Attention in Natural Language Processing

Andrea Galassi[1], Marco Lippi[2], Paolo Torroni[1]

[1]DISI, University of Bologna

[2]DISMI, University of Modena and Reggio Emilia

Published in:
IEEE Transactions on Neural Networks and Learning Systems
http://doi.org/10.1109/TNNLS.2020.3019893

Purposes of Attention

- Improve neural networks

  ⇒Make them focus on specific parts of the imput

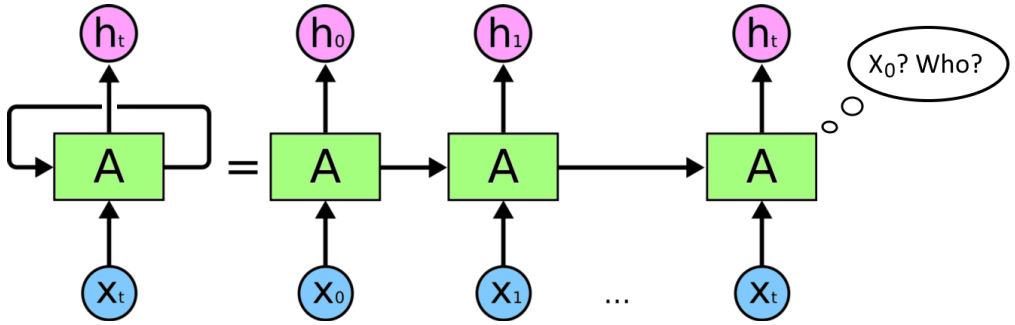
- Make deep networks (sort of) interpretable

  ⇒Allow human to see which part of the imput are considered relevant

- Alternative to RNNs for sequence-to-sequence

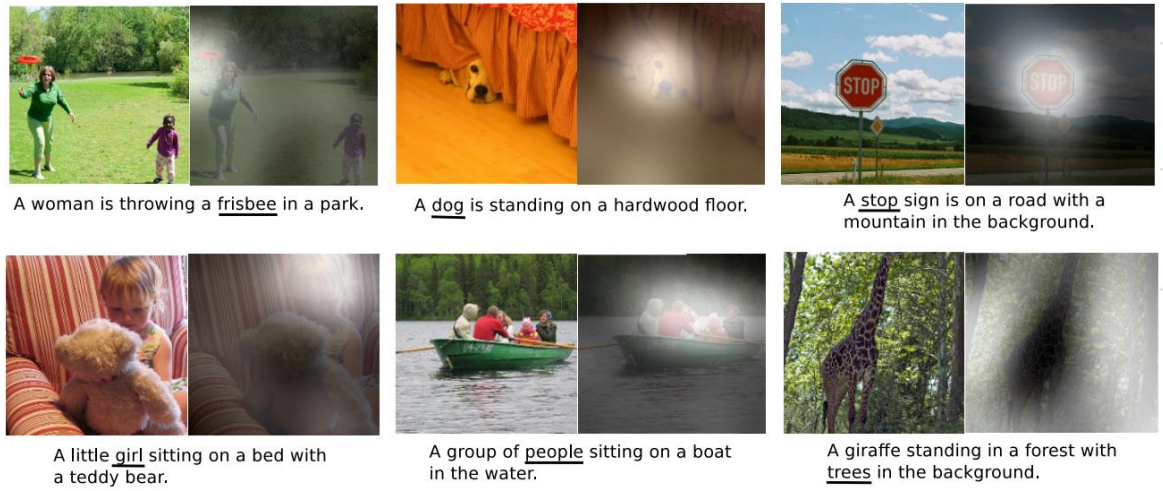
  ⇒Do not forget long range dependencies

Learning long-term dependencies with gradient descent is difficult (Bengio et al., 1994)



Interpretability

Task: Image Captioning



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (Xu et al., 2015)

Task: Aspect-based sentiment analysis

Task: Hotel location

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! **for location and price , this ca n't be beaten** , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel cleanliness

