Attention
The problem

- For very long sentences, the score for machine translation really goes down after 30-40 words.

Bahdanau et al. 2014. Neural machine translation by jointly learning to align and translate.
Basic structure of a RNN

- We want to have notion of “time” or “sequence”
Basic structure of a RNN

- We want to have notion of “time” or “sequence”

\[ A_t = \theta_c A_{t-1} + \theta_x x_t \]

Parameters to be learned

Prof. Leal-Taixé and Prof. Niessner
Basic structure of a RNN

- We want to have notion of “time” or “sequence”

\[ A_t = \theta_c A_{t-1} + \theta_x x_t \]

\[ h_t = \theta_h A_t \]

Same parameters for each time step = generalization!
Basic structure of a RNN

- Unrolling RNNs

Hidden state is the same
Basic structure of a RNN

- Unrolling RNNs
Long-term dependencies

I moved to Germany ... so I speak German fluently
ATTENTION: Which hidden states are more important to predict my output?

I moved to Germany ... so I speak German fluently
Attention: intuition

I moved to Germany ...
so I speak German fluently
Attention: architecture

- A decoder processes the information

- Decoders take as input:
  - Previous decoder hidden state
  - Previous output
  - Attention
Attention

- $\alpha_{1,t+1}$ indicates how much the word in the position 1 is important to translate the work in position $t + 1$

- The context aggregates the attention

\[
c_{t+1} = \sum_{k=1}^{t+1} \alpha_{k,t+1} a_k
\]

- **Soft** attention: All attention masks alpha sum up to 1
Computing the attention mask

• We can train a small neural network

\[
\begin{align*}
\alpha_{1, t+1} &= \frac{\exp f_{1, t+1}}{\sum_{k=1}^{t+1} \exp f_{k, t+1}} \\
\end{align*}
\]

Previous state of the decoder

Hidden state of the encoder

Prof. Leal-Taixé and Prof. Niessner
Attention for vision
Why do we need attention?

• We use the whole image to make the classification

• Are all pixels equally important?
Why do we need attention?

- Wouldn’t it be easier and computationally more efficient to just run our classification network on the patch?
Soft attention for captioning
Xu et al. 2015. Show attention and tell: neural image caption generation with visual attention.
Image captioning

• Input: image
• Output: a sentence describing the image.
• **Encoder**: a classification CNN (VGGNet, AlexNet). This computes a feature maps over the image.
• **Decoder**: an attention-based RNN
  – In each time step, the decoder computes an attention map over the entire image, effectively deciding which regions to focus on.
  – It receives a context vector, which is the weighted average of the conv net features.
Conventional captioning

Image from: https://blog.heuritech.com/2016/01/20/attention-mechanism/
Attention mechanism

A girl is throwing a frisbee in the park
Attention mechanism

A girl is throwing a frisbee in the park
Attention mechanism

A girl is throwing a frisbee in the park
Attention mechanism

A girl is throwing a frisbee in the park

Prof. Leal-Taixé and Prof. Niessner
Attention mechanism

$y_i$: Output of encoder are the image features which still retain spatial information (no FC layer!)

$Z_i$: Output of attention model

$h_i$: Hidden state of LSTM
Attention mechanism

How does the attention model look like?
Attention model

• Attention architecture

Any past hidden state

Visual features

Image: https://blog.heuritech.com/2016/01/20/attention-mechanism/
Attention model

• Inputs = feature descriptor for each image patch
Attention model

- Inputs = feature descriptor for each image patch

Still related to the spatial location of the image
Attention model

- We want a bounded output

\[ m_i = \tanh(W_{cm}c + W_{ym}y_i) \]
Attention model

• Softmax to create the attention values between 0 and 1
Attention model

- Multiplied by the image features $\rightarrow$ ranking by importance
Hard attention model

• Choosing one of the features by sampling with probabilities $s_i$
Types of attention

- **Soft attention**: deterministic process that can be backproped.

- **Hard attention**: stochastic process, gradient is estimated through Monte Carlo sampling.

- Soft attention is the most commonly used since it can be incorporated into the optimization more easily.
Types of attention

- Soft vs hard attention
Types of attention: soft

- Can be backproped
- Uses all the image

Prof. Leal-Taixé and Prof. Niessner
Types of attention: hard

- You can view it as an image cropping!

- If we cannot use gradient descent, what alternative could we use to train this function?

