Attention

a useful tool to improve and understand neural networks

Sala Riunioni DISI

V.le Risorgimento 2

Bologna

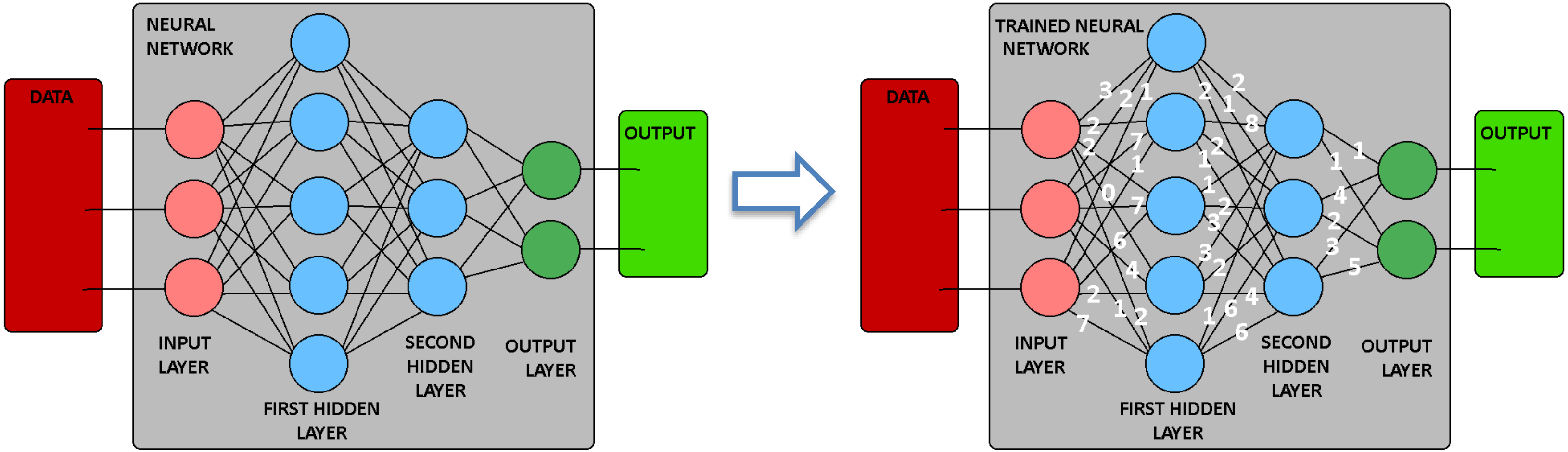
Jan 18th, 2019

Why do we need attention?

- Neural Networks are cool. They can learn lot of stuff and do amazing things.
- BUT! They are sub-symbolic system: knowledge is stored as numerical values

Andrea Galassi

Why do we need attention?



Knowledge acquired:

3, 2, 2, 0, 2; 2, 7, 7, 4, 1; 1, 1, 6, 2, 7;
2, 1, 2; 8, 2, 1; 1, 2, 3; 3, 2, 4; 1, 6, 6;
1, 1; 4, 2; 3, 5

CLEAR, RIGHT?

Andrea Galassi

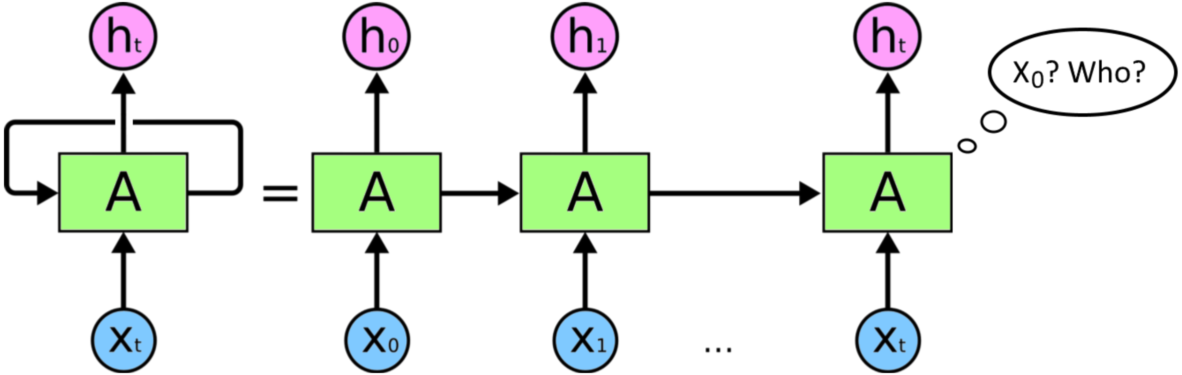
PhD ~~candidate~~ survivor

Why do we need attention?

