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

1. 1-hot encoding + projection + update \textbf{forward} RNN hidden state
Neural Machine Translation

1bis. update backward RNN hidden state
2. **Annotation** = concat **forward** and **backward** vectors

Every $h_i$ encodes the whole sentence with a focus on the i-th word
2. Decoder gets the annotations.
NMT principle

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

\[ \tilde{h}_j = \sum_i \alpha_{ij} h_i \]
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

\[ \tilde{h}_j = \sum_i \alpha_{ij} h_i \]
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
NMT principle

2. Decoder gets the annotations.
3. **Attention weights** are computed
   a. Feed forward NN
   b. Weighted mean
   \[ \tilde{h}_j = \sum_i \alpha_{ij} h_i \]
4. Update hidden state of GRU
5. Probability distribution for all words
6. Generate next word
   a. Most probable of beam
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 skakající 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](http://vision03.csail.mit.edu/cnn_art/index.html)
Representing an image: quick look to CNNs

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

- 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.

\(<\text{bos}>\) Ein brauner Hund spielt mit dem Sand.
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 $\text{softmax}()$
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

-log(P(Ein)) = -log(0.8)

<eos> Ein brauner Hund spielt mit dem Sand.

a brown dog is playing with the sand.
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.
1. Initialize the source sentence encoder.

- \(-\log(P(Ein)) = -\log(0.8)\)
Simple Multimodal NMT

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

\[ -\log(P(Ein)) = -\log(0.8) \]
Simple Multimodal NMT

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

- \( \text{softmax} \cdot \log(P(\text{Ein})) = -\log(0.8) \)
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.

\[ \text{-log}(P(Ein)) = -\log(0.8) \]
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.

\[
\text{-log}(P(\text{Ein})) = -\log(0.8)
\]
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


< bos > Ein brauner Hund spielt mit dem Sand.
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
Attentive Multimodal NMT

- A brown dog is playing with the sand.

- $<\text{bos}>$ Ein brauner Hund spielt mit dem Sand.

- $\log(-P(Ein)) = -\log(0.8)$
Attentive Multimodal NMT

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

\[
\log(\text{softmax}(\text{Ein})) = \log(0.8)
\]

- A brown dog is playing with the sand.
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 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.

- 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.
Some Results

<table>
<thead>
<tr>
<th>En→De Flickr</th>
<th># Params</th>
<th>Test2016 (μ ± σ/Ensemble)</th>
<th>Test2017 (μ ± σ/Ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Caglayan et al. (2016a)</td>
<td>62.0M</td>
<td>29.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Huang et al. (2016)</td>
<td>-</td>
<td>36.5</td>
<td>54.1</td>
</tr>
<tr>
<td>Calixto et al. (2017a)</td>
<td>213M</td>
<td>36.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Calixto et al. (2017b)</td>
<td>-</td>
<td>37.3</td>
<td>55.1</td>
</tr>
<tr>
<td>Elliott and Kádár (2017)</td>
<td>-</td>
<td>36.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>57.3 ± 0.5 / 59.2</td>
</tr>
<tr>
<td>(D1) fusion-conv</td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>57.0 ± 0.3 / 59.1</td>
</tr>
<tr>
<td>(D2) dec-init-ctx-trg-mul</td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>57.3 ± 0.3 / 59.3</td>
</tr>
<tr>
<td>(D3) dec-init</td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>57.5 ± 0.2 / 59.4</td>
</tr>
<tr>
<td>(D4) encdec-init</td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>57.6 ± 0.3 / 59.5</td>
</tr>
<tr>
<td>(D5) ctx-mul</td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>57.8 ± 0.5 / 59.6</td>
</tr>
<tr>
<td>(D6) trg-mul</td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td><strong>57.7 ± 0.5 / 60.4</strong></td>
</tr>
</tbody>
</table>
Some Results

<table>
<thead>
<tr>
<th>En→De Flickr</th>
<th># Params</th>
<th>Test2016 (μ ± σ/Ensemble)</th>
<th>Test2017 (μ ± σ/Ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Caglayan et al. (2016a)</td>
<td>62.0M</td>
<td>29.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Huang et al. (2016)</td>
<td>-</td>
<td>36.5</td>
<td>54.1</td>
</tr>
<tr>
<td>Calixto et al. (2017a)</td>
<td>213M</td>
<td>36.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Calixto et al. (2017b)</td>
<td>-</td>
<td>37.3</td>
<td>55.1</td>
</tr>
<tr>
<td>Elliott and Kádár (2017)</td>
<td>-</td>
<td>36.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>57.3 ± 0.5 / 59.2</td>
</tr>
<tr>
<td>(D1) fusion-conv</td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>57.0 ± 0.3 / 59.1</td>
</tr>
<tr>
<td>(D2) dec-init-ctx-trg-mul</td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>57.3 ± 0.3 / 59.3</td>
</tr>
<tr>
<td>(D3) dec-init</td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>57.5 ± 0.2 / 59.4</td>
</tr>
<tr>
<td>(D4) encdec-init</td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>57.6 ± 0.3 / 59.5</td>
</tr>
<tr>
<td>(D5) ctx-mul</td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>57.8 ± 0.5 / 59.6</td>
</tr>
<tr>
<td>(D6) trg-mul</td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td>57.7 ± 0.5 / 60.4</td>
</tr>
</tbody>
</table>

