Introduction to Multimodal Machine Translation

Loïc Barrault, University of Le Mans

Thanks to Ozan Caglayan for sharing some slides

Motivations

- Semantics still poorly used in MT systems
  - Embeddings seem to convey such information
- Can meaning be modelled from text only?
- We argue that we can’t learn everything from books!
  - Language grounding
  - Use of multiple modalities

Example 1: 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.

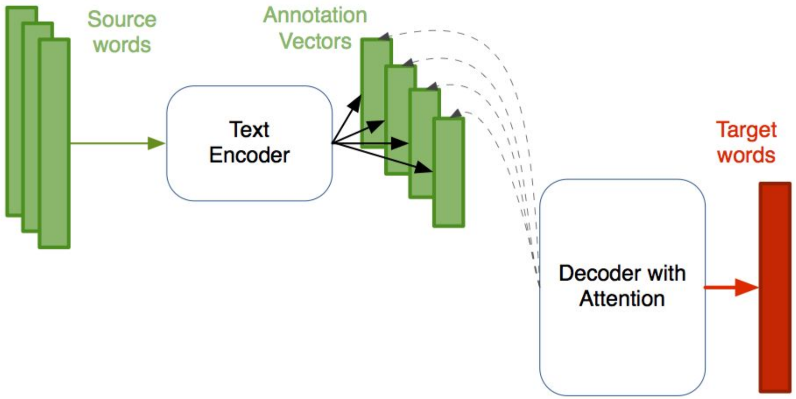
Example 2: semantics

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

Neural Machine Translation (quick recap)

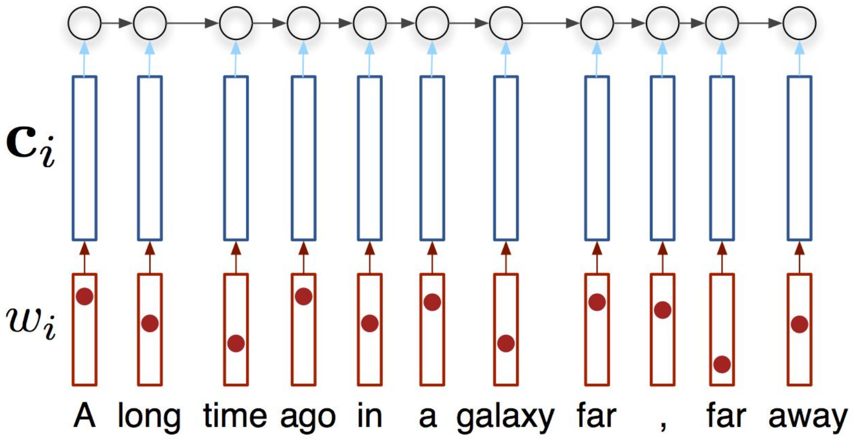
Neural Machine Translation: General picture

- Encoder decoder architecture equipped with attention mechanism
  - Encode the source sentence (generally using a bidirectional-RNN)
  - Generate an intermediate representation (source context vector)
    - used to be static
    - becomes dynamic with the attention mechanism
  - Decoder is a conditional target language model
    - conditioned on source context



Should remind you the presentation by P. Koehn earlier this week!

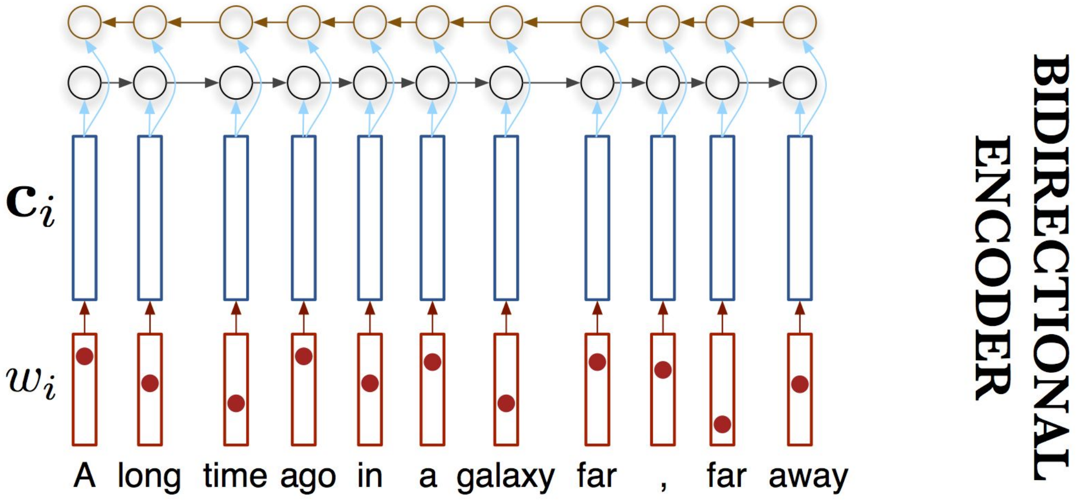
Neural Machine Translation



BIDIRECTIONAL ENCODER

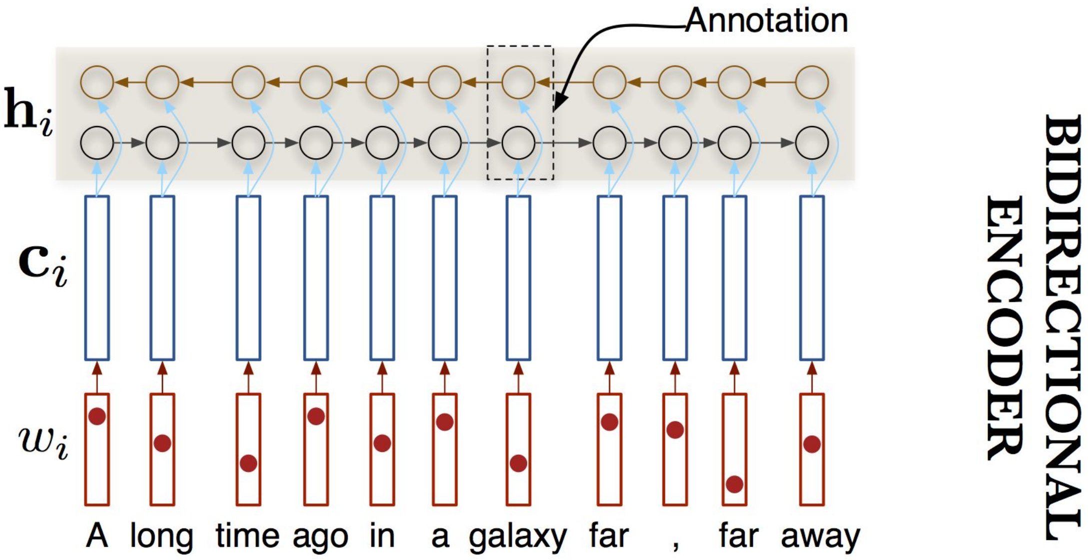
1. 1-hot encoding + projection + update **forward** RNN hidden state

Neural Machine Translation



1bis. update **backward** RNN hidden state

Neural Machine Translation

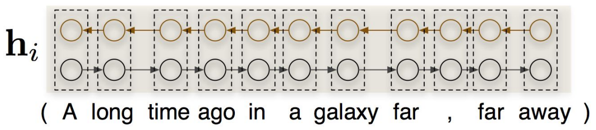


2. **Annotation** = concat **forward** and **backward** vectors
   Every $\mathbf{h}_i$ encodes the whole sentence with a focus on the i-th word

NMT principle

2. Decoder gets the annotations.

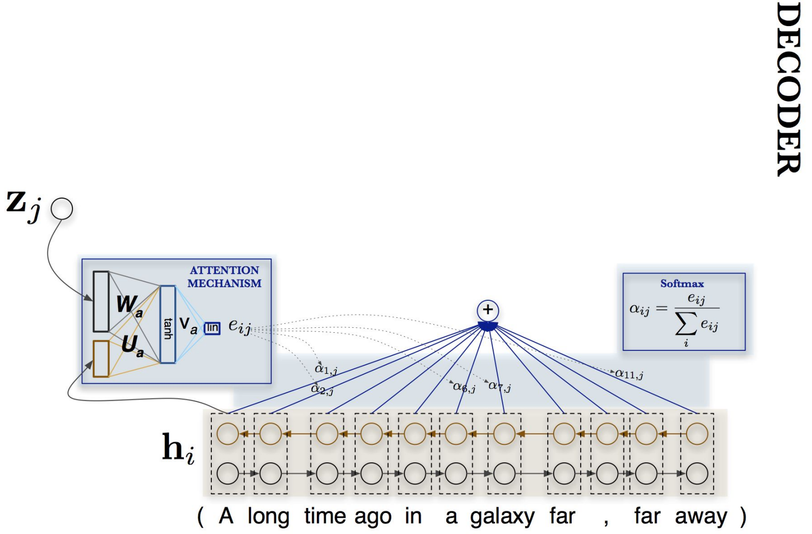
DECODER



NMT principle

2. Decoder gets the annotations.
3. Attention weights are computed
   a. Feed forward NN
   b. Weighted mean

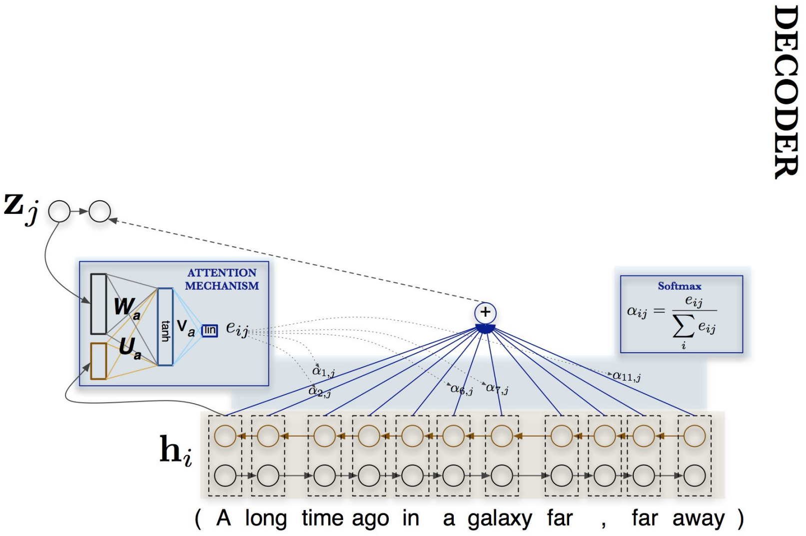
$$\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$$



NMT principle

2. Decoder gets the annotations.
3. Attention weights are computed
   a. Feed forward NN
   b. Weighted mean
4. Update hidden state of GRU

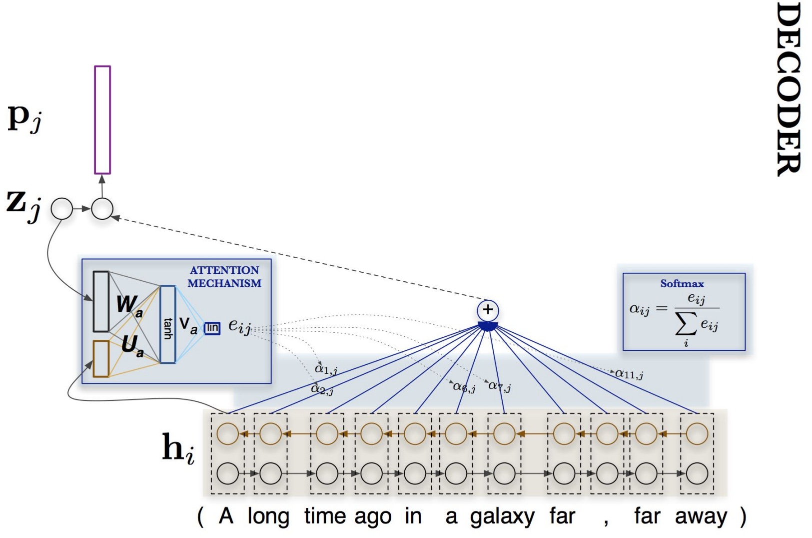
$$\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$$



NMT principle

2. Decoder gets the annotations.
3. Attention weights are computed
   a. Feed forward NN
   b. Weighted mean
4. Update hidden state of GRU
5. Probability distribution for all words

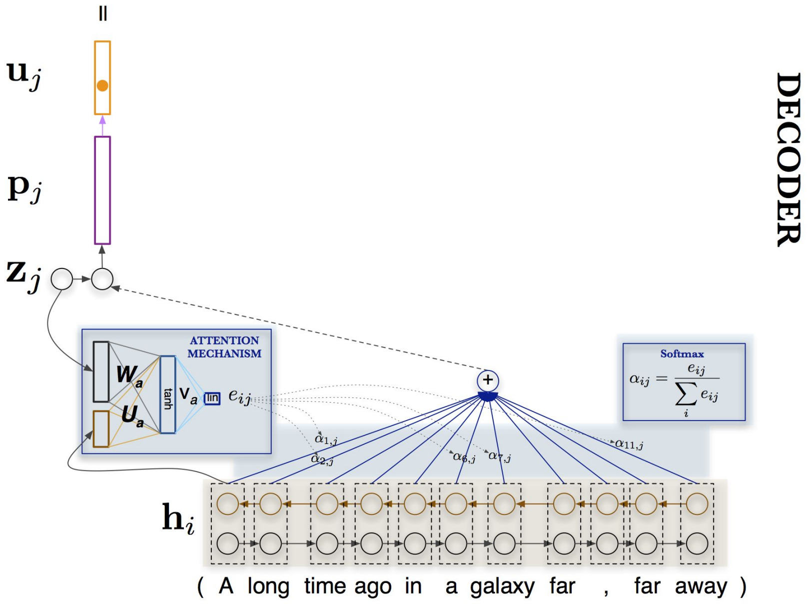
$$\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$$



