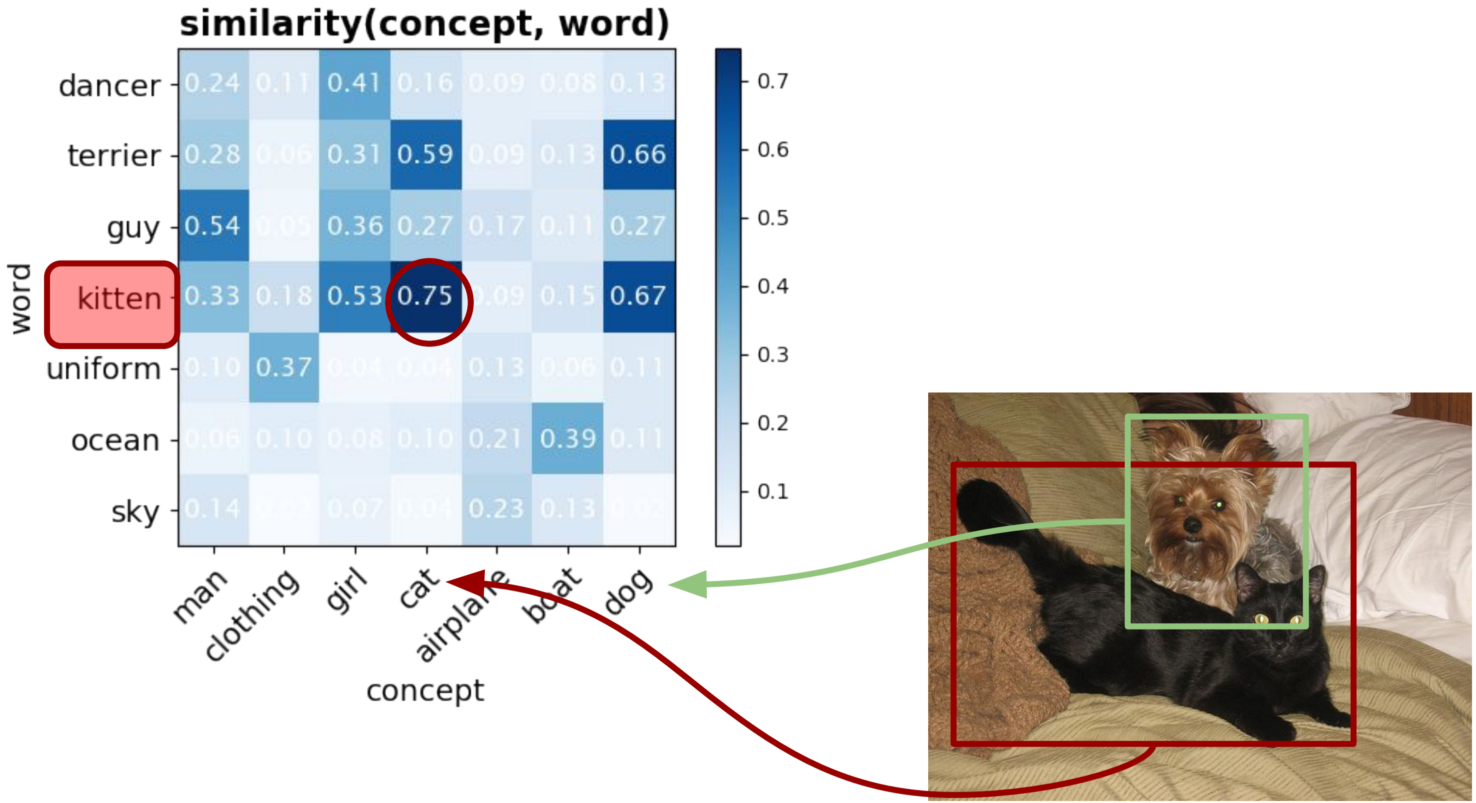


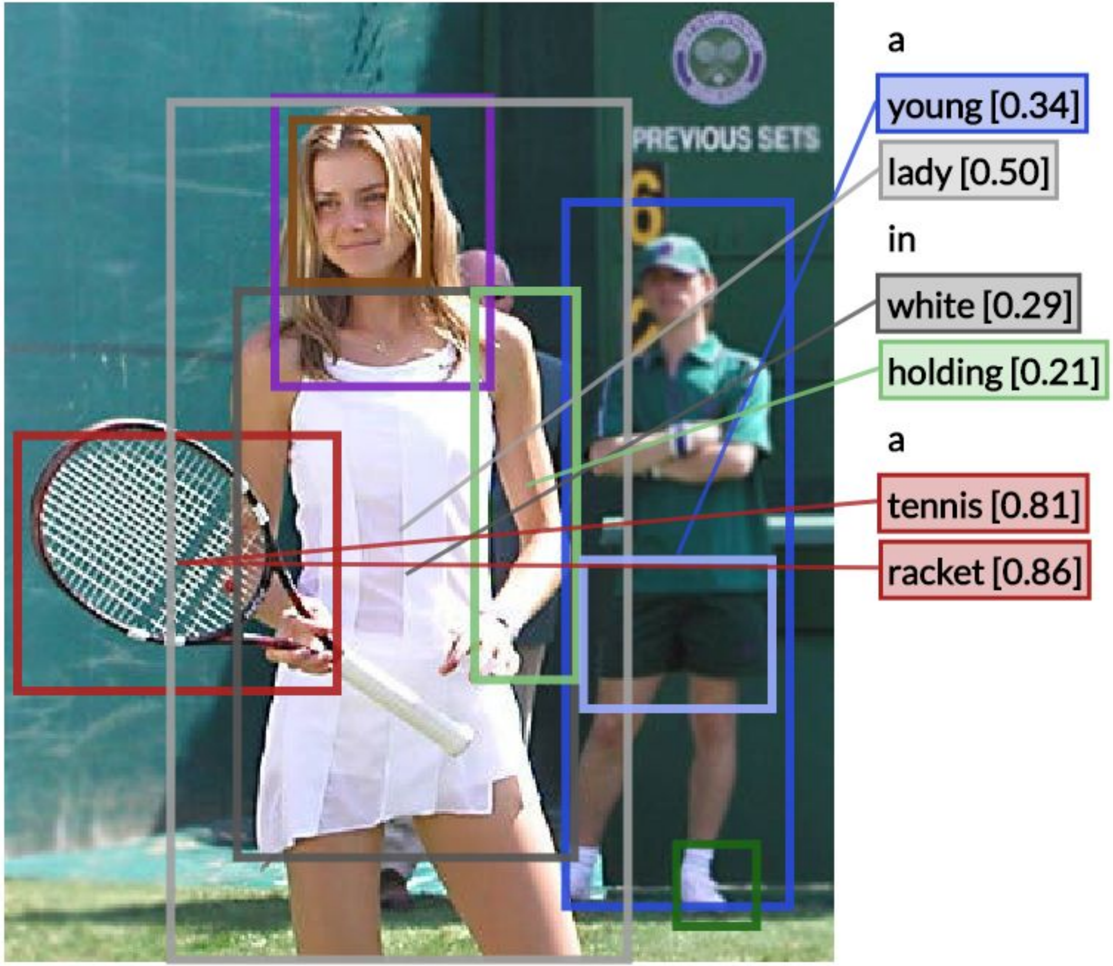
Representing image regions

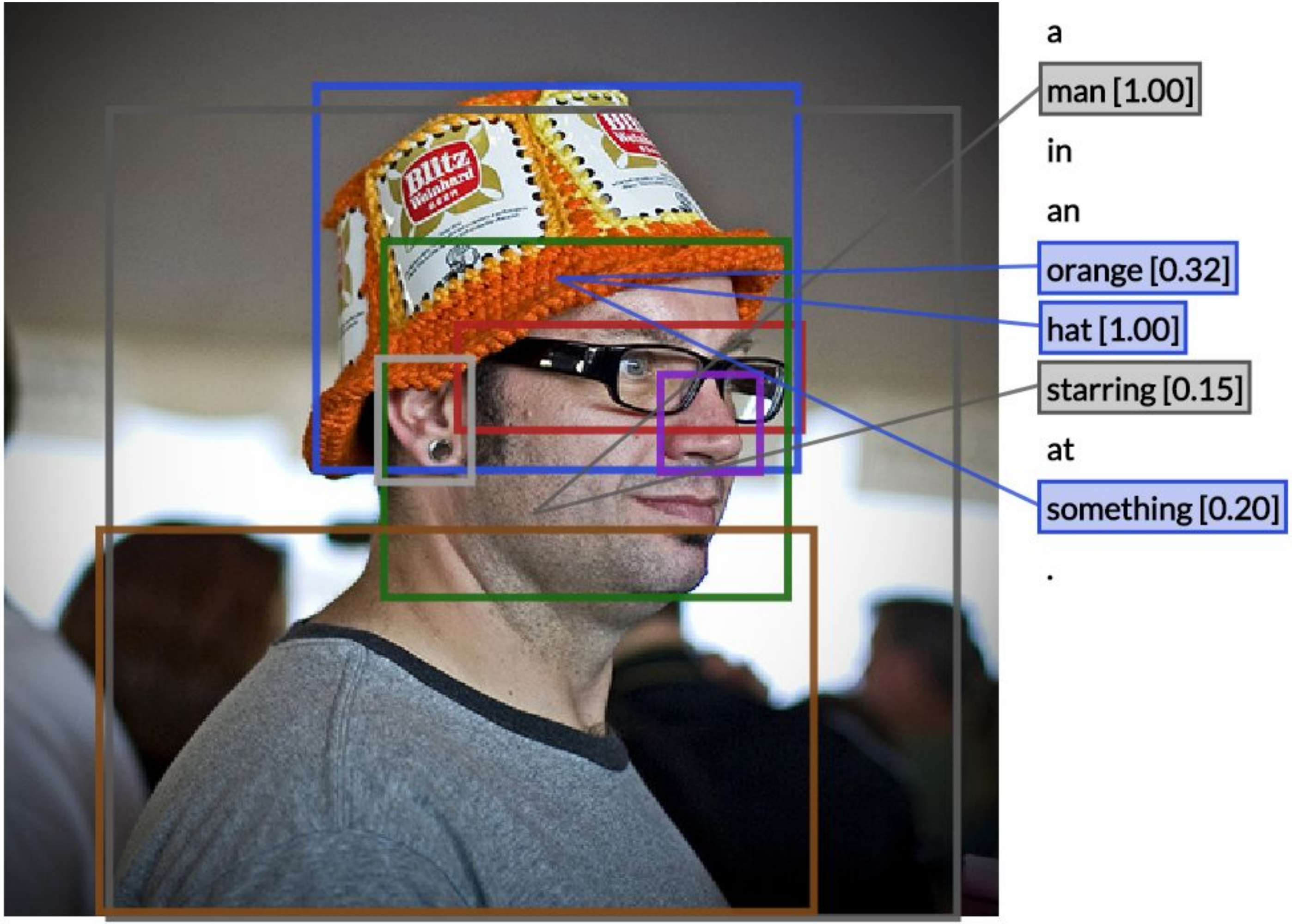


Explicit alignments

Alignments learnt explicitly




PREVIOUS SETS
a
young [0.34]
lady [0.50]
in