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

Simple MNMT variants

<table>
<thead>
<tr>
<th>En→De Flickr</th>
<th># Params</th>
<th>Test2016 ($\mu \pm \sigma$/Ensemble)</th>
<th>Test2017 ($\mu \pm \sigma$/Ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Caglayan et al. (2016a)</td>
<td>62.0M</td>
<td>29.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Huang et al. (2016)</td>
<td>-</td>
<td>36.5</td>
<td>54.1</td>
</tr>
<tr>
<td>Calixto et al. (2017a)</td>
<td>213M</td>
<td>36.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Calixto et al. (2017b)</td>
<td>-</td>
<td>37.3</td>
<td>55.1</td>
</tr>
<tr>
<td>Elliott and Kádár (2017)</td>
<td>-</td>
<td>36.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>57.3 ± 0.5 / 59.2</td>
</tr>
<tr>
<td>(D1) fusion-conv</td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>57.0 ± 0.3 / 59.1</td>
</tr>
<tr>
<td>(D2) dec-init-ctx-trg-mul</td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>57.3 ± 0.3 / 59.3</td>
</tr>
<tr>
<td>(D3) dec-init</td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>57.5 ± 0.2 / 59.4</td>
</tr>
<tr>
<td>(D4) encdec-init</td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>57.6 ± 0.3 / 59.5</td>
</tr>
<tr>
<td>(D5) ctx-mul</td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>57.8 ± 0.5 / 59.6</td>
</tr>
<tr>
<td>(D6) trg-mul</td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td>57.7 ± 0.5 / 60.4</td>
</tr>
</tbody>
</table>
### Some Results

Multiplicative interaction with target embeddings

<table>
<thead>
<tr>
<th>Model</th>
<th>Params</th>
<th>BLEU (μ ± σ/Ensemble)</th>
<th>METEOR</th>
<th>BLEU (μ ± σ/Ensemble)</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline NMT</td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>57.3 ± 0.5 / 59.2</td>
<td>30.8 ± 1.0 / 33.2</td>
<td>51.6 ± 0.5 / 53.8</td>
</tr>
<tr>
<td>(D1) fusion-conv</td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>57.0 ± 0.3 / 59.1</td>
<td>29.8 ± 0.9 / 32.7</td>
<td>51.2 ± 0.3 / 53.4</td>
</tr>
<tr>
<td>(D2) dec-init-ctx-trg-mul</td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>57.3 ± 0.3 / 59.3</td>
<td>30.9 ± 1.0 / 33.2</td>
<td>51.4 ± 0.3 / 53.7</td>
</tr>
<tr>
<td>(D3) dec-init</td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>57.5 ± 0.2 / 59.4</td>
<td>31.2 ± 0.7 / 33.4</td>
<td>51.3 ± 0.3 / 53.2</td>
</tr>
<tr>
<td>(D4) encdec-init</td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>57.6 ± 0.3 / 59.5</td>
<td>31.4 ± 0.4 / 33.5</td>
<td>51.9 ± 0.4 / 53.7</td>
</tr>
<tr>
<td>(D5) ctx-mul</td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>57.8 ± 0.5 / 59.6</td>
<td>31.1 ± 0.7 / 33.5</td>
<td>51.9 ± 0.2 / 53.8</td>
</tr>
<tr>
<td>(D6) trg-mul</td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td>57.7 ± 0.5 / 60.4</td>
<td>30.7 ± 1.0 / 33.4</td>
<td>52.2 ± 0.4 / 54.0</td>
</tr>
</tbody>
</table>
Some Results