- Recurrent Networks can be used to create sequence-to-sequence models

- BUT! They tend to forget long-range dependencies

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



Andrea Galassi

What is Neural Attention?

- Technique that can be applied in neural networks models to compute a specific weight for each input element, which assess its **relevance**
- Filter of the input => <u>better results</u> ☺
- <u>Interpretable result</u>: the higher the weight, the more relevant is the input ☺
- Seq-to-seq models that remember <u>long-range dependencies</u> ☺
- (most of the cases) Computationally <u>cheap</u> ☺

Andrea Galassi

PhD ~~candidate~~ survivor

Explainability!

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

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

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

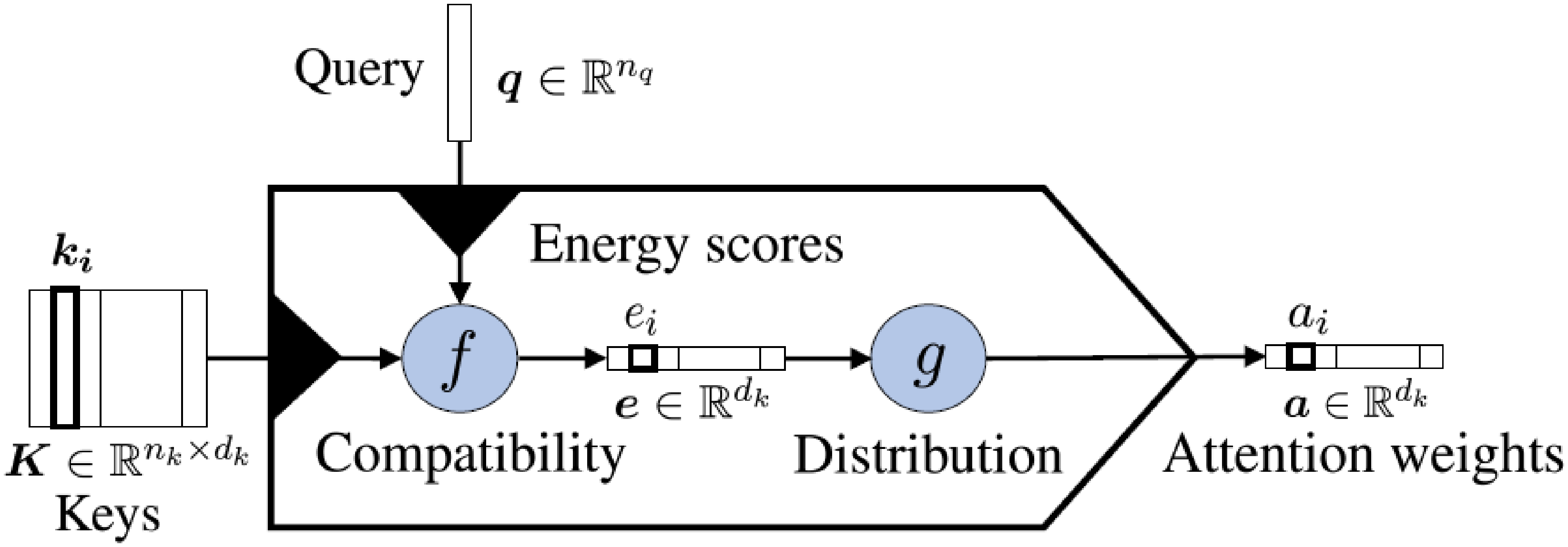
you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . **service was excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton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