NMT principle

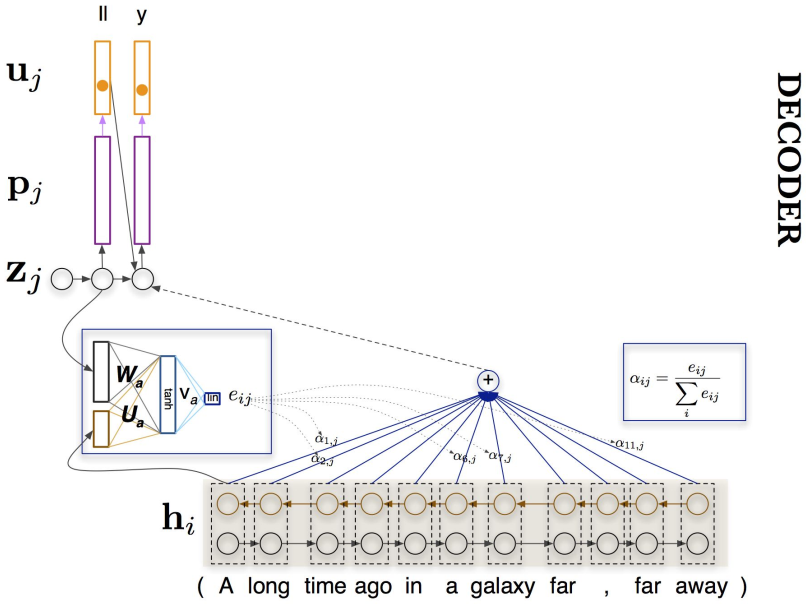
2. Decoder gets the annotations.
3. Attention weights are computed
    a. Feed forward NN
    b. Weighted mean
4. Update hidden state of GRU
5. Probability distribution for all words
6. Generate next word
    a. Most probable of beam

$$\tilde{\mathbf{h}}_{\mathbf{j}} = \sum_i \alpha_{ij} \mathbf{h}_i$$



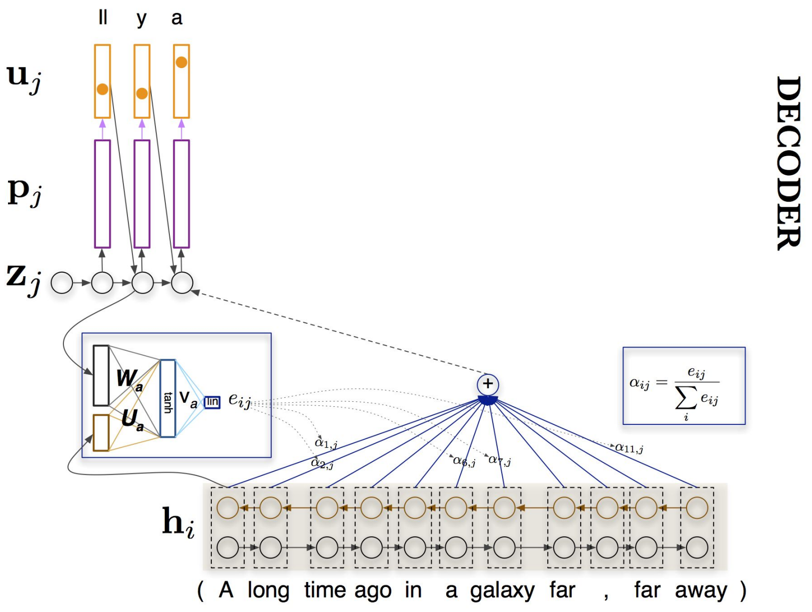
NMT principle

- Decoder RNN is using the source context and embedding of previous generated word
- This is a simplified view, see Ozan's part



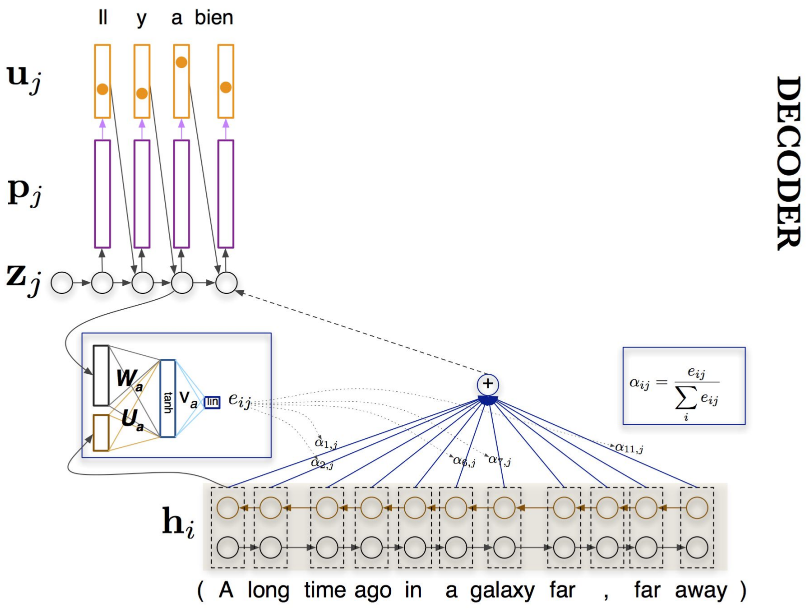
NMT principle

- At each timestep, a new set of attention weights is computed
- **Annotations don't change**
- **Hidden state of decoder RNN has changed!**



NMT principle

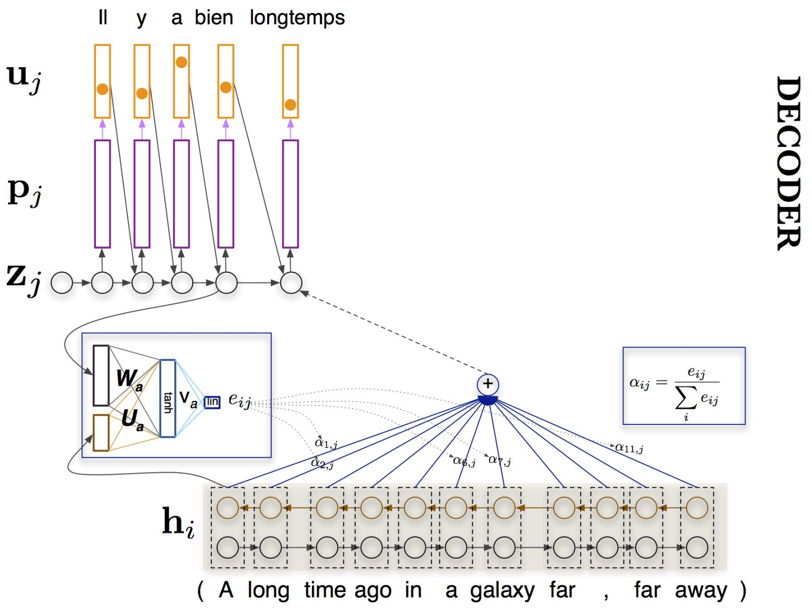
- At each timestep, a new set of attention weights is computed
- Annotations don't change
- Hidden state of decoder RNN has changed!
- Decoder is a conditional LM



NMT principle

- At each timestep, a new set of attention weights is computed
- Annotations don’t change
- Hidden state of decoder RNN has changed!
- **Decoder is a conditional LM**

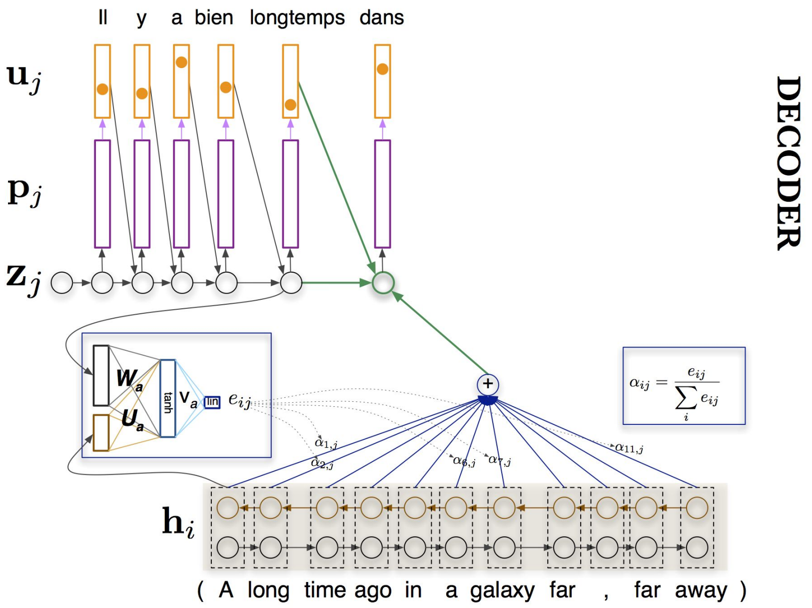
- And so on and so forth...



NMT principle

- At each timestep, a new set of attention weights is computed
- Annotations don't change
- Hidden state of decoder RNN has changed!
- **Decoder is a conditional LM**

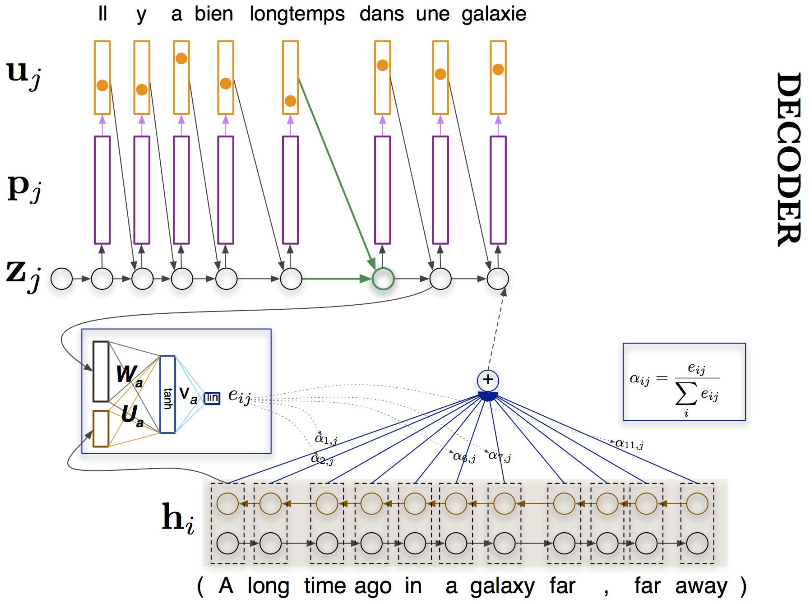
- And so on and so forth...



NMT principle

- At each timestep, a new set of attention weights is computed
- Annotations don't change
- Hidden state of decoder RNN has changed!
- Decoder is a conditional LM

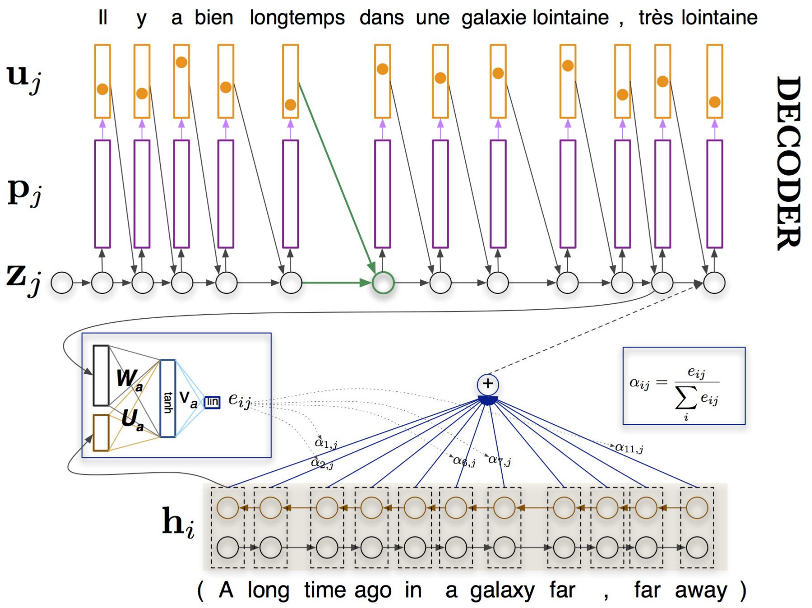
- And so on and so forth...



NMT principle

- At each timestep, a new set of attention weights is computed
- Annotations don't change
- Hidden state of decoder RNN has changed!
- Decoder is a conditional LM

- And so on and so forth...
- ... until end of sequence token is generated.



Multimodal Neural Machine Translation

Multimodal Machine Translation

- 2 modalities: text and images
- Context: Multimodal MT challenge at WMT (3rd edition this year)
- Data: Multi30k

Descriptions

- 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.

MMT: research questions

- How to represent both modalities? Which architecture?
- How/where to integrate them in the model?
- Can we create visually grounded representations?
- Can we improve the MT system performance with images?