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel service

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . **service was excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Deriving Machine Attention from Human Rationales (Bao et al., 2018)

Historical Note

- CV: Glimpses, Visual Attention => focus on specific parts of images

  Learning to combine foveal glimpses with a third-order boltzmann machine (Larochelle and Hinton 2010), Recurrent models of visual attention (Mnih et al 2014)

- RNNsearch => first formulation of neural attention, encoder/decoder architecture for machine translation

  Neural machine translation by jointly learning to align and translate (Bahdanau et al 2015)

- Memory Networks => allow to exploit a virtual memory

  End-to-end memory networks (Sukhbaatar et al 2015)

  ...

- Transformer => completely replace recurrent/convolutional modules with attention for machine translation

  Attention is all you need (Vaswani et al 2017)

- BERT => exploits transformer to create word embeddings

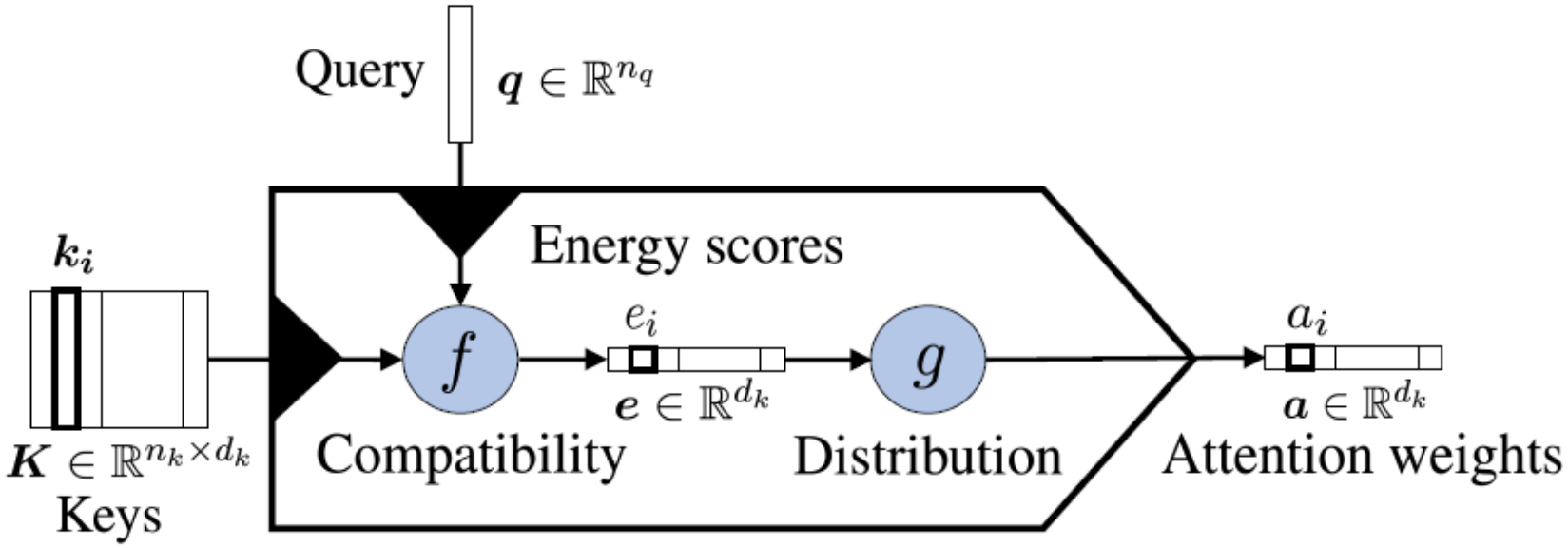
  BERT: pre-training of deep bidirectional transformers for language understanding (Devlin et al 2019)

Core Attention Model

Given a set of input elements (called **keys**) and a input **query**, the relevance of each key is computed as an attention **weight**.

Compatibility function *f* assesses the relevance of the keys.