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

<table>
<thead>
<tr>
<th>En→De Flickr</th>
<th># Params</th>
<th>Test2016 ($\mu \pm \sigma$/Ensemble)</th>
<th>Test2017 ($\mu \pm \sigma$/Ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Caglayan et al. (2016a)</td>
<td>62.0M</td>
<td>29.2</td>
<td>48.5</td>
</tr>
<tr>
<td>Huang et al. (2016)</td>
<td>-</td>
<td>36.5</td>
<td>54.1</td>
</tr>
<tr>
<td>Calixto et al. (2017a)</td>
<td>213M</td>
<td>36.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Calixto et al. (2017b)</td>
<td>-</td>
<td>37.3</td>
<td>55.1</td>
</tr>
<tr>
<td>Elliott and Kádár (2017)</td>
<td>-</td>
<td>36.8</td>
<td>55.8</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>57.3 ± 0.5 / 59.2</td>
</tr>
<tr>
<td>(D1) fusion-conv</td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>57.0 ± 0.3 / 59.1</td>
</tr>
<tr>
<td>(D2) dec-init-ctx-trg-mul</td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>57.3 ± 0.3 / 59.3</td>
</tr>
<tr>
<td>(D3) dec-init</td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>57.5 ± 0.2 / 59.4</td>
</tr>
<tr>
<td>(D4) encdec-init</td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>57.6 ± 0.3 / 59.5</td>
</tr>
<tr>
<td>(D5) ctx-mul</td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>57.8 ± 0.5 / 59.6</td>
</tr>
<tr>
<td>(D6) trg-mul</td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td>57.7 ± 0.5 / 60.4</td>
</tr>
</tbody>
</table>
Some Results

Models are early-stopped w.r.t METEOR

BLEU seems more unstable than METEOR

Best METEOR does not guarantee best BLEU

<table>
<thead>
<tr>
<th>En→De Flickr</th>
<th># Params</th>
<th>Test2016 ($\mu \pm \sigma$/Ensemble)</th>
<th>Test2017 ($\mu \pm \sigma$/Ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU / METEOR</td>
<td>BLEU / METEOR</td>
</tr>
<tr>
<td>Caglayan et al. (2016a)</td>
<td>62.0M</td>
<td>29.2 / 48.5</td>
<td></td>
</tr>
<tr>
<td>Huang et al. (2016)</td>
<td>-</td>
<td>36.5 / 54.1</td>
<td></td>
</tr>
<tr>
<td>Calixto et al. (2017a)</td>
<td>213M</td>
<td>36.5 / 55.0</td>
<td></td>
</tr>
<tr>
<td>Calixto et al. (2017b)</td>
<td>-</td>
<td>37.3 / 55.1</td>
<td></td>
</tr>
<tr>
<td>Elliott and Kádár (2017)</td>
<td>-</td>
<td>36.8 / 55.8</td>
<td></td>
</tr>
<tr>
<td><strong>Baseline NMT</strong></td>
<td>4.6M</td>
<td>38.1 ± 0.8 / 40.7</td>
<td>30.8 ± 1.0 / 33.2</td>
</tr>
<tr>
<td><strong>(D1) fusion-conv</strong></td>
<td>6.0M</td>
<td>37.0 ± 0.8 / 39.9</td>
<td>29.8 ± 0.9 / 32.7</td>
</tr>
<tr>
<td><strong>(D2) dec-init-ctx-trg-mul</strong></td>
<td>6.3M</td>
<td>38.0 ± 0.9 / 40.2</td>
<td>30.9 ± 1.0 / 33.2</td>
</tr>
<tr>
<td><strong>(D3) dec-init</strong></td>
<td>5.0M</td>
<td>38.8 ± 0.5 / 41.2</td>
<td>31.2 ± 0.7 / 33.4</td>
</tr>
<tr>
<td><strong>(D4) encdec-init</strong></td>
<td>5.0M</td>
<td>38.2 ± 0.7 / 40.6</td>
<td>31.4 ± 0.4 / 33.5</td>
</tr>
<tr>
<td><strong>(D5) ctx-mul</strong></td>
<td>4.6M</td>
<td>38.4 ± 0.3 / 40.4</td>
<td>31.1 ± 0.7 / 33.5</td>
</tr>
<tr>
<td><strong>(D6) trg-mul</strong></td>
<td>4.7M</td>
<td>37.8 ± 0.9 / 41.0</td>
<td>30.7 ± 1.0 / 33.4</td>
</tr>
</tbody>
</table>
(Compact) Bilinear Pooling

- (Compact) Bilinear Pooling
- In MT: Delbrouck et al, 2017
## Integrating textual and visual features

<table>
<thead>
<tr>
<th></th>
<th>Concat</th>
<th>Element-wise multiplication</th>
<th>Bilinear Pooling (outer product)</th>
<th>Compact Bilinear Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>All elements can interact</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Multiplicative interaction</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Low #activations &amp; computation</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Low #parameters</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>
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

Average spatial attention

Sequential 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(?)

Loss function encouraging image and text representations to be close
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

References

References