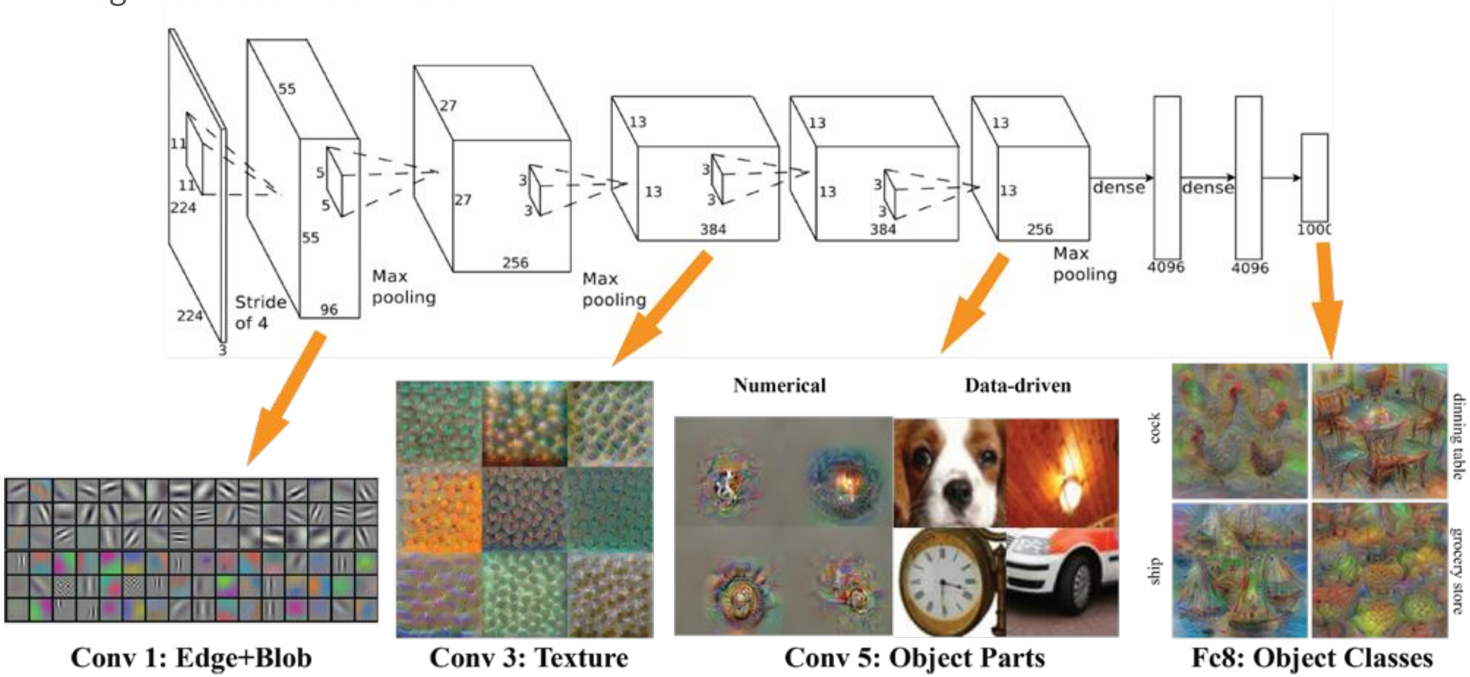
Representing textual input

- See NMT
- RNN
  - bidirectional RNN
  - Can use several layers : more abstract representation?
  - Last state: fixed-size vector representation
  - All states: matrix representation
- Convolutional Networks, etc.

➢ “General purpose sentence representation learning” project during JSALT(?)

Representing an image: quick look to CNNs

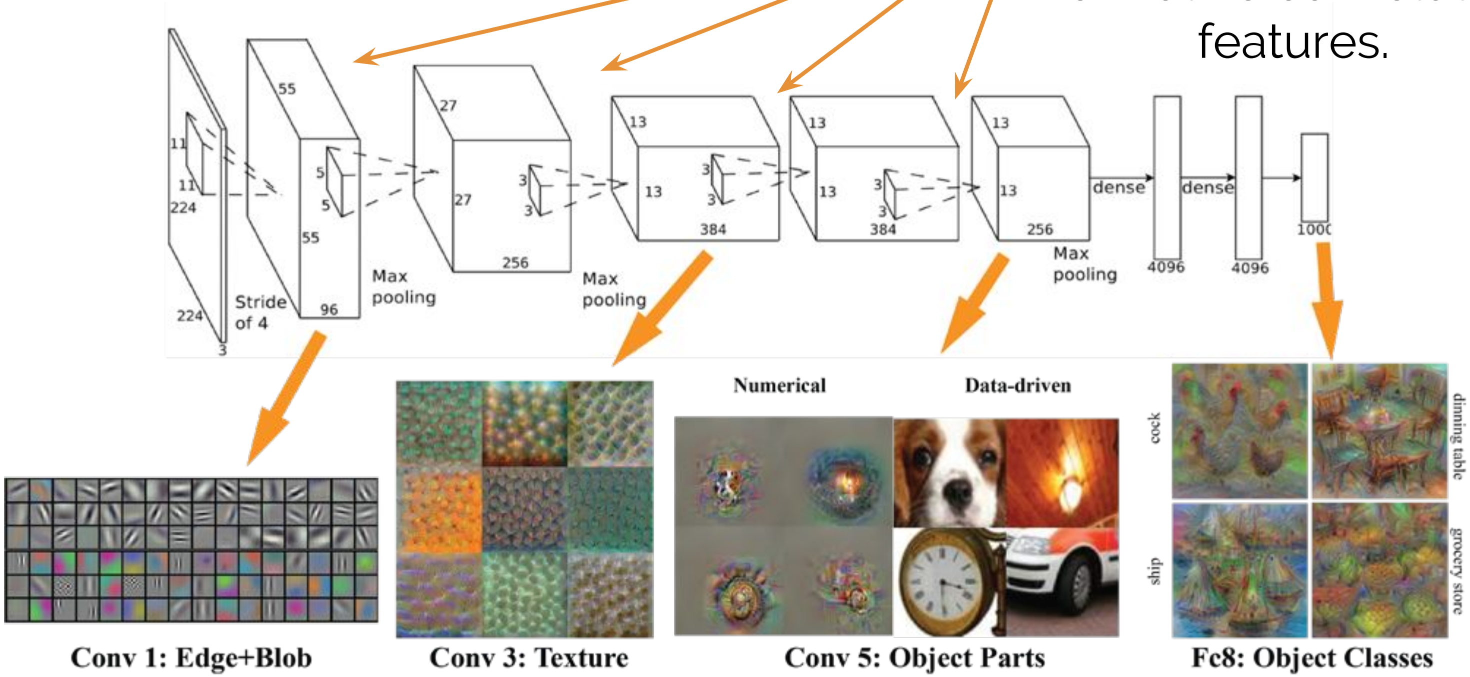
- ImageNet classification task



A visualization of AlexNet architecture: http://vision03.csail.mit.edu/cnn_art/index.html

Representing an image: quick look to CNNs

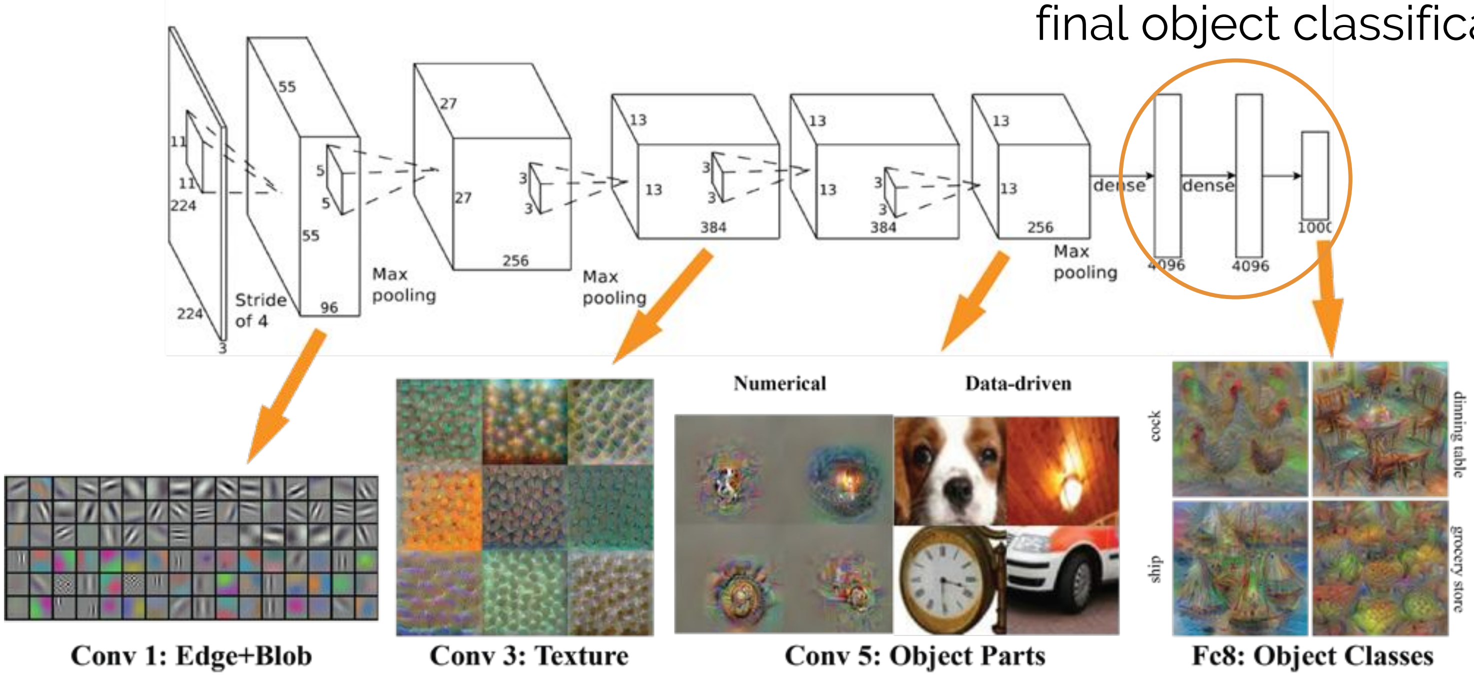
Fine grained, spatially informative convolutional features.



A visualization of AlexNet architecture: http://vision03.csail.mit.edu/cnn_art/index.html

Representing an image: quick look to CNNs

Fixed-size global features guided more towards the final object classification task

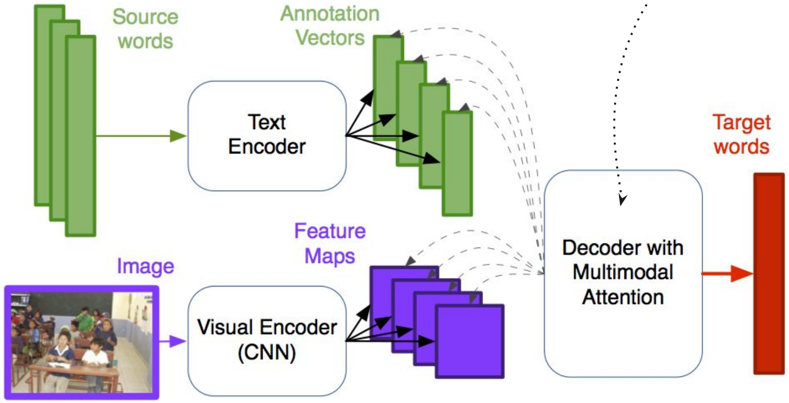


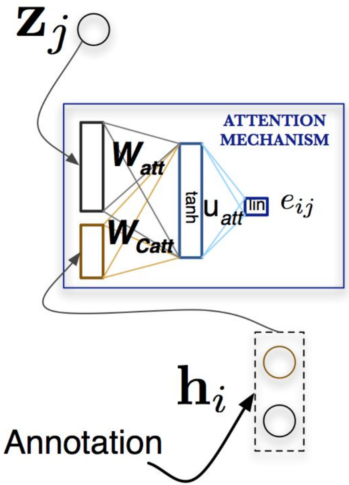
A visualization of AlexNet architecture: http://vision03.csail.mit.edu/cnn_art/index.html

Fusion, multimodal attention

Caglayan et al., 2016a, 2016b

Calixto et al, 2016, Libovický and Helcl, 2017

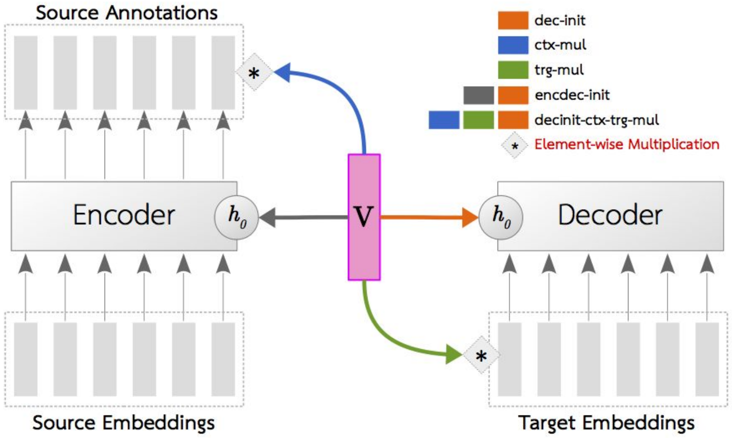




Shared vs. distinct weights for both modalities

Integration of fixed size visual information

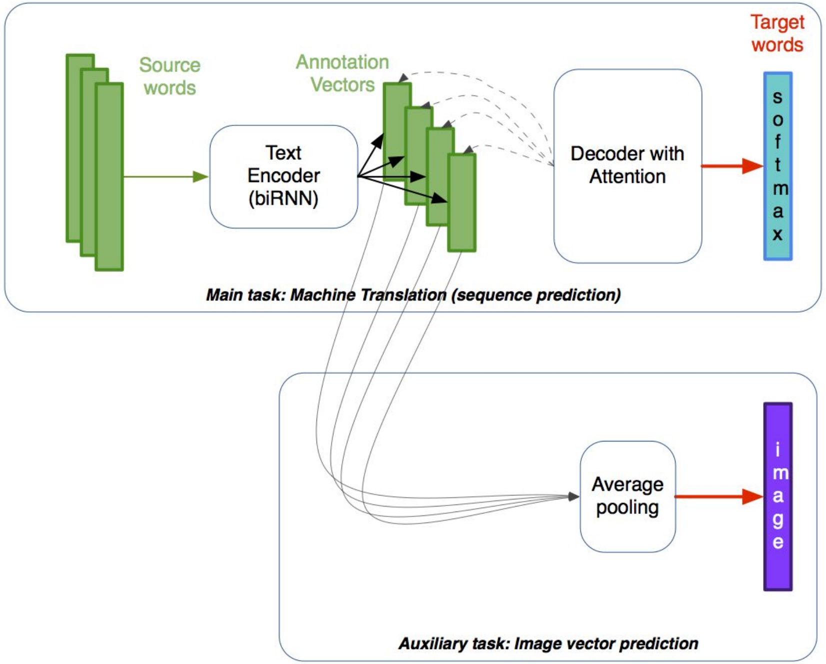
- Prepending and/or appending visual vectors to source sequence
  - Huang et al., 2016
- Decoder initialization
  - Calixto et al., 2016
- Encoder/decoder initialization, multiplicative interaction schemes
  - Caglayan et al., 2017, Delbrouck and Dupont, 2017
- ImageNet class probability vector as a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