white [0.29]
holding [0.21]
a
tennis [0.81]
racket [0.86]


a
man [1.00]
in
an
orange [0.32]
hat [1.00]
starring [0.15]
at
something [0.20]
.

Idea

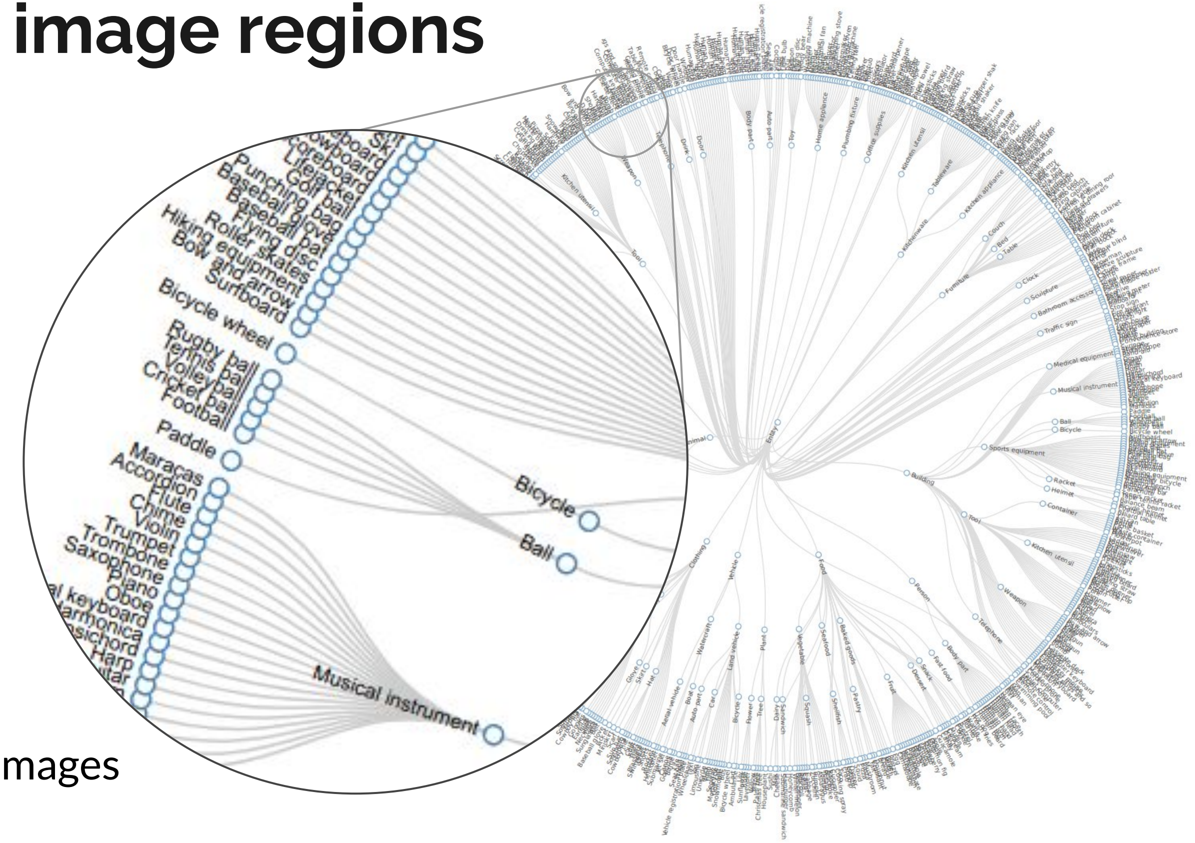
Further specify source words with respective image region visual info