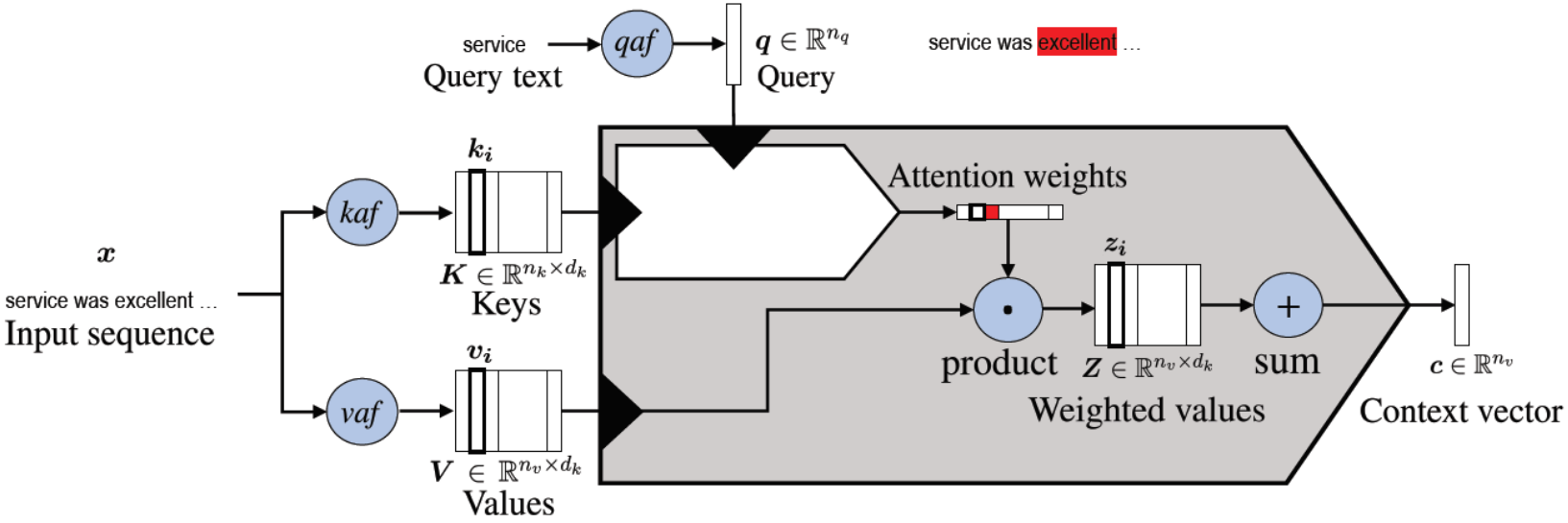
Distribution function *g* enforces desired properties.



General Attention Model

The attention weights are used to compute a weighted sum of the input set, so to obtain a compact representation: the **context vector**.

Two different representation of the input set are computed: keys are used to compute the weights, **values** are used to create the context vector.



Uses of attention

- Creation of a contextual embedding
  Machine Translation, Sentiment Analysis, Information Extraction
- Feature selection / Word selection
  Multi-modal tasks / Dependency parsing, Cloze question answering
- Seq-to-seq annotation
  Machine translation
- Multiple Input Processing
  Question Answering

Compatibility Functions

- **Input**: keys and query.
- **Output**: energy scores.
- **Purpose**: evaluate the relevance of each key according to the query.