- Detailed later

Multitask learning: Imagination

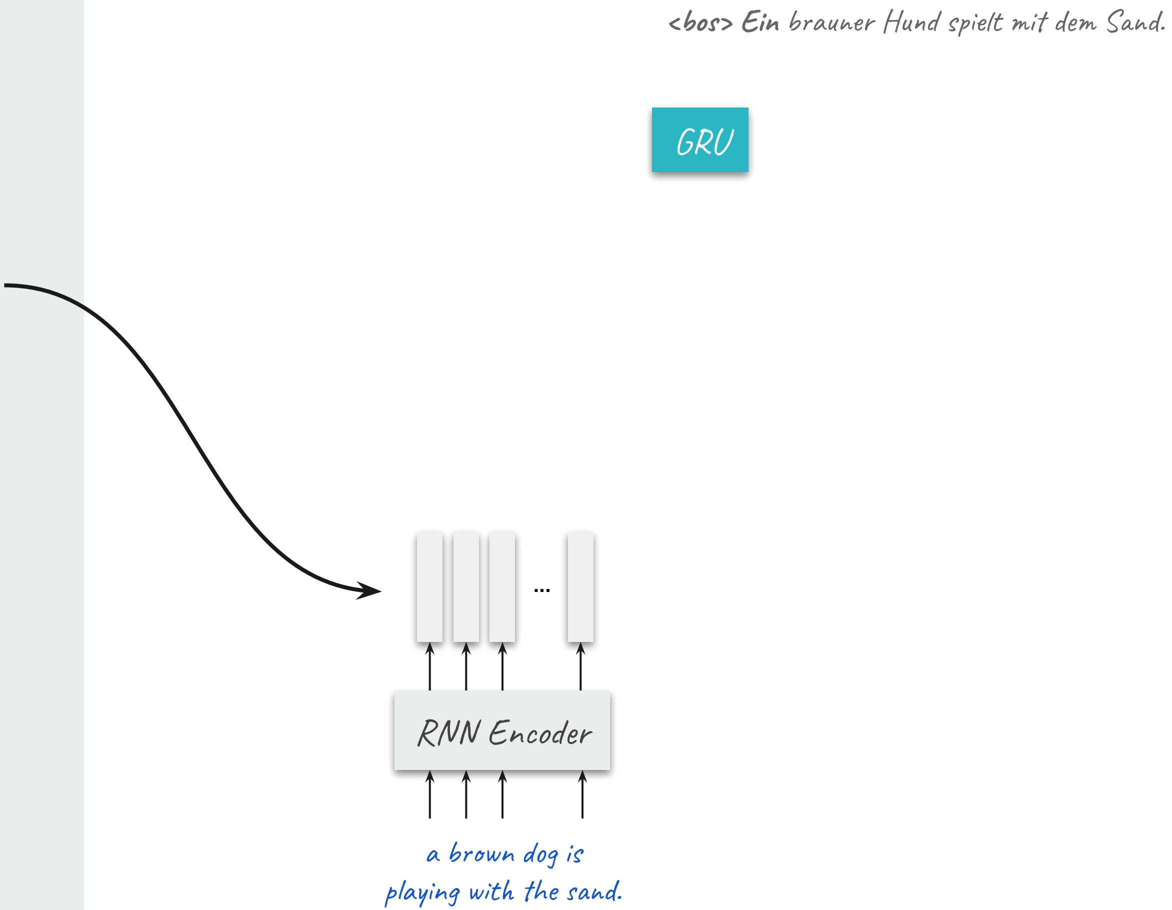
- Elliott, D. and Kádár, A. (2017).
- Predict image vector from source sentence
    - during training only
- ➢ Gradient flow from image vector impact the source text encoder and embeddings.



A More Detailed Look into Multimodal NMT

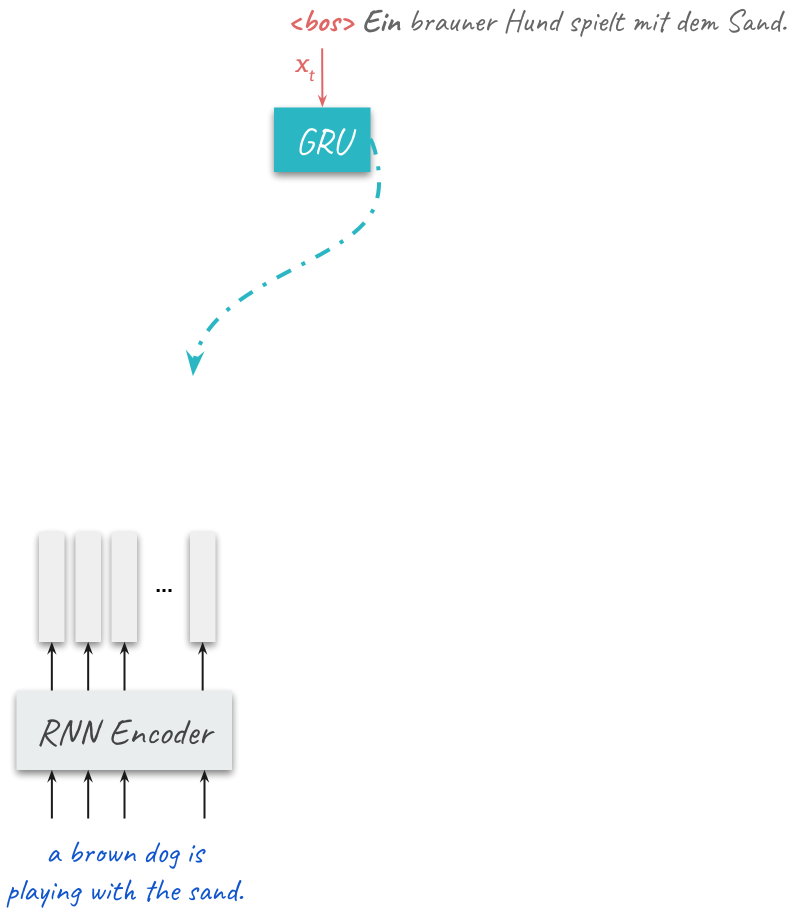
NMT with conditional GRU

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



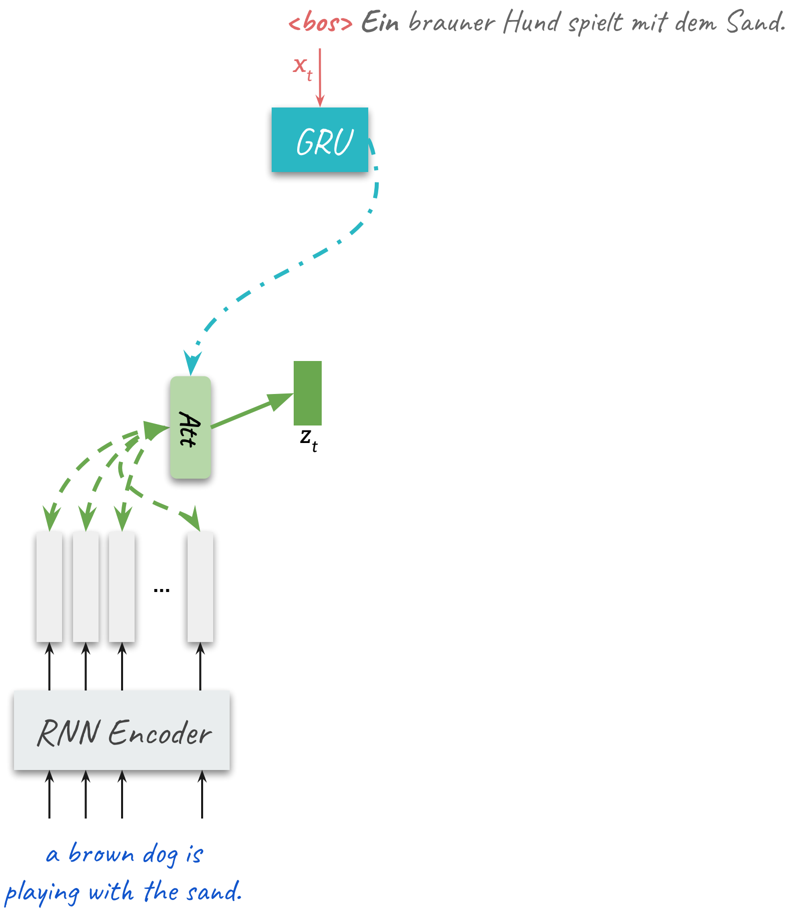
NMT with conditional GRU

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



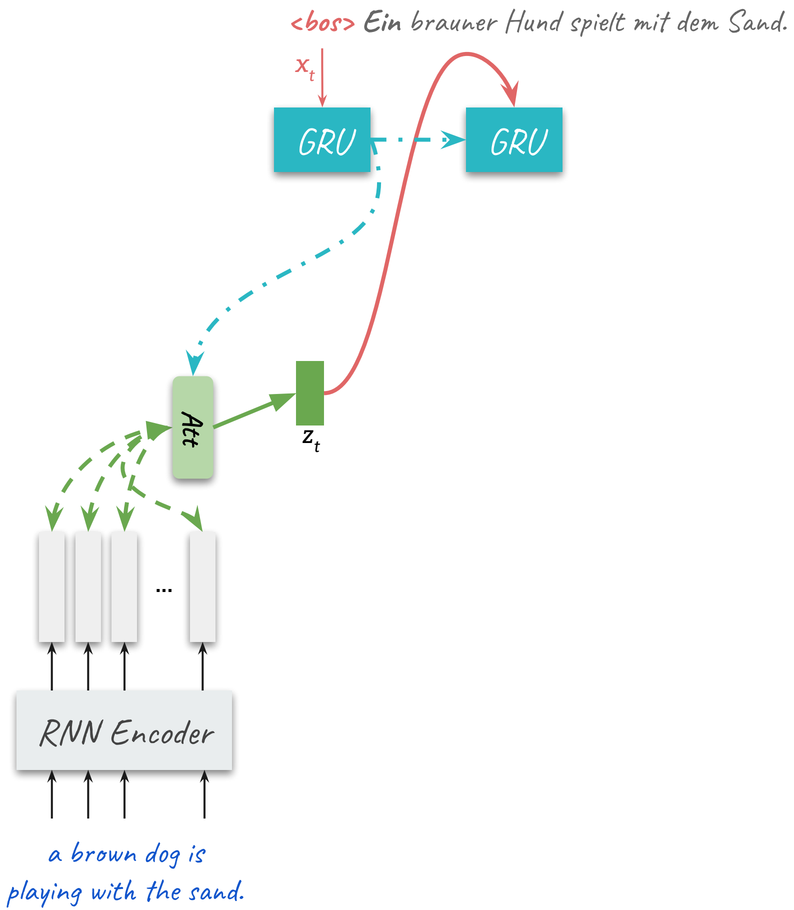
NMT with conditional GRU

- 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.