Category:
clothing

The man in yellow pants is raising his arms

Categories from image regions

- Oracle (8)
  - People
  - Clothing
  - Scene
  - Animals
  - Vehicles
  - Instruments
  - Body parts
  - Other



- Predicted (545) - Open Images

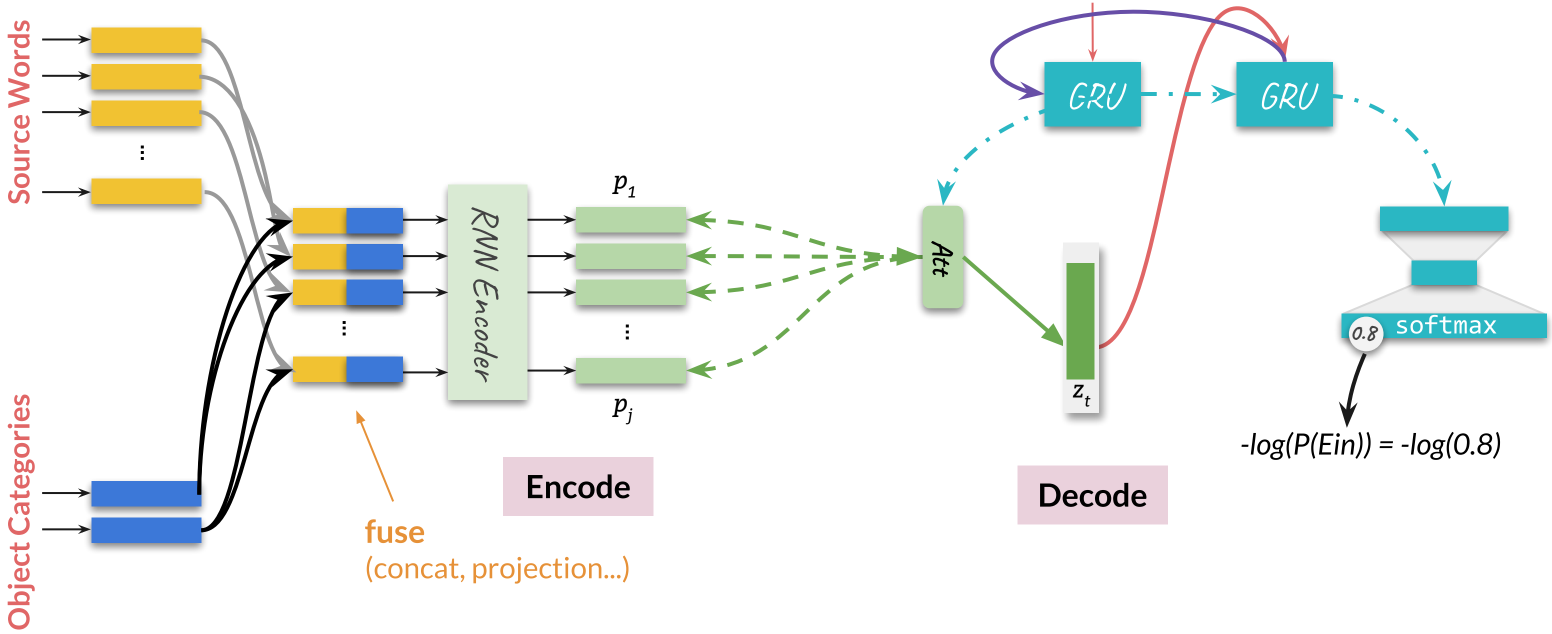
Category embeddings for grounding

- Take category of image region to describe nouns

Sentence:	The	man	in	yellow	pants	is	raising	his	arms
		⇩			⇩				⇩
Categories:		people			clothing				body part

- Take pre-trained word embeddings of category to be visual info
- For any other word, set category to “empty” or to word itself

Category embeddings for grounding



Results (test2016)

METEOR	Features	en-de	en-fr	en-cs
Text-only (no image)	-	57.35	75.16	29.35
Decoder init. (full image)	Pool5	56.97	74.82	29.04
Attention over regions (decoder)	Pool5	56.77	74.74	28.86
Attention over regions (decoder)	Cat. embeddings	56.48	73.65	28.42
Encoder attention over regions	Pool5	57.30	75.36	**30.48**
Encoder attention over regions	Cat. embeddings	57.29	**75.97**	**30.78**
Supervised attention over regions	Pool5	56.34	75.07	**30.19**
Supervised attention over regions	Cat. embeddings	56.64	75.56	**30.39**
Explicit alignment - projection	Cat. embeddings	57.39	75.25	**30.64**
Explicit alignment - concatenation	Cat. embeddings	57.44	75.47	**30.77**

Results - human eval

- Proportion of times each system is better (meaning preservation)

	Features	en-de	en-fr	en-cs
Text-only (no image)	-	22%	32%	20%
Multimodal	Pool5	[illegible]3%	37%	34%
	Cat. em[illegible]ed	+ [illegible]% = 78%	+ 32% = 68%	+ 46% = 80%