W and *w* are (learned) parameters

Name	Equation
similarity	$f(q, K) = sim(q, K)$
multiplicative or *dot*	$f(q, K) = q^\intercal K$
scaled multiplicative	$f(q, K) = \frac{q^\intercal K}{\sqrt{n_k}}$
general or *bilinear*	$f(q, K) = q^\intercal W K$
biased general	$f(q, K) = K^\intercal (Wq + b)$
activated general	$f(q, K) = act(q^\intercal W K + b)$
concat	$f(q, K) = w_{imp}{}^\intercal act\big(W[K; q] + b\big)$
additive	$f(q, K) = w_{imp}{}^\intercal act(W_1 K + W_2 q + b)$
deep	$f(q, K) = w_{imp}{}^\intercal E^{(L-1)} + b^L$ $E^{(l)} = act\Big(W_l E^{(l-1)} + b^l\Big)$ $E^{(1)} = act\Big(W_1 K + W_0 q + b^1\Big)$
convolution-based	$f(q, K) = [e_0; \ldots; e_{d_k}]$ $e_j = \frac{1}{l} \sum_{i=j-l}^{j} e_{j,i}$ $e_{j,i} = act\Big(w_{imp}{}^\intercal [k_i; \ldots; k_{i+l}] + b\Big)$
location-based	$f(q, K) = f(q)$

Relevance of a key

Known:

similarity to the query

- Sentiment analysis
- Abusive speech

Unknown:

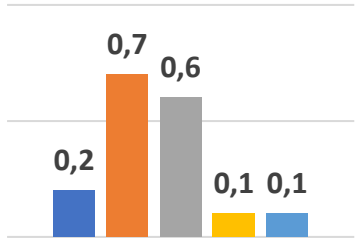
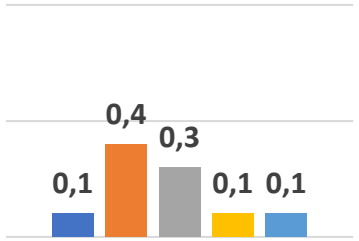
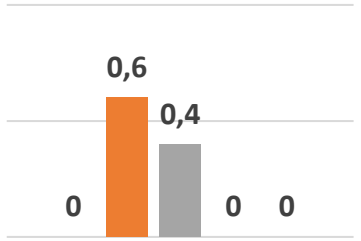
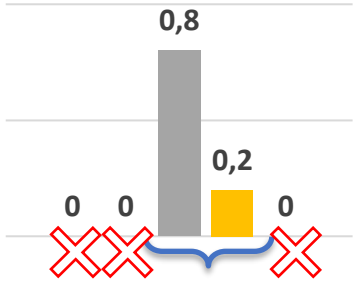
Similarity to a (learned) model w_{imp}

- Document classification
- Document summarization

Distribution Functions

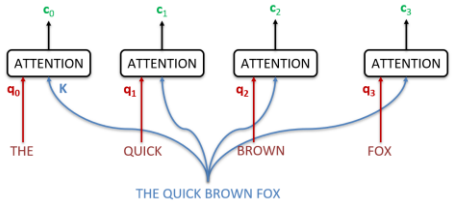
Input: energy scores. Output: weights.

Purpose: enforce some desired properties or criterion in the distribution

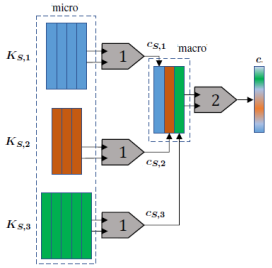
Properties		Sparsity Speeds up the computation Document summarization	Locality Selection Windows Gaussians Machine translation
Logistic sigmoid	Softmax	Sparsemax	Hard/Local Attention
			
Kim and Kim, 2018		Martins & Astudillo, 2016	Gregor et al., 2015; Luong et al., 2015; Xu et al., 2015; Yang et al., 2018

Other topics

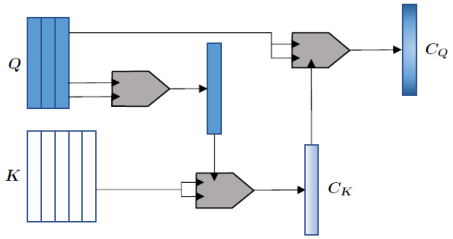
- Seq-to-seq annotation



- Hierarchical-input attention



- Interaction between two set of data (co-attention)