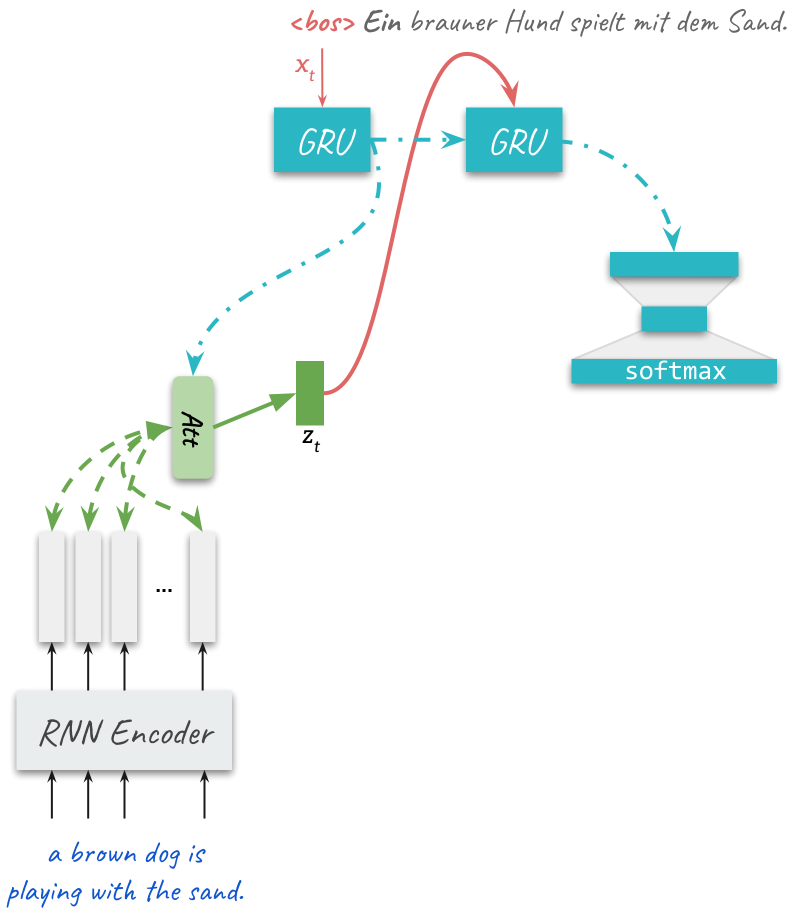
NMT with conditional GRU

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



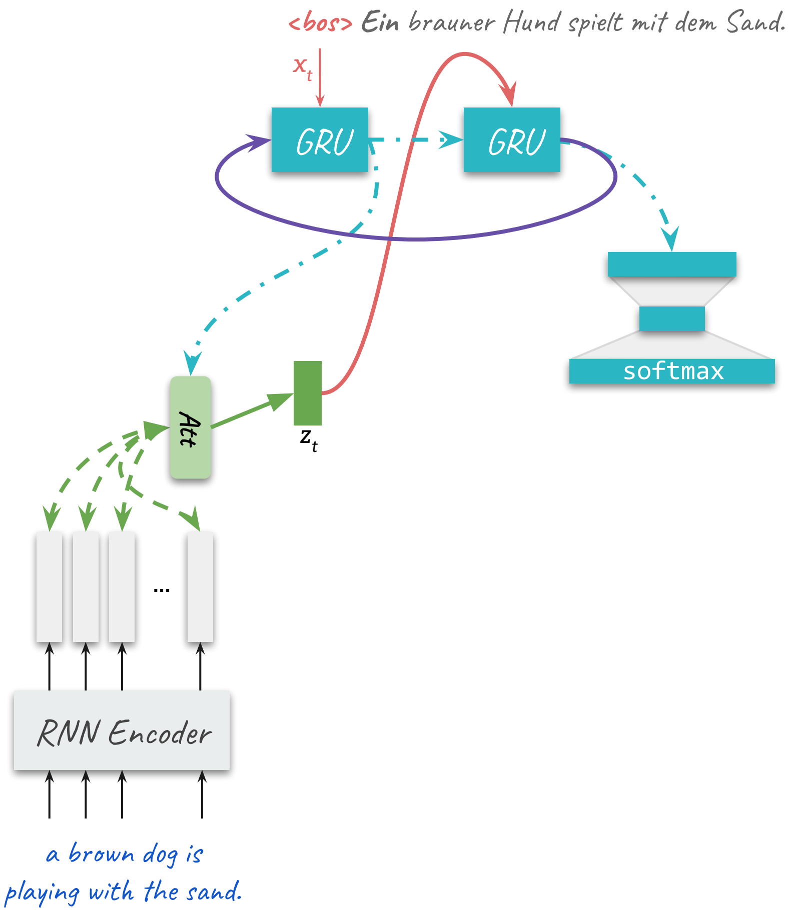
NMT with conditional GRU

- 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()*



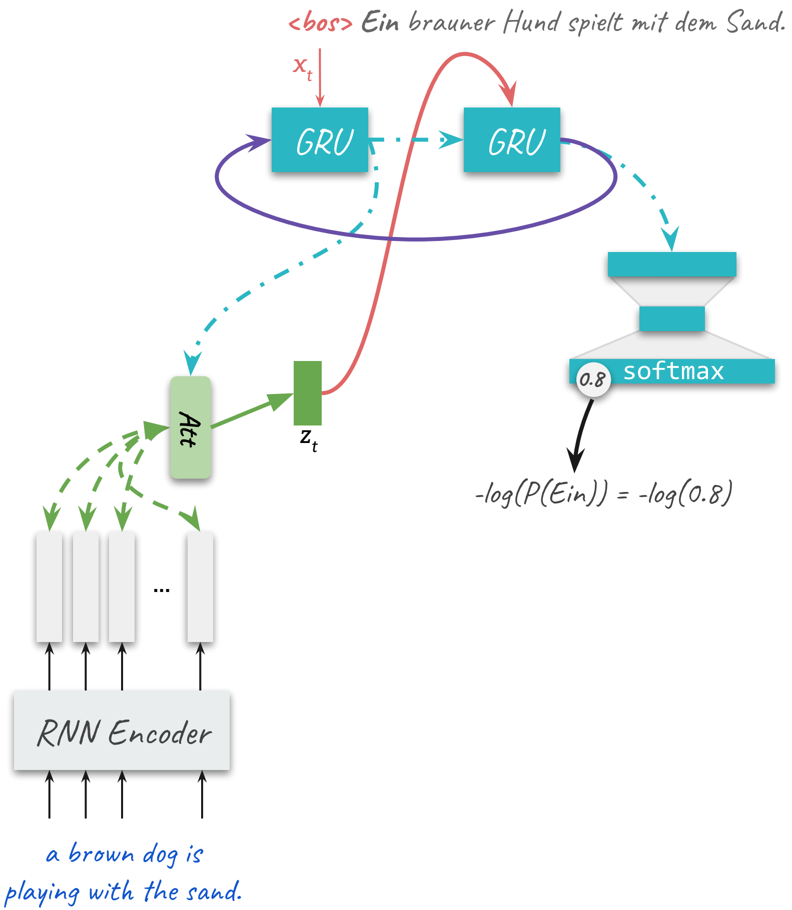
NMT with conditional GRU

- 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.

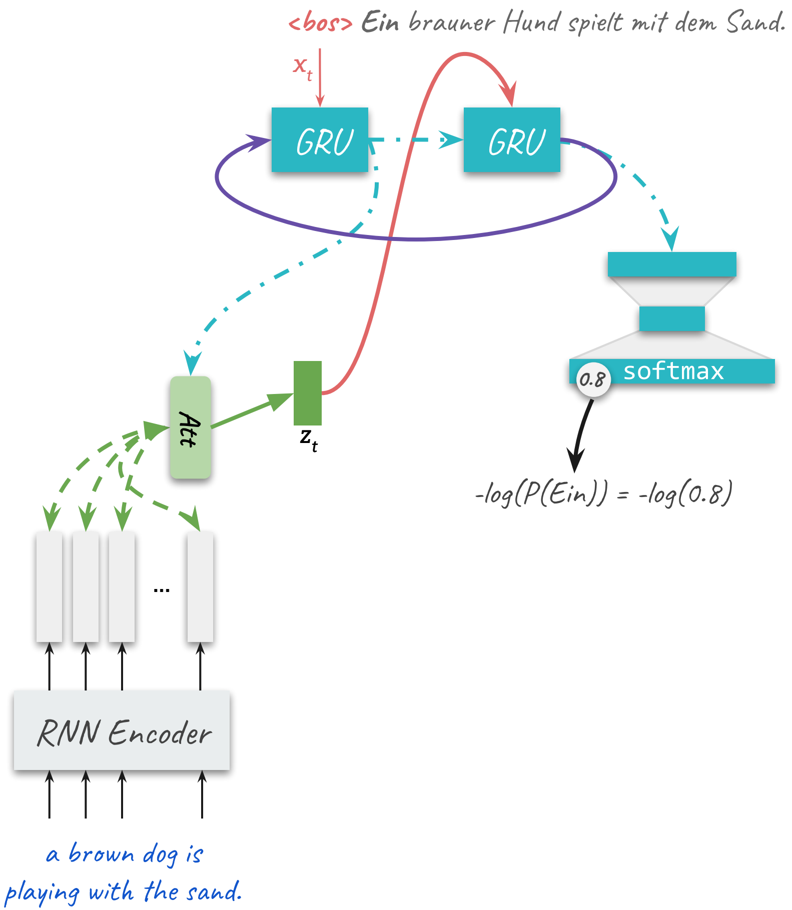


NMT with conditional GRU

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

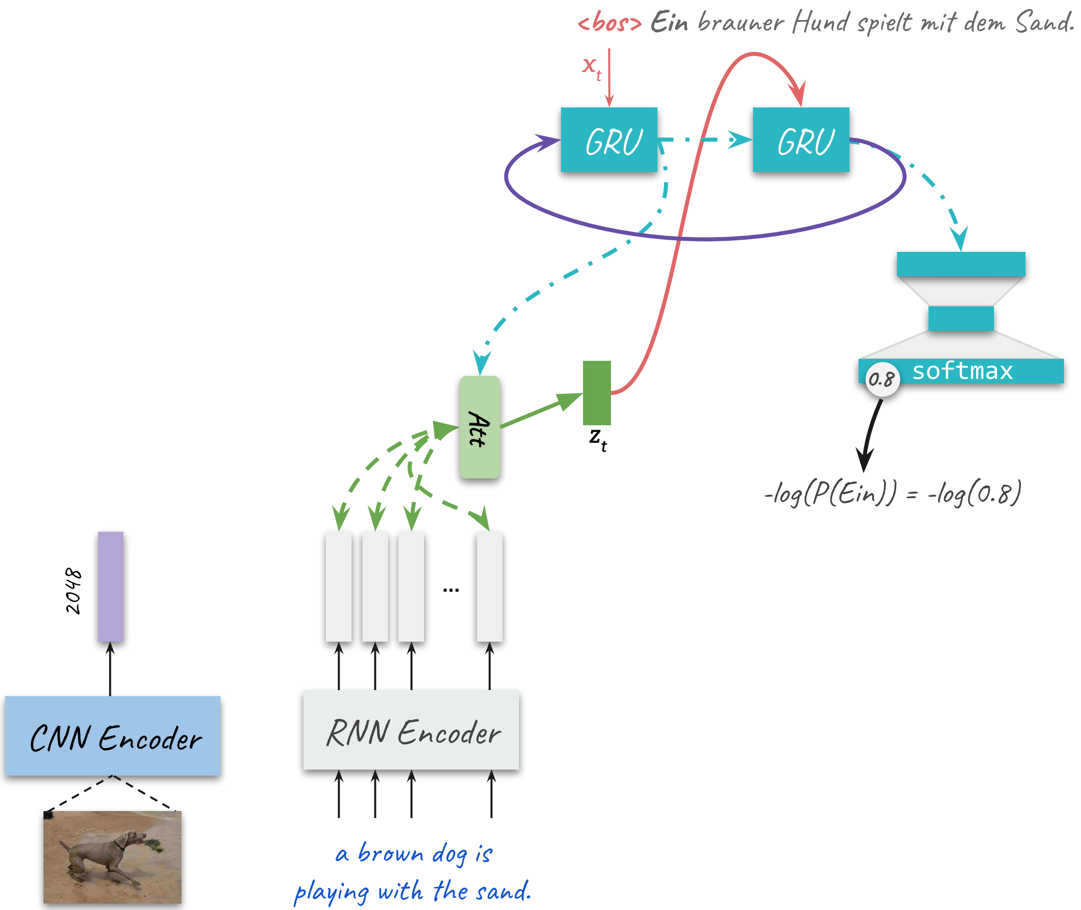


Simple Multimodal NMT



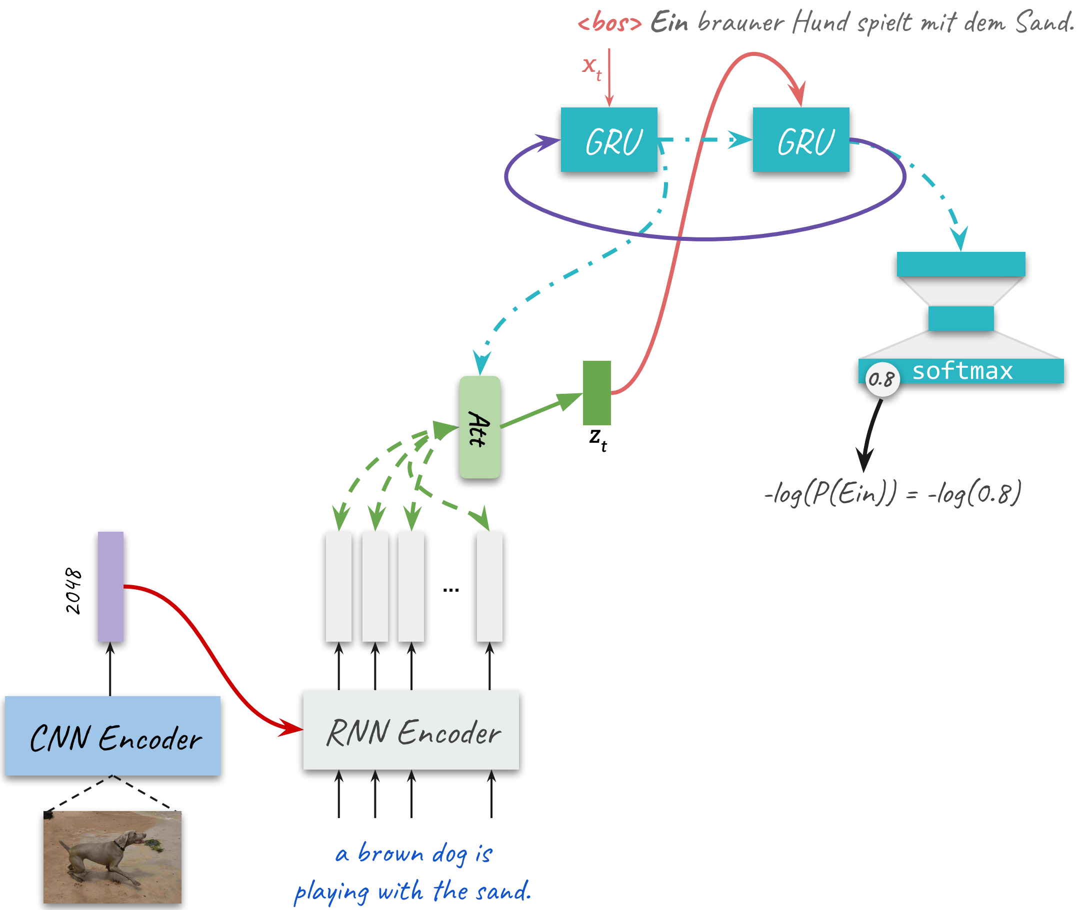
Simple Multimodal NMT

- Here we extract a single global feature vector from some later layers of the CNN.
- This vector will be further used throughout the network to contextualize language representations.