- Text-only system is more fluent but has less correct content words

Conclusions

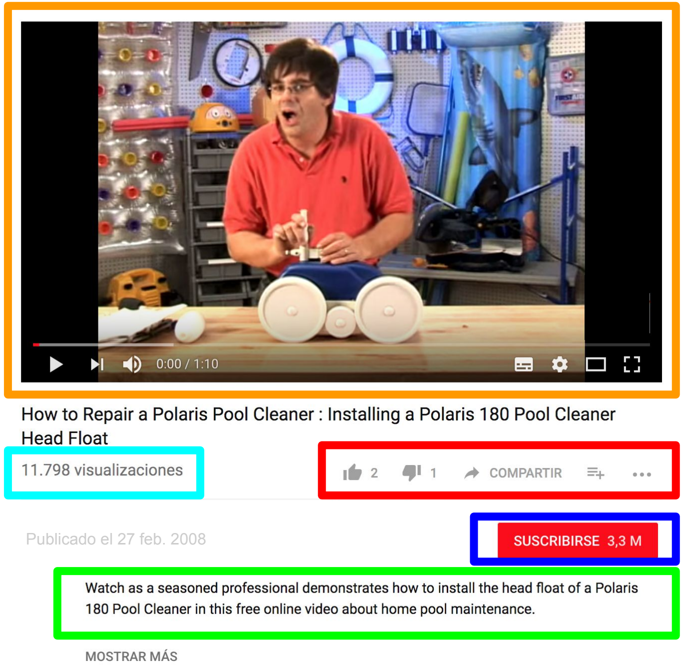
- **Text-only** vs **region-specific**
  - Region-specific always better
- **Oracle** vs **predicted** regions and alignment
  - Predictions do not degrade performance substantially
- Representations: **pool5** vs **category embeddings**
  - Similar but category embeddings more interpretable
- **Meteor/BLEU** are not indicative of performance variations
  - Human evaluation: much more telling

Future of MMT: better use of explicit & implicit **alignments**, better **evaluation**, more challenging **data**

New dataset

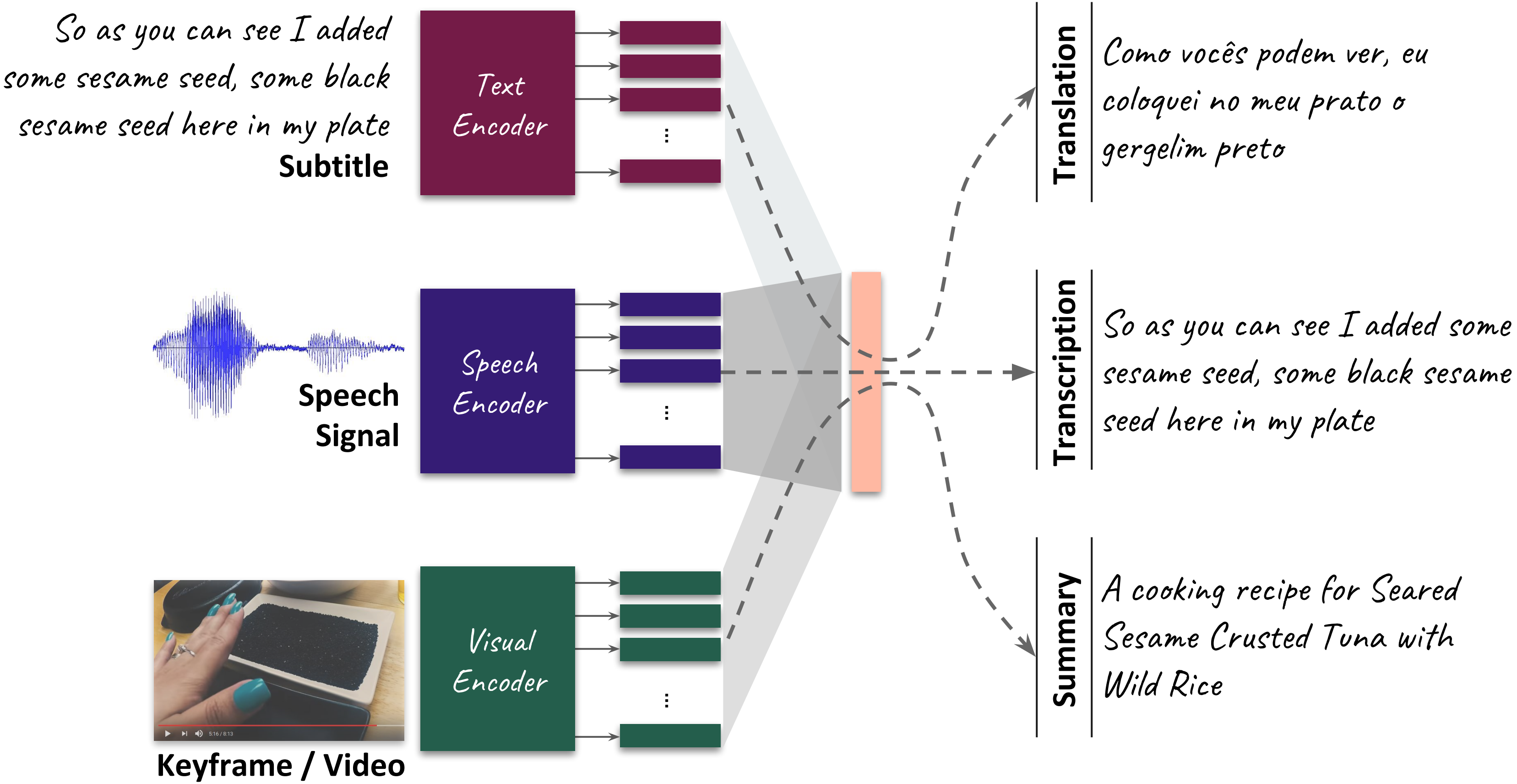
How2 dataset

- 2000h of **how-to** videos (Yu et al., 2014)
  - 300h for MT
- Ground truth English captions
- Metadata
  - Number of likes / dislikes
  - Visualizations
  - Uploader, Date
  - Tags
- Video descriptions (“summaries”)
  - 80K descriptions for 2000h
- Very different topics
  - Cooking, fixing things, playing instruments, etc.
- 300,000 segments translated into Portuguese



How2 dataset - example

How2 dataset - what can one do?



Questions?

References

- Bahdanau, D., Cho, K., and Bengio, Y. (2014). **Neural machine translation by jointly learning to align and translate**. In ICLR 2014.
- 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 Proc. of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pages 432–439, Copenhagen, Denmark.
- Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016a). **Does multimodality help human and machine for translation and image captioning?** In Proc. of the First Conference on Machine Translation, pages 627–633, Berlin, Germany.
- Caglayan, O., Barrault, L., and Bougares, F. (2016b). **Multimodal attention for neural machine translation**. CoRR, abs/1609.03976.
- Calixto, I., Elliott, D., and Frank, S. (2016). **DCU-UVA multimodal mt system report**. In Proc. of the First Conference on Machine Translation, pages 634–638, Berlin, Germany.