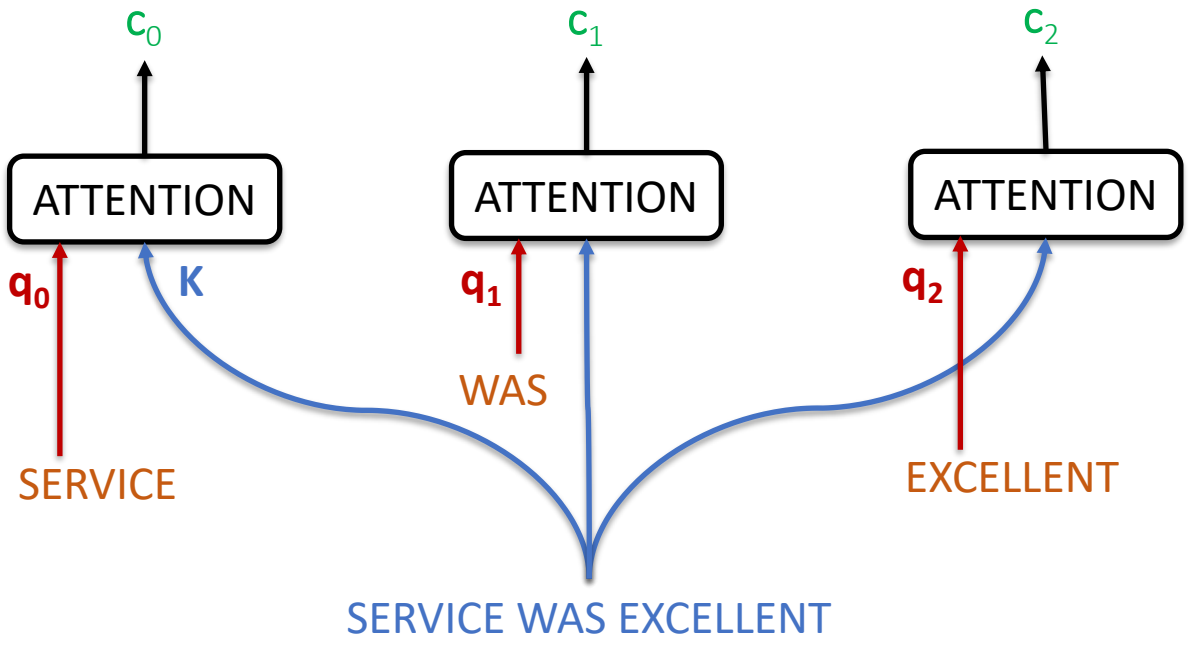


- Multi-output attention

- Exploiting knowledge: supervised attention

Sequence-to-sequence Annotation

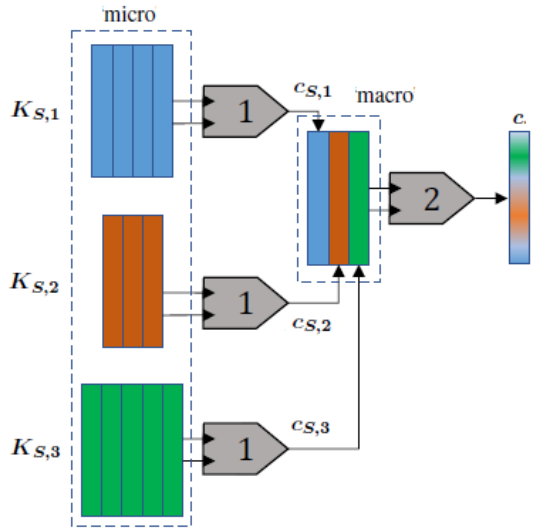
- Used for machine translation
- Perform attention multiple times
- Each time, one of the keys is used as query