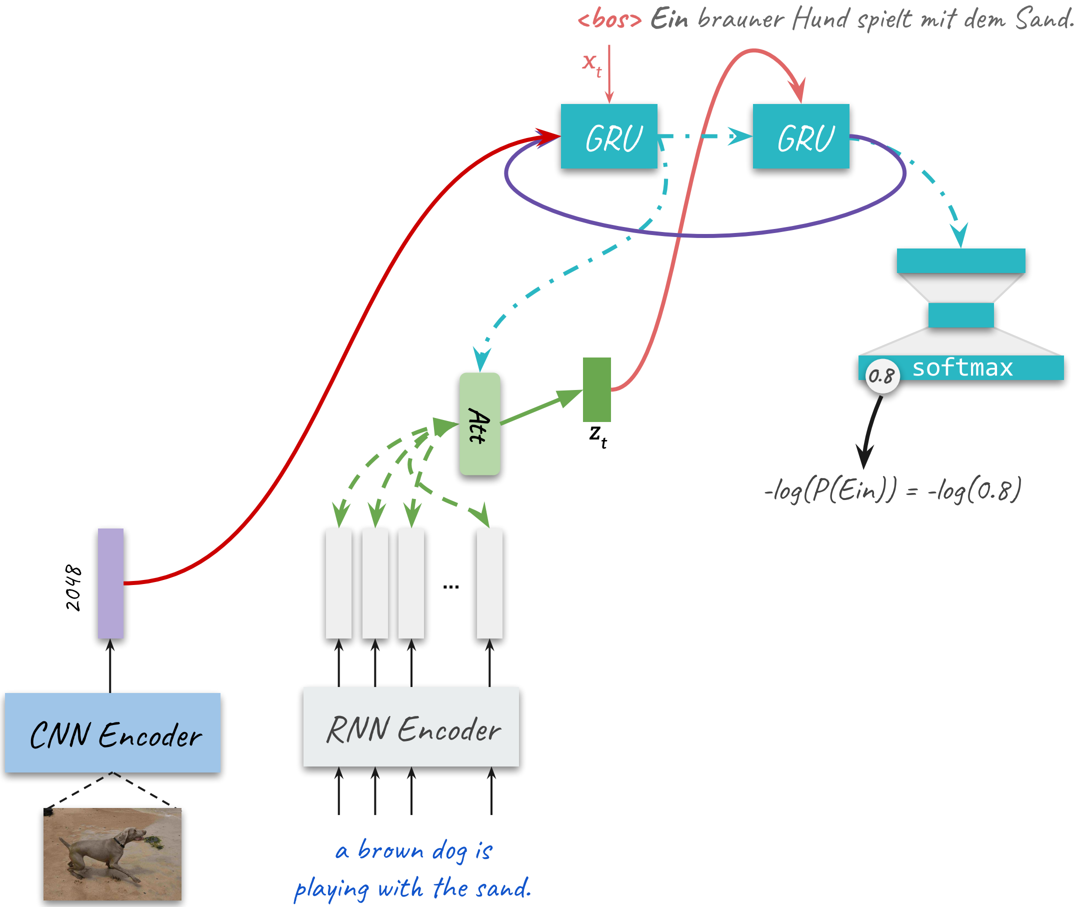
Simple Multimodal NMT

1. Initialize the source sentence encoder.



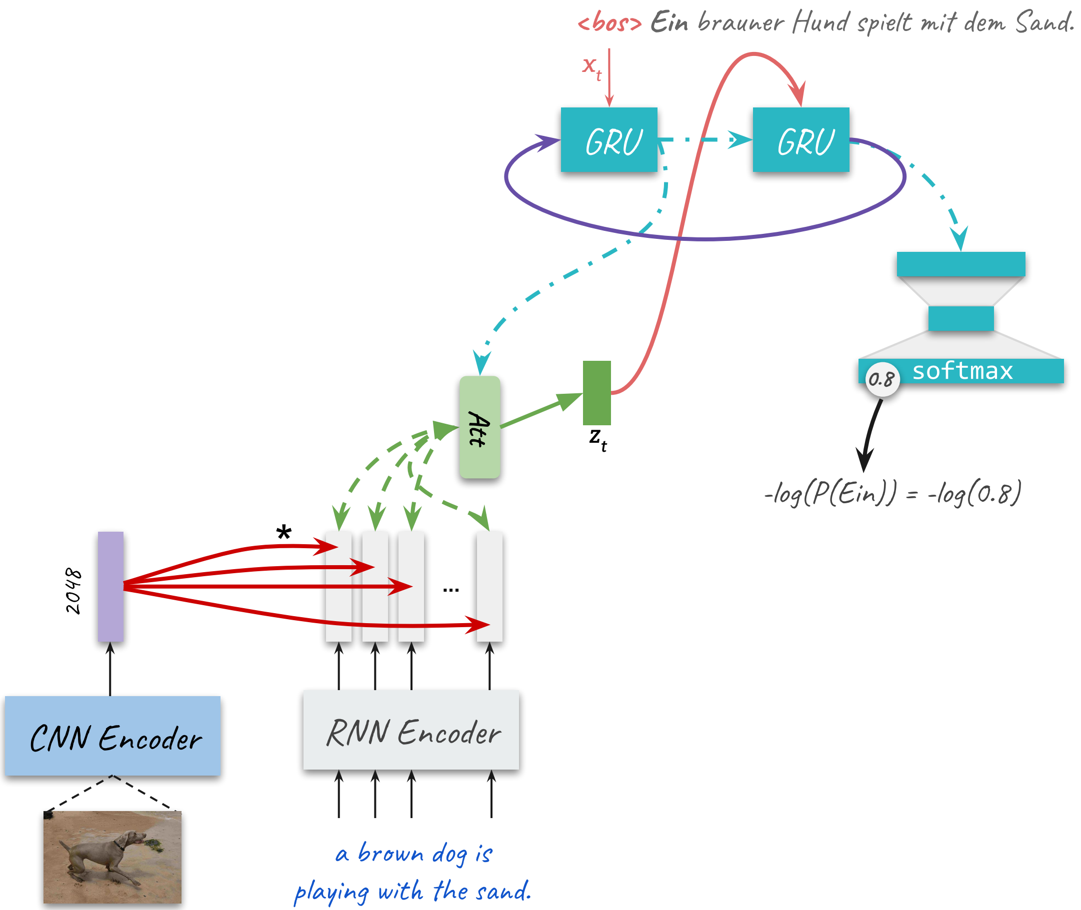
Simple Multimodal NMT

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



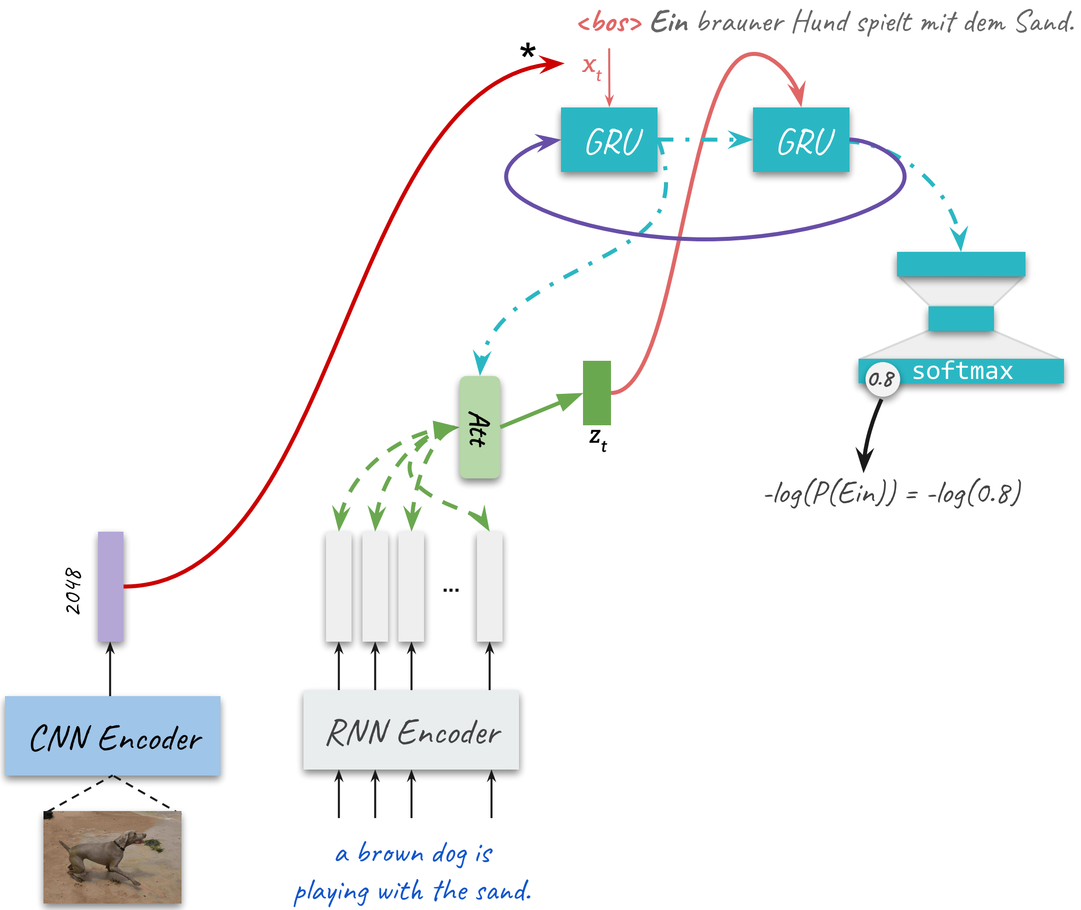
Simple Multimodal NMT

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



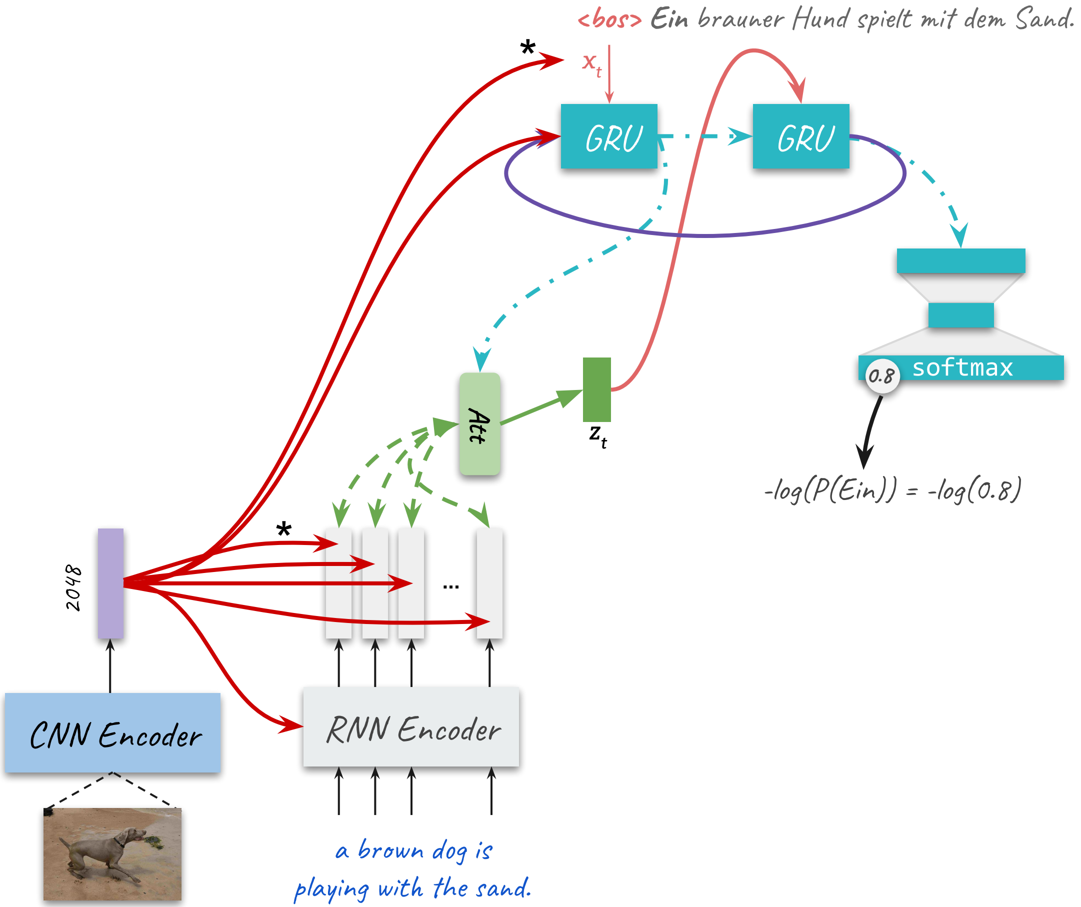
Simple Multimodal NMT

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

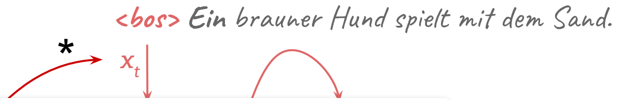
- Initialize the source sentence encoder
- Initialize the decoder
- Element-wise multiplicative interaction with source annotations.
- Element-wise multiplicative interaction with target embeddings.

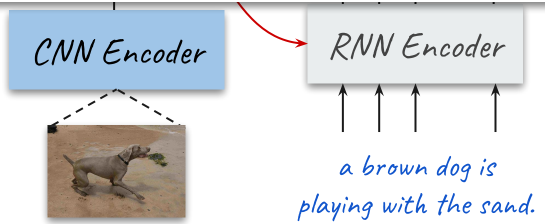


Simple Multimodal NMT

- Initia enco
- Initia
- Elen inter anno
- Elen 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.