Reinforcement Learning
Image captioning with attention

Xu et al. 2015. Show attention and tell: neural image caption generation with visual attention.
Interesting works on attention

• Luong et al, “Effective Approaches to Attention-based Neural Machine Translation,” EMNLP 2015
• Chorowski et al, “Attention-based models for Speech Recognition”, NIPS 2015
• Yao et al, “Describing Videos by Exploiting Temporal Structure”, ICCV 2015
• Chu et al, „Online Multi-Object Tracking Using CNN-based Single Object Tracker with Spatial-Temporal Attention Mechanism“. ICCV 2017
Conditioning
When do we need conditioning?

- Scene understanding from an image and an audio source. Both need to be processed!

"Are you sure there's no mistake?"
When do we need conditioning?

- Visual Question and Answering: the sentence (question) needs to be understood, the image is needed to create the answer.
When do we need conditioning?

• Visual Question and Answering: the sentence (question) needs to be understood, the image is needed to create the answer.
When do we need conditioning?

• We have two sources, can we process one in the context of the other?

• Conditioning: the computation carried out by a model is conditioned or modulated by information extracted from an auxiliary input.

• Note: a similar thing can be obtained with attention (see p. 39)
When do we need conditioning?

- Generate images based on a word
- Do we need to retrain a model for each word?

![Diagram](https://distill.pub/2018/feature-wise-transformations/)

Concatenation-based conditioning

“puppy”

 Concatenation-based conditioning
simply concatenates the conditioning representation to the input.

Concatenation-based conditioning

"puppy"

Concatenation-based conditioning simply concatenates the conditioning representation to the input.

Concatenation-based conditioning

The result is passed through a linear layer to produce the output.

Concatenation-based conditioning

- Source: image (high-dimensional) and pose (low-dimensional) → expressed as an image (same dimensionality)

L. Ma et al. „Pose guided person image generation“. NIPS 2017
Concatenation-based conditioning

- Source: image (high-dimensional) and pose (low-dimensional) → expressed as an image (same dimensionality)

Wait for the GAN intro in a few weeks!

L. Ma et al. „Pose guided person image generation“. NIPS 2017
Concatenation-based conditioning

- Sources: image (high-dimensional) and measurements (low-dimensional)

A. Dosovitskiy and V. Koltun. Learning to act by predicting the future. ICLR 2017
Conditional biasing

Think about the similarities with concatenation-based conditioning.

Conditional scaling

Conditional scaling first maps the conditioning representation to a scaling vector.

The scaling vector is then multiplied with the input.

Conditional scaling

• Reminds you of…. Gating
  – Long-Short Term Memory units

• Gating allows you to learn which inputs are more related between e.g. the two sources

• All conditioning so far is on a feature level → efficient and effective → number of parameters to be learned scales linearly with the number of features of the NN
Conditional scaling

• Can one do both conditional scaling and biasing?

Conditional Affine Transformation
The FiLM generator processes the conditioning information and produces parameters that describe how the target network should alter its computation.

Here, the FiLM-ed network’s computation is conditioned by two FiLM layers.

Information coming from e.g. the other source


In a **fully-connected** network, FILM applies a different affine transformation to each feature.

First, each feature (or channel) is scaled by the corresponding $\gamma$ parameter.

Then, each feature (or channel) is shifted by the corresponding $\beta$ parameter.

In a **convolutional** network, FILM applies a different affine transformation to each channel, consistent across spatial locations.


What can we do with conditioning?

• GuessWhat?! Visual object discovery through multi-modal dialogue. de Vries et al CVPR 2017
• A learned representation for artistic style. Dumoulin et al ICLR 2017
• Conditional image generation with PixelCNN decoders. van den Oord et al. NIPS 2016
Visual Question and Answering

The linguistic pipeline acts as the FiLM generator.

FiLM layers in each residual block modulate the visual pipeline.

Each residual block has a FiLM layer added to it.
Attention vs Conditioning

**Attention** computes a probability distribution over locations.

**Attention pools** over locations.

**Attention** summarizes the input into a vector.

**FiLM** computes a scaling vector applied to the feature axis.

**FiLM** conserves input dimensions.

Attention vs Conditioning

• Attention: assumes that specific locations contain the most useful information

• Conditioning: assumes that specific feature maps contain the most useful information
Next lecture

• No session on Friday

• Next Monday: no lecture – CVPR break –