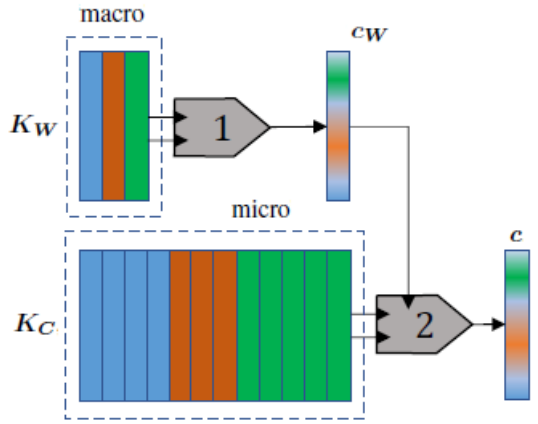
Hierarchical-input Attention

If the input has a hierarchical structure (e.g. characters, words, sentences, documents)

- Only the lower (micro) representation is available:
  - Iteratively apply attention to obtain higher-level representations
- Multiple representation are available:
  - Compute attention on one and use the output as query to compute attention on the other



Hierarchical attention networks for document classification (Yang et al., 2016)



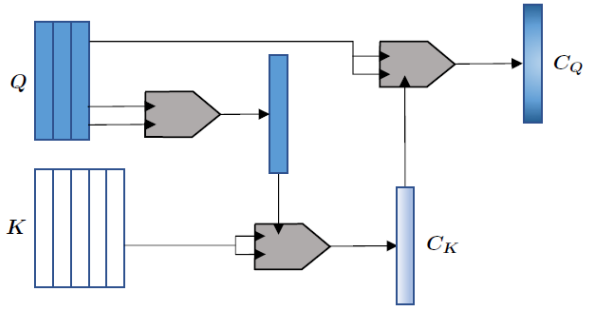
Attention-via-attention neural machine translation (Zhao et al., 2018)

Multi-input Attention

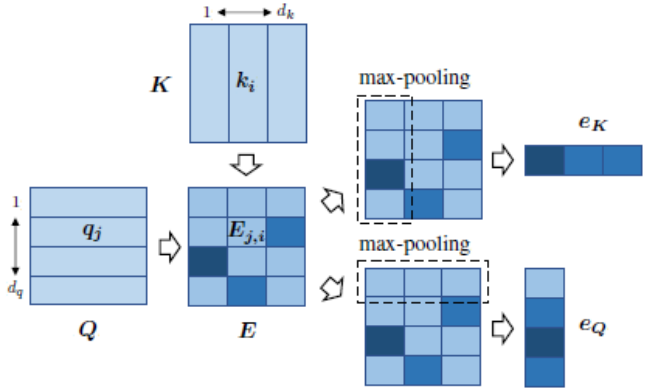
What if the query is a matrix, not a vector? For example, if the query is a set of words (=> multi-target sentiment analysis, question answering).

Attention can be applied on both, modelling interaction between the two sets.

- Coarse-grained co-attention:
  - Embed one set to perform attention on the other
- Fine-grained co-attention:
  - Compute a co-attention matrix E
  - Evaluate relevance of each pair of elements



Hierarchical question-image co-attention for visual question answering (Lu et al., 2016)



Attentive Pooling Networks (dos Santos et al., 2016)

Multi-output Attention

Purpose: obtain more than one relevance distribution

How:

- Change of parameters size

  A structured self-attentive sentence embedding (Lin et al., 2017)

- Multiple attention in parallel: Multi-head attention

  Attention is all you need (Vaswani et al., 2017)

- In classification task: a different attention for each possible class
  - Better error analysis

  Interpretable emoji prediction via label-wise attention lstms (Barbieri et al., 2018)

Also: possible to enforce different attention distributions through regularization

Multi-head attention with disagreement regularization (Li et al., 2018)

Supervised Attention

- Pre training, to model some knowledge which is already available
  - Detection of relevant parts

    Rationale-augmented convolutional neural networks for text classification (Zhang et al., 2016)

- Attention as an auxiliary task
  - Model specific knowledge
    - Relevance information

      Neural machine translation with supervised attention (Liu et al., 2016)
    - Semantic information

      Linguistically-informed self-attention for semantic role labeling (Strubell et al., 2018)
  - Mimic an existing attention model: Transfer Learning!