Pre-Conclusion

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

Pre-Conclusion

- Prof Ray Mooney (U. Texas), controversial claim:

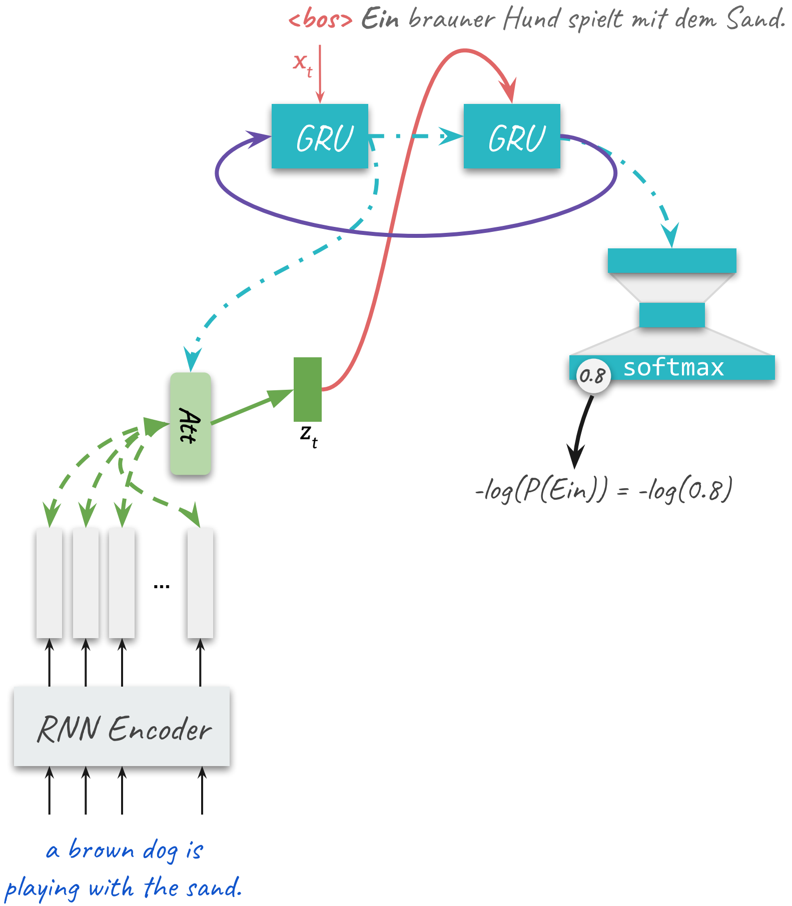
You can't cram the meaning of a whole *$#*! sentence into a single *$#*! vector!

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

- From **coarse** to **fine** visual information
- **Parsimony**:
    - use only **relevant parts** of the image, **when needed**
    - e.g. objects related to the input words
    - cf. Karpathy and Fei-Fei, 2015

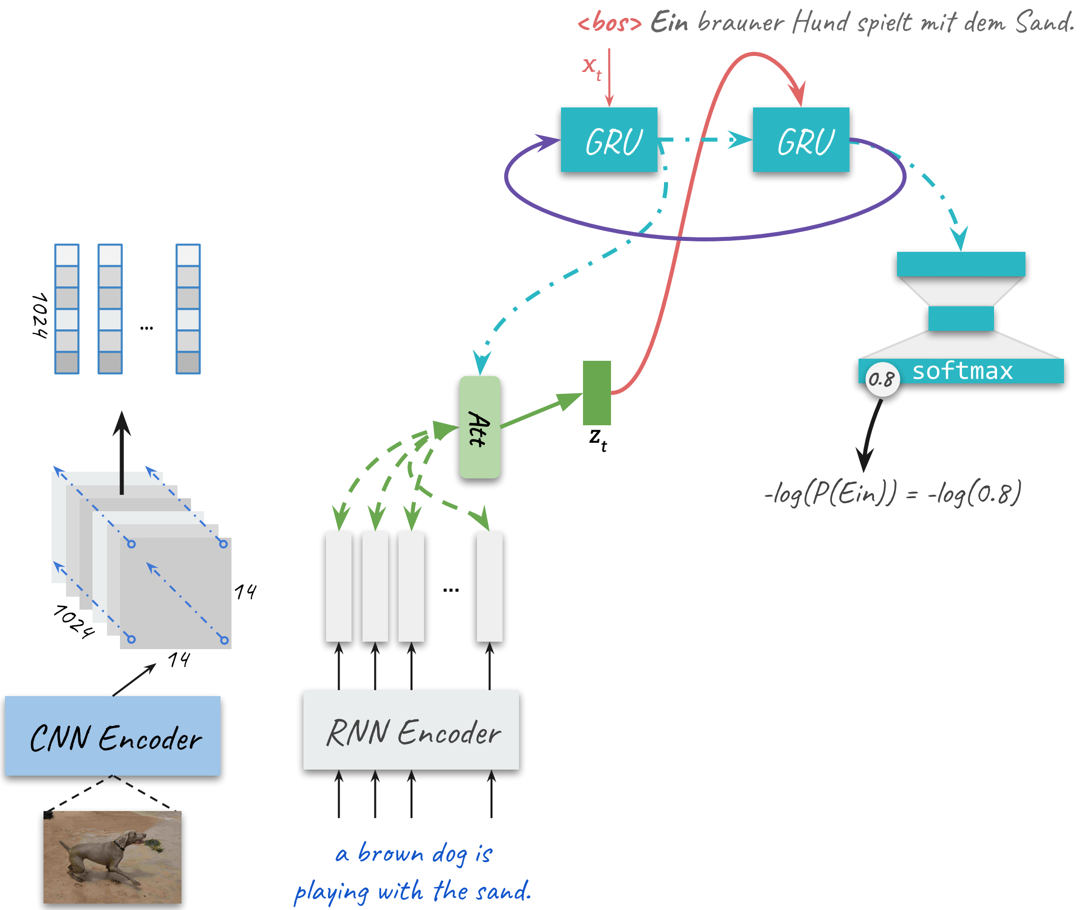
Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

Attentive Multimodal NMT



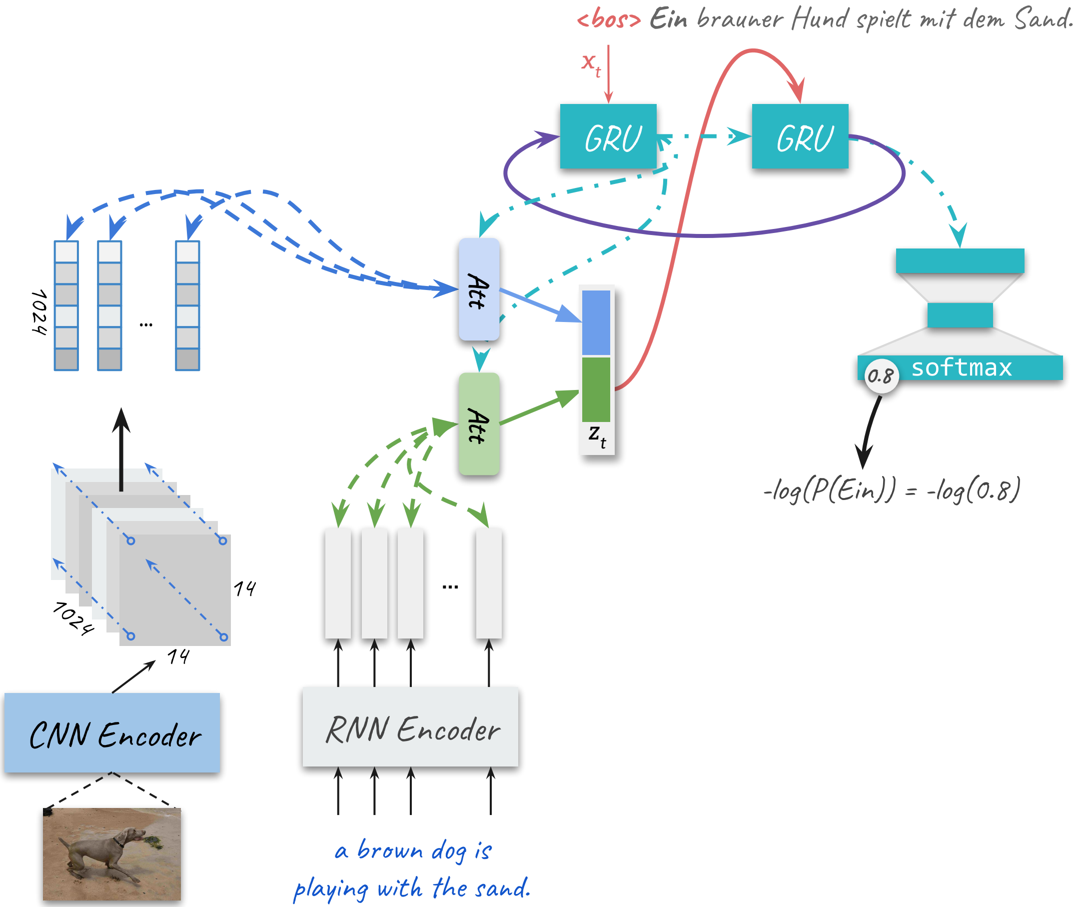
Attentive Multimodal NMT

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



Attentive Multimodal NMT

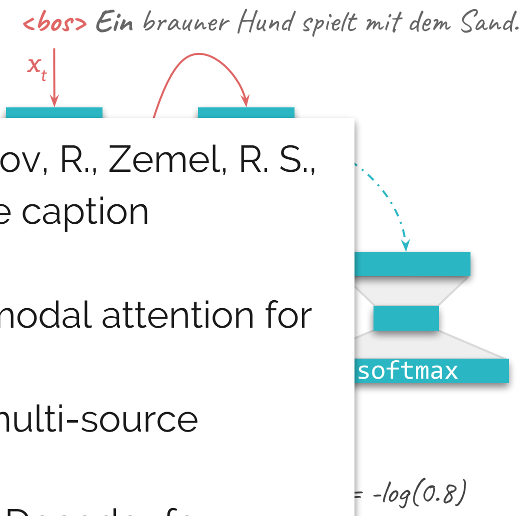
- Use a CNN to extract **convolutional features** from the image
  - preserve spatial correspondence with the input image
- A new attention block for the visual annotations
- z_t becomes the concatenation of both contexts.

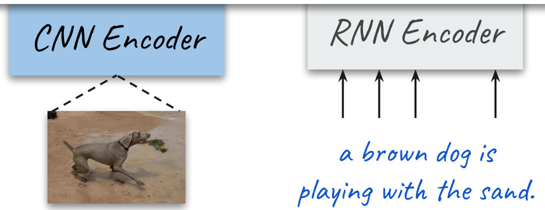


Attentive Multimodal NMT

- Use
  conv
  the i
  -
- A ne
  visua
- z_t be
  conc
  contexts.

- 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.





Some Results

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	$\underline{51.9} \pm 0.4$ / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6	31.1 ± 0.7 / 33.5	$\underline{51.9} \pm 0.2$ / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**	30.7 ± 1.0 / 33.4	$\underline{52.2} \pm 0.4$ / **54.0**

Some Results

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

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	$\underline{51.9} \pm 0.4$ / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6	31.1 ± 0.7 / 33.5	$\underline{51.9} \pm 0.2$ / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**	30.7 ± 1.0 / 33.4	$\underline{52.2} \pm 0.4$ / **54.0**

Some Results

Simple MNMT variants

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	51.9 ± 0.4 / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	57.8 ± 0.5 / 59.6	31.1 ± 0.7 / 33.5	51.9 ± 0.2 / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	57.7 ± 0.5 / **60.4**	30.7 ± 1.0 / 33.4	52.2 ± 0.4 / **54.0**

Some Results

Multiplicative interaction with target embeddings

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	$\underline{51.9} \pm 0.4$ / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6	31.1 ± 0.7 / 33.5	$\underline{51.9} \pm 0.2$ / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**	30.7 ± 1.0 / 33.4	$\underline{52.2} \pm 0.4$ / **54.0**

Some Results

Huge models are overfitting and slow.
Small dimensionalities are better for small datasets. (no need for a strong regularization)

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	$\underline{51.9} \pm 0.4$ / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6	31.1 ± 0.7 / 33.5	$\underline{51.9} \pm 0.2$ / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**	30.7 ± 1.0 / 33.4	$\underline{52.2} \pm 0.4$ / **54.0**

Some Results

Models are early-stopped w.r.t METEOR

BLEU seems more unstable than METEOR

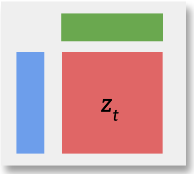
Best METEOR does not guarantee best BLEU

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	Test2016 ($\mu \pm \sigma$/Ensemble) METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	Test2017 ($\mu \pm \sigma$/Ensemble) 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	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	$\underline{51.9} \pm 0.4$ / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	$\underline{57.8} \pm 0.5$ / 59.6	31.1 ± 0.7 / 33.5	$\underline{51.9} \pm 0.2$ / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	$\underline{57.7} \pm 0.5$ / **60.4**	30.7 ± 1.0 / 33.4	$\underline{52.2} \pm 0.4$ / **54.0**

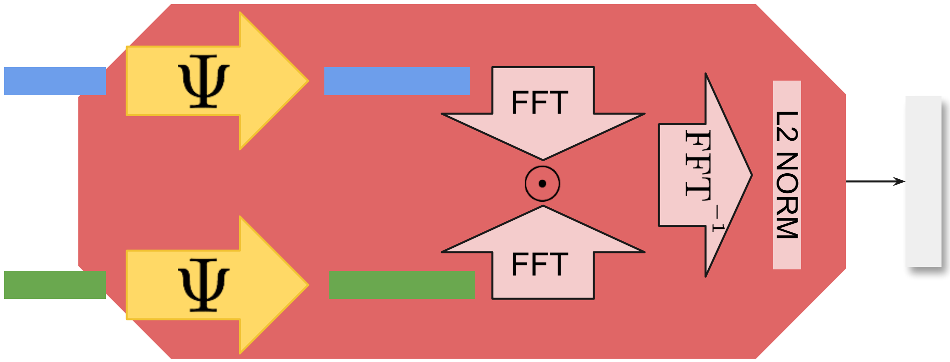
(Compact) Bilinear Pooling

- (Compact) Bilinear Pooling
- Used in VQA: Fukui et al, 2016, slides: http://visualqa.org/static/slides/vqa_final.pdf
- In MT: Delbrouck et al, 2017