References

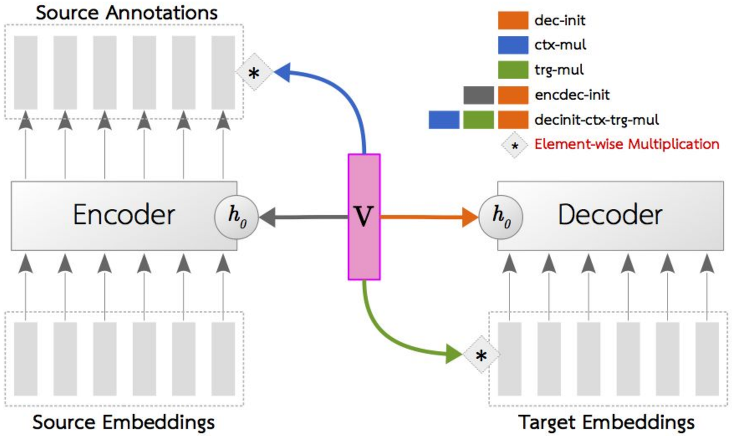
- Delbrouck, J. and Dupont, S. (2017). **Multimodal compact bilinear pooling for multimodal neural machine translation**. CoRR, abs/1703.08084.
- Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). **Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description**. In Proc. of the Second Conference on Machine Translation, Copenhagen, Denmark.
- Elliott, D. and Kádár, A. (2017). **Imagination improves multimodal translation**. In Proc. of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 130–141, Taipei, Taiwan.
- Firat, O., Cho, K., Sankaran, B., Yarman Vural, F. T., and Bengio, Y. (2017). **Multi-way, multilingual neural machine translation**. Computer Speech and Language., 45(C):236–252.
- Fukui, A. , Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., **Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding**, EMNLP 2016
- Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). **Attention-based multimodal neural machine translation.** In Proc. of the First Conference on Machine Translation, pages 639–645, Berlin, Germany. Association for Computational Linguistics.
- Libovický, J. and Helcl, J. (2017). **Attention strategies for multi-source sequence-to-sequence learning.** In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 196–202.

References

- Madhyastha, P. S., Wang, J., and Specia, L. (2017). **Sheffield multimt: Using object posterior predictions for multimodal machine translation**. In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pages 470–476, Copenhagen, Denmark.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). **Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models**. International Journal of Computer Vision, 123(1):74–93
- Shah, K., Wang, J., and Specia, L. (2016). **Shef-multimodal: Grounding machine translation on images**. In Proc. of the First Conference on Machine Translation, pages 660–665, Berlin, Germany. Association for Computational Linguistics.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). **Show, attend and tell: Neural image caption generation with visual attention**. CoRR, abs/1502.03044.

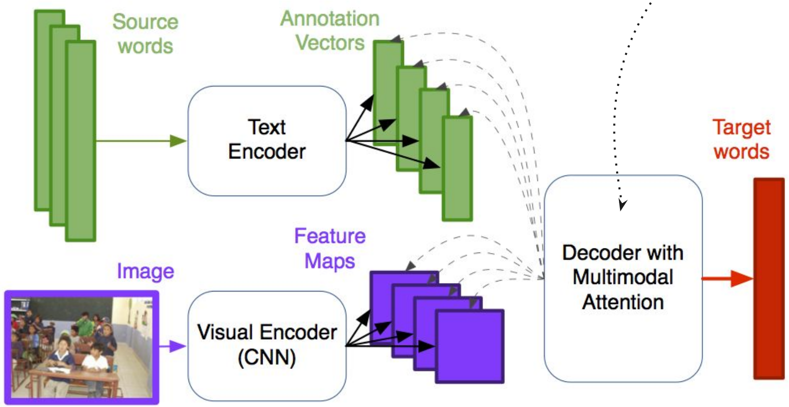
Integration: fixed size visual information

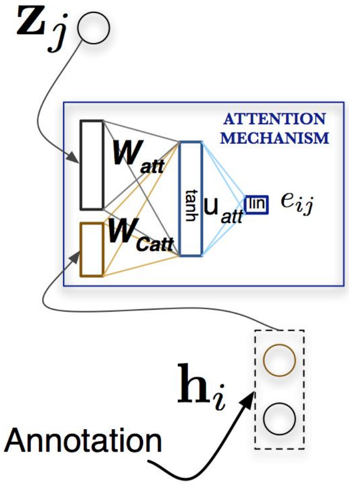
- Prepending and/or appending visual vectors to source sequence
  - Huang et al., 2016
- Decoder initialization
  - Calixto et al., 2016
- Multiplicative interaction schemes
  - Caglayan et al., 2017, Delbrouck and Dupont, 2017
- ImageNet class probability vector as features
  - Madhyastha et al., 2017



Integration: fusion, multimodal attention

- Two attention mechanisms
  - Caglayan et al., 2016a, 2016b
  - Calixto et al, 2016
  - Libovický and Helcl, 2017





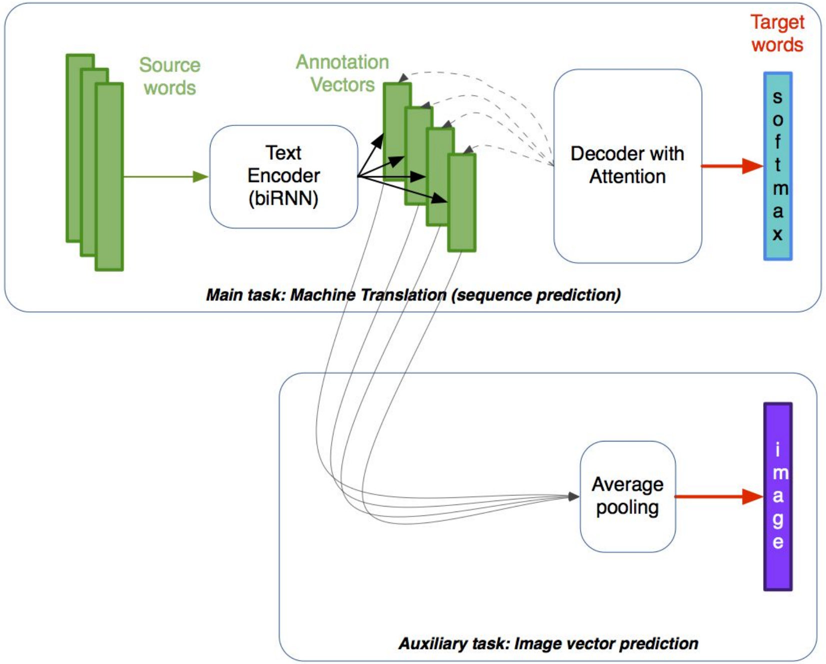
Shared vs. distinct weights for both modalities