    1) Train attention model on a source task/domain
    2) Use the this model for supervised learning on a target task/domain

    Deriving machine attention from human rationales (Bao et al., 2018)

    Improving multi-label emotion classification via sentiment classification with dual attention transfer network (Yu et al., 2018)

Is Attention Explanation?

Can attention be used to explain neural networks? Debatable:

Attention is not exactly explanation

It might be, depending be the definition and the setting

Consistency: is attention consistent with other measures of feature importance?

No: it does not correlate with gradient-based and leave-one-out methods

Attention is not Explanation (Jain and Wallace 2019)

Plausibility: may attention provide a possible explanation?

Yes: attention weights may contain information regarding feature importance

Their use in context-less attention-less architectures leads to improvements

Attention is not not Explanation (Wiegreffe and Pinter 2019)

Faithfulness: may attention be the only true explanation of the outcome?

No: other different distributions may lead to the same outcome

It is possible to create adversarial distributions, but when attention contribution is stronger, it is more difficult to create an effective adversarial distribution

Attention is not not Explanation (Wiegreffe and Pinter 2019)

How to evaluate if attention is useful and/or reliable?

- Try to use the model with random and uniform weights distribution
- Train a simpler context-less architecture and see if the use of (pre-learned) attention weights improves it
- Try to train adversarial models with multi-objective learning:
  - Similar outcome
  - Different weights distribution

Attention is not not Explanation (Wiegreffe and Pinter 2019)

References

- Learning long-term dependencies with gradient descent is difficult (Bengio et al., 1994)
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (Xu et al., 2015)
- Deriving Machine Attention from Human Rationales (Bao et al., 2018)
- Learning to combine foveal glimpses with a third-order boltzmann machine (Larochelle and Hinton 2010)
- Recurrent models of visual attention (Mnih et al 2014)
- **Neural machine translation by jointly learning to align and translate (Bahdanau et al., 2015)**
- End-to-end memory networks (Sukhbaatar et al 2015)
- **Attention is all you need (Vaswani et al., 2017)**
- **BERT: pre-training of deep bidirectional transformers for language understanding (Devlin et al 2019)**
- **Neural turing machines (Graves et al., 2014)**
- **Effective approaches to attention-based neural machine translation (Luong et al., 2015)** **<= ArXiv version!**
- Iterative alternating neural attention for machine reading (Sordoni et al., 2016)
- Convolution-based neural attention with applications to sentiment classification (Du et al., 2019)
- Hierarchical attention networks for document classification (Yang et al., 2016)
- Attention-via-attention neural machine translation (Zhao et al., 2018)
- Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM (Ma et al., 2018)

References

- Deeper attention to abusive user content moderation (Pavlopoulos et al., 2017)
- Supervised domain enablement attention for personalized domain classification (Kim and Kim, 2018)
- From softmax to sparsemax: A sparse model of attention and multi-label classification (Martins & Astudillo, 2016)
- Draw: A recurrent neural network for image generation (Gregor et al., 2015)
- Modeling localness for self-attention networks (Yang et al., 2018)
- Hierarchical question-image co-attention for visual question answering (Lu et al., 2016)
- Attentive Pooling Networks (dos Santos et al., 2016)
- A structured self-attentive sentence embedding (Lin et al., 2017)
- Interpretable emoji prediction via label-wise attention lstms (Barbieri et al., 2018)
- Multi-head attention with disagreement regularization (Li et al., 2018)
- Rationale-augmented convolutional neural networks for text classification (Zhang et al., 2016)
- Neural machine translation with supervised attention (Liu et al., 2016)
- Linguistically-informed self-attention for semantic role labeling (Strubell et al., 2018)
- Deriving machine attention from human rationales (Bao et al., 2018)
- Improving multi-label emotion classification via sentiment classification with dual attention transfer network (Yu et al., 2018)
- Attention is not not Explanation (Wiegreffe and Pinter 2019)
- Attention is not Explanation (Jain and Wallace 2019)