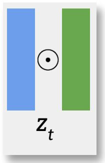
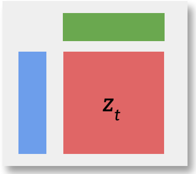
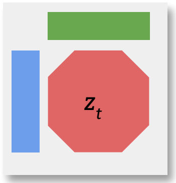
Bilinear Pooling
(outer product)



Compact Bilinear Pooling



Integrating textual and visual features

	Concat 	Element-wise multiplication 	Bilinear Pooling (outer product) 	Compact Bilinear Pooling 
All elements can interact	YES	NO	YES	YES
Multiplicative interaction	NO	YES	YES	YES
Low #activations & computation	YES	YES	NO	YES
Low #parameters	YES	YES	NO	YES

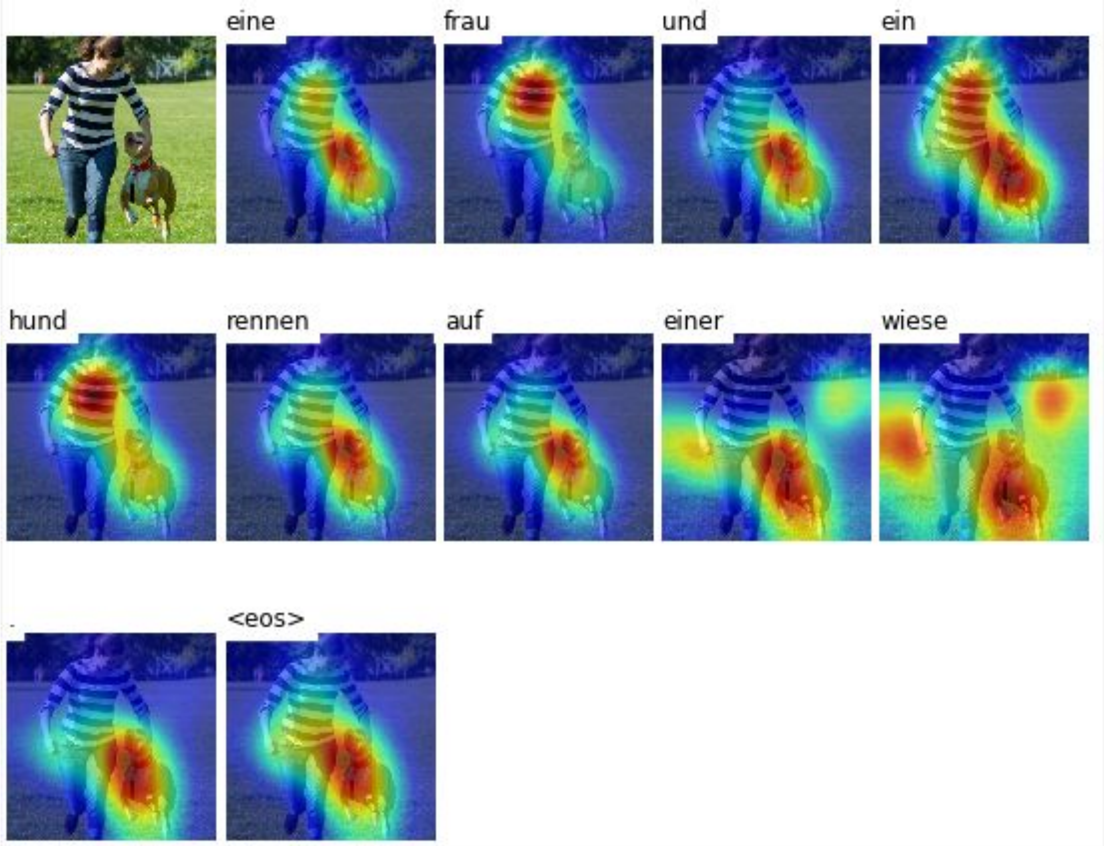
Grounding?

Attention mechanism

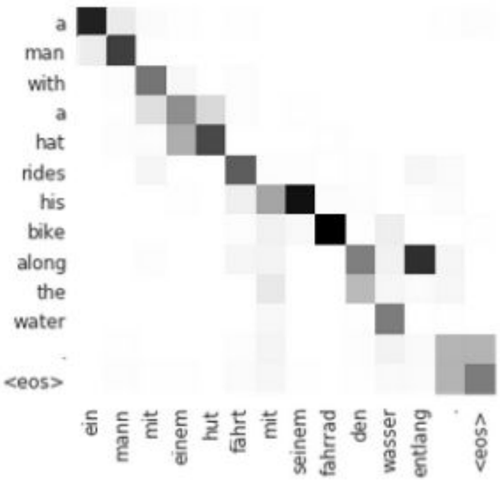
- Attention weights can be used as **link** between modalities

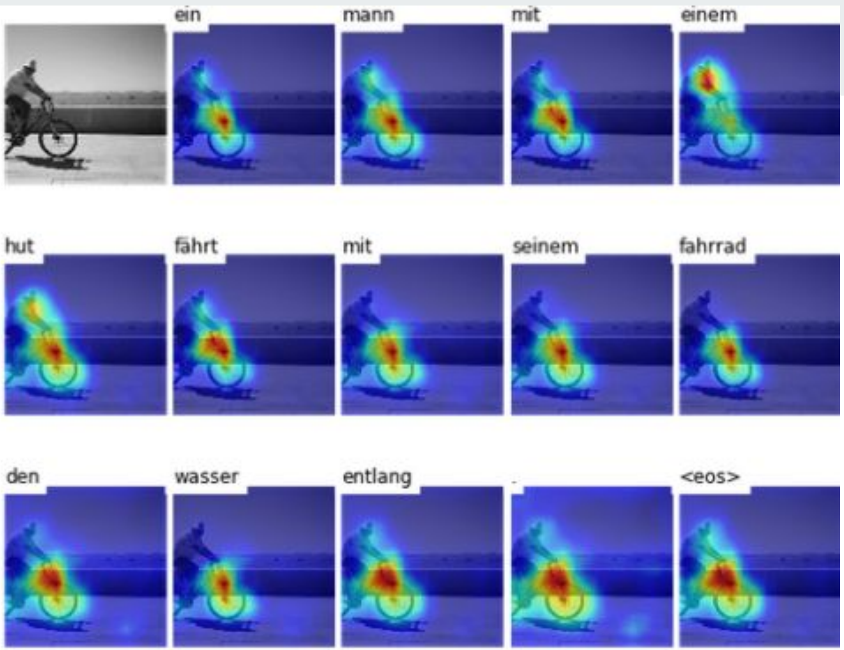
Attentive Multimodal NMT

- Attention over spatial regions while translating from English → German



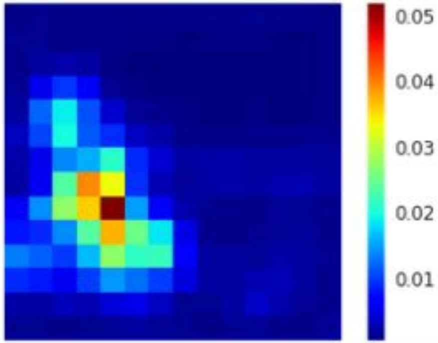
Textual Attention





Sequential spatial attention

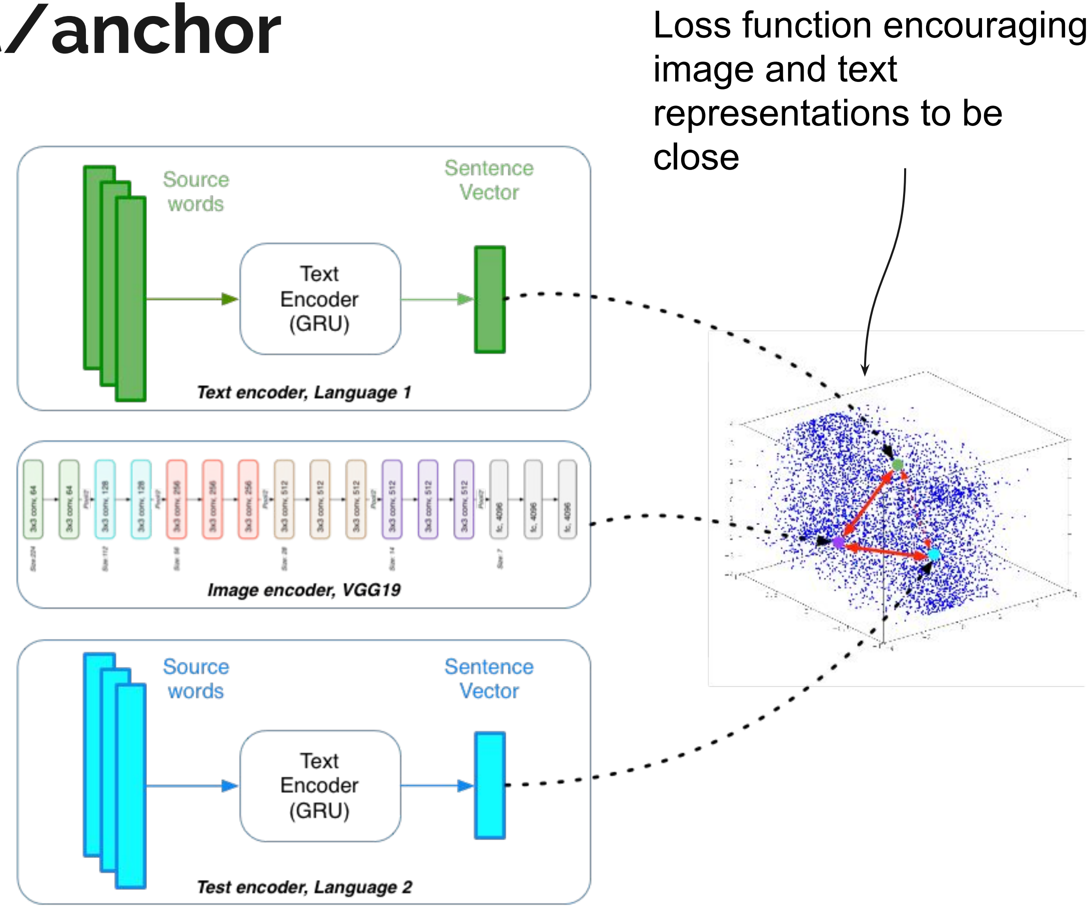
Average spatial attention



Use image as pivot/anchor

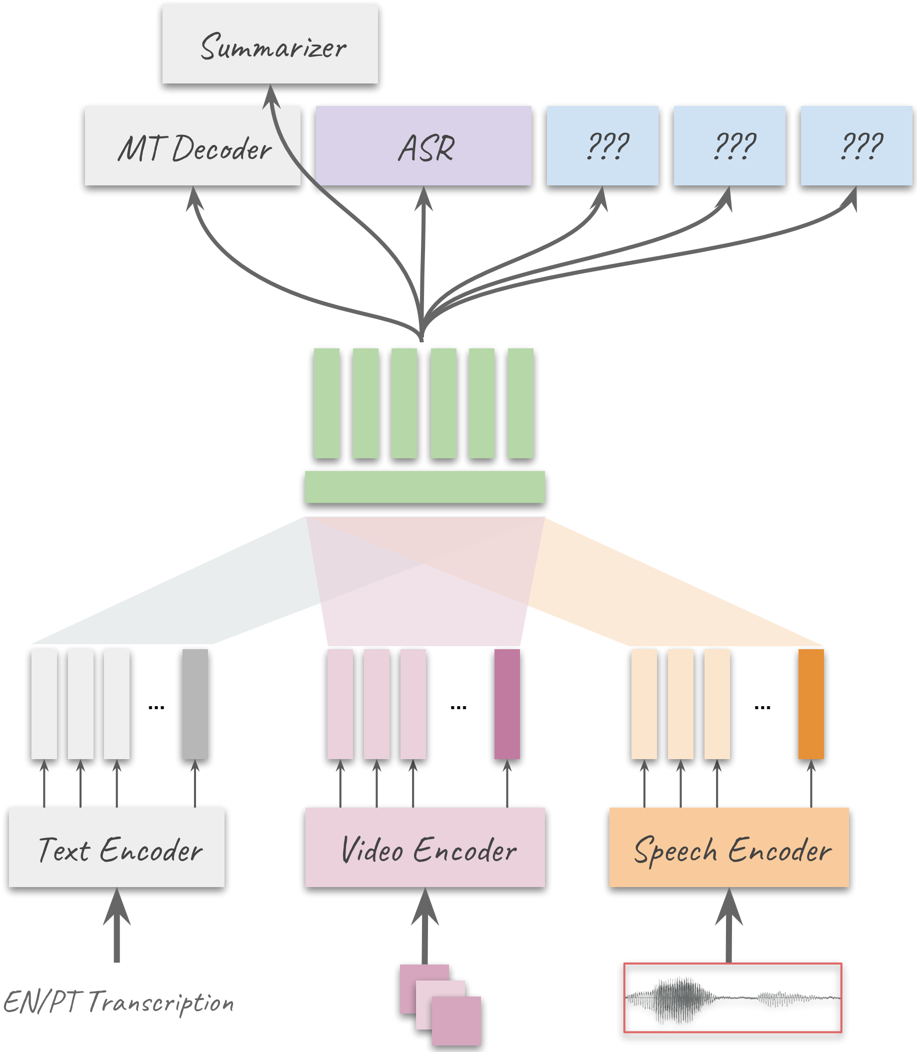
- Visually grounded representation
- Used for
  - Image retrieval
  - description retrieval
- Gella et al, 2017

- Not used for MT… yet(?)



Multitasking for JSALT

- Jointly optimize auxiliary **tasks** along with the NMT.



Conclusion

- Various ways of integrating textual and visual features
- Results in terms of BLEU are only slightly impacted
  - Multi30k has some biases
  - Not all sentences need visual information to produce a good translation
- Multi-task systems are promising

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.