Integration: multitask learning -- Imagination

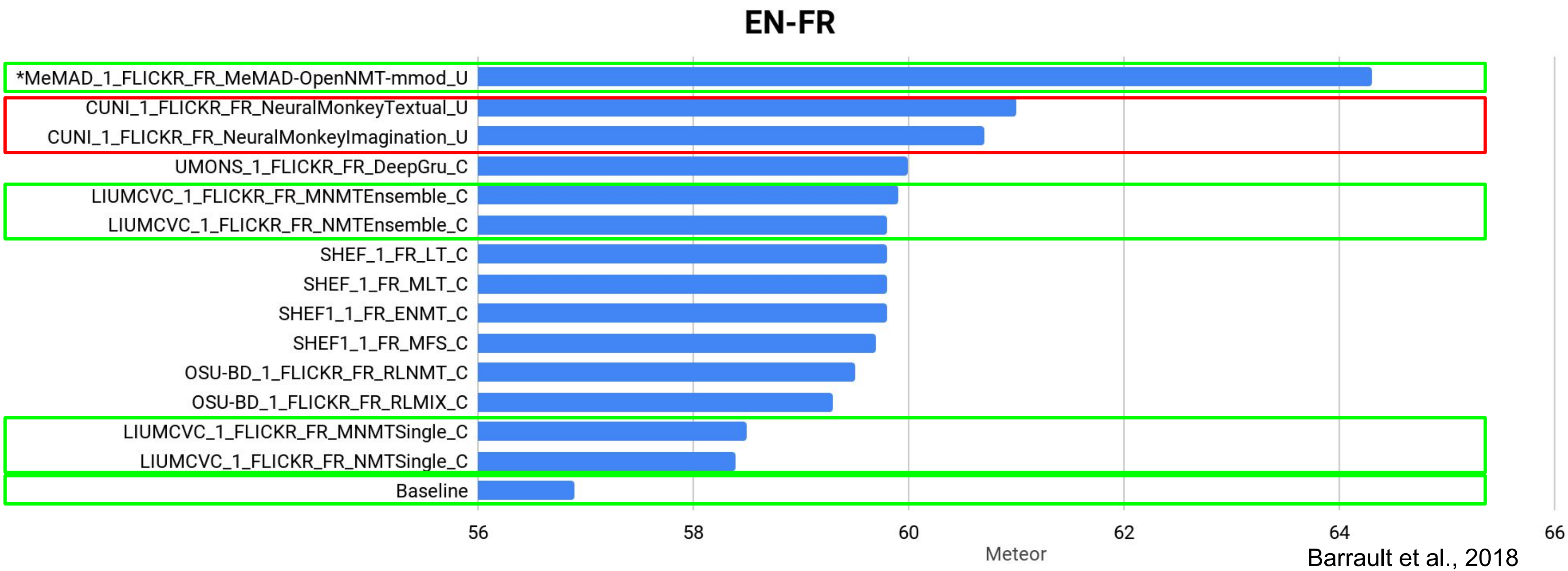
- Predict image vector from source sentence during training only

➢ Gradient flow from image vector impact the source text encoder and embeddings

  ○ Elliott and Kádár (2017)

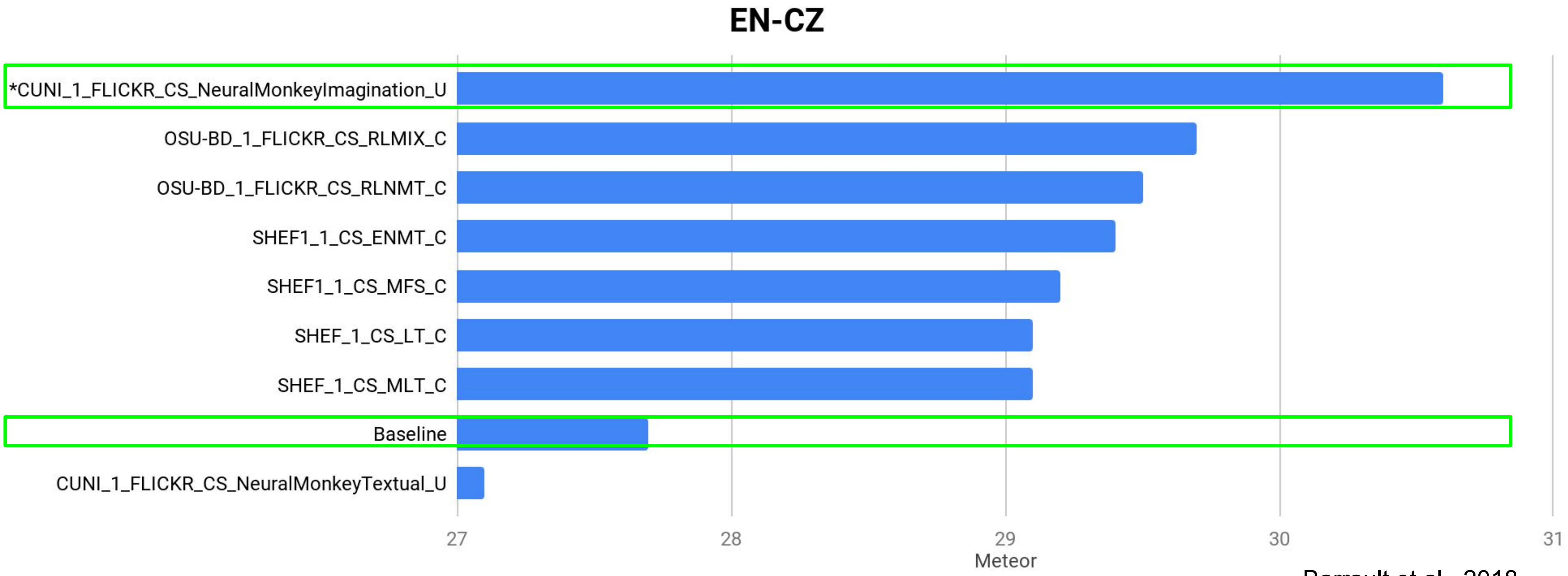


Results from WMT shared task - 2018



Barrault et al., 2018

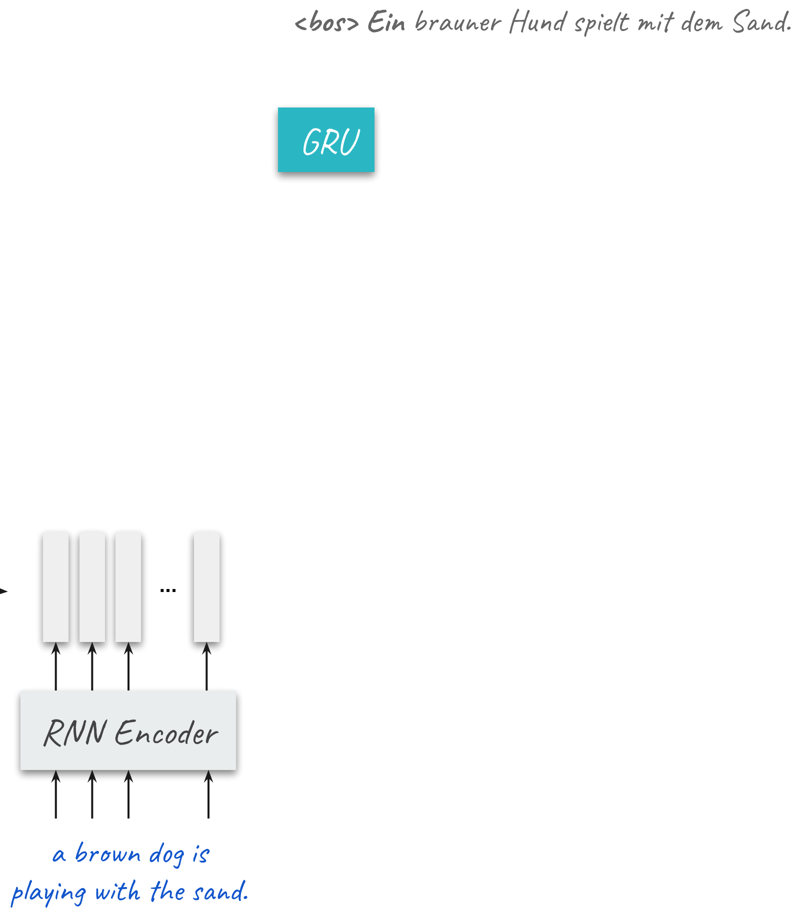
Results from WMT shared task - 2018



Barrault et al., 2018

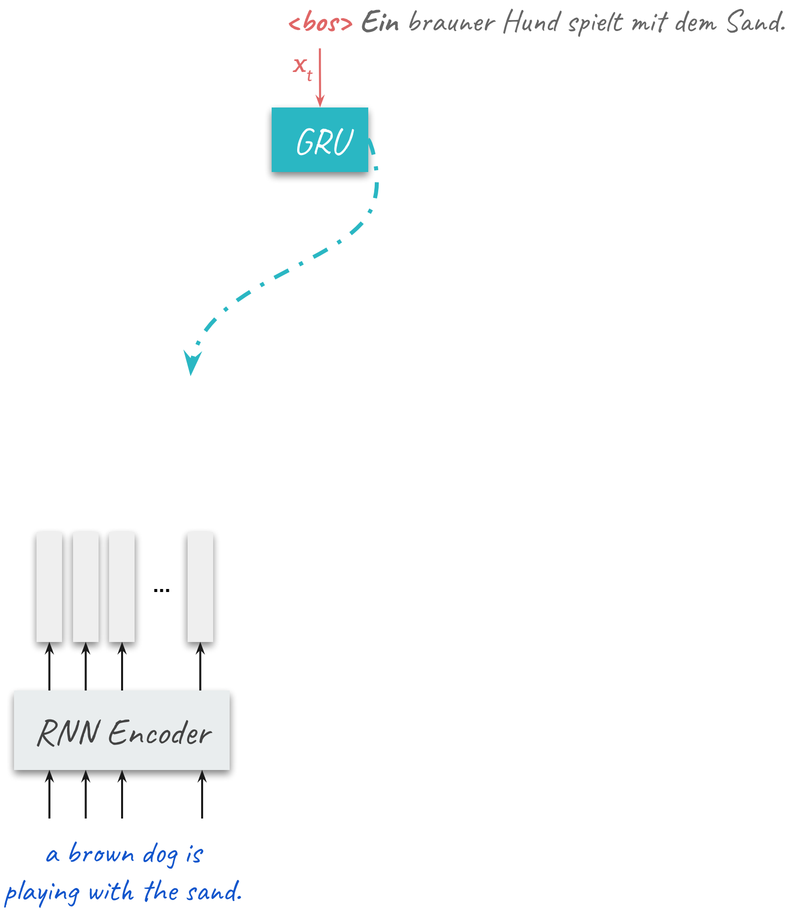
NMT with conditional GRU

- Encode source sentence with an RNN to obtain the annotations.



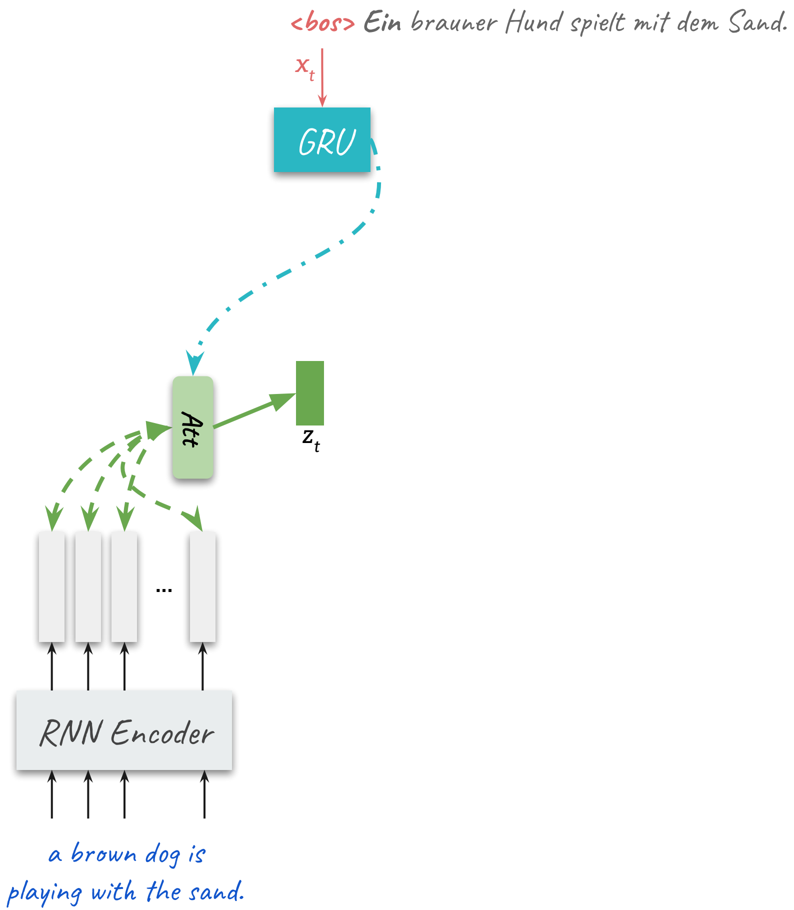
NMT with conditional GRU

- Encode source sentence with an RNN to obtain annotations.
- First decoder RNN consumes a target embedding to produce a hidden state.



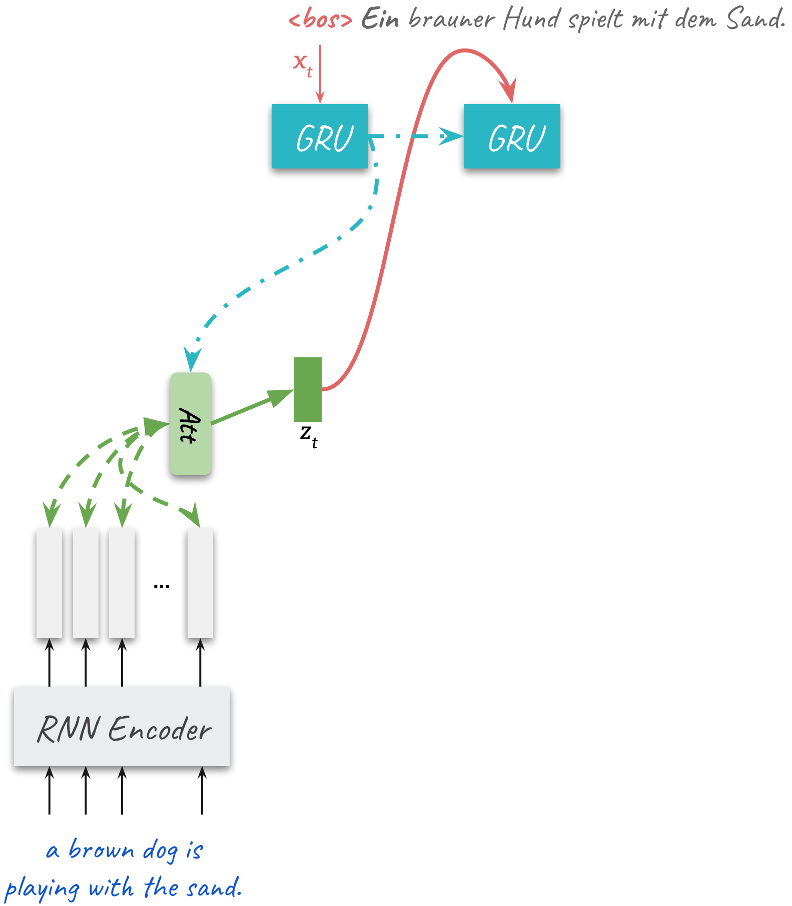
NMT with conditional GRU

- Encode source sentence with an RNN to obtain annotations.
- First decoder RNN consumes a target embedding to produce a hidden state.
- Attention block takes this hidden state and the annotations to compute the so-called "context vector" z_t which is the weighted sum of annotations.



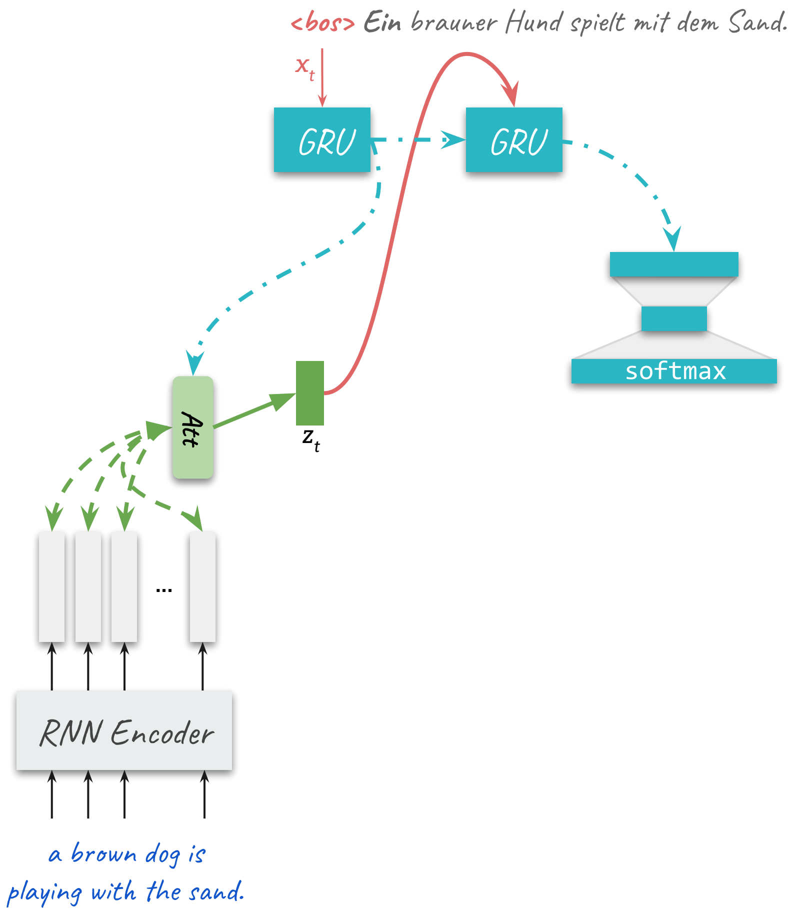
NMT with conditional GRU

- z_t becomes the input for the second RNN. (The hidden state is carried over as well.)



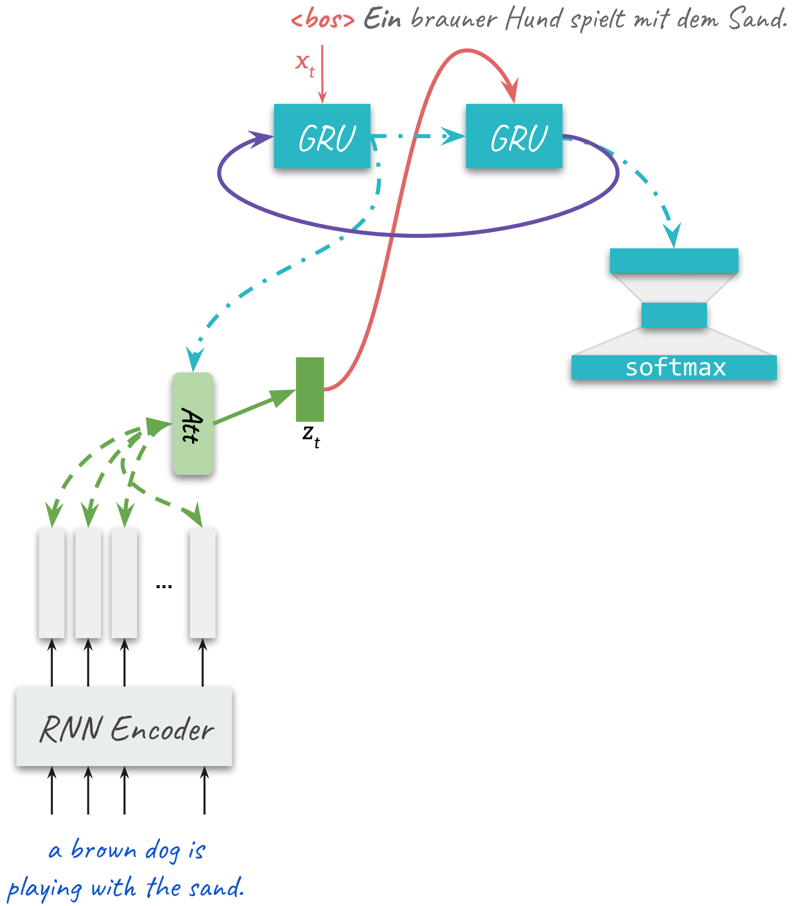
NMT with conditional GRU

- z_t becomes the input for the second RNN. (The hidden state is carried over as well.)
- The final hidden state is then projected to the size of the vocabulary and target token probability is obtained with *softmax()*



NMT with conditional GRU

- z_t becomes the input for the second RNN. (The hidden state is carried over as well.)
- The final hidden state is then projected to the size of the vocabulary and target token probability is obtained with *softmax()*
- Same hidden state is fed back to first RNN for the next timestep.



NMT with conditional GRU

- The loss for a decoding timestep is the negative log-likelihood of the ground-truth token.