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

Andrea Galassi

PhD ~~candidate~~ survivor

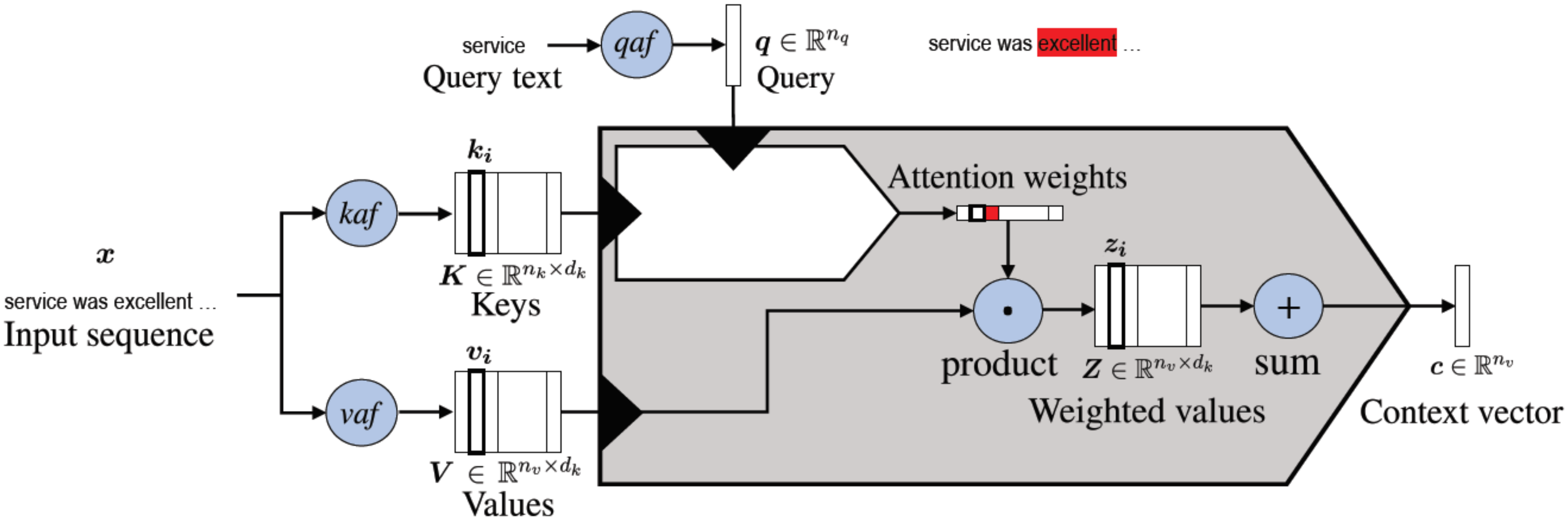
Core Attention Model



Andrea Galassi

PhD ~~candidate~~ survivor

General Attention Model



Andrea Galassi

PhD ~~candidate~~ survivor

Uses

- Embedding: the context is way smaller than the input
- Dynamic representation: if **q** changes, **c** changes !
- Selection: the weights can be used to classify the keys
- Seq-to-seq models
- Interaction between two set of data (co-attention)

Andrea Galassi

Compatibility Functions

- Compute the energy scores

Relevance of a key

Name	Equation	Reference
similarity	$f(q, K) = sim(q, K)$	Graves et al., 2014
multiplicative or *dot*	$f(q, K) = q^\intercal K$	Luong et al., 2015
scaled multiplicative	$f(q, K) = \frac{q^\intercal K}{\sqrt{d_k}}$	Vaswani et al., 2017
general or *bilinear*	$f(q, K) = q^\intercal W K$	Luong et al., 2015
biased general	$f(q, K) = K^\intercal (W q + b)$	Sordoni et al., 2016
activated general	$f(q, K) = act(q^\intercal W K + b)$	Ma et al., 2017
concat	$f(q, K) = w_{imp}{}^\intercal act(W[K; q] + b)$	Luong et al., 2015
additive	$f(q, K) = w_{imp}{}^\intercal act(W_1 K + W_2 q + b)$	Bahdanau et al., 2015
deep	$f(q, K) = w_{imp}{}^\intercal E^{(L-1)} + b^L$ $E^{(l)} = act(W_l E^{(l-1)} + b^l)$ $E^{(1)} = act(W_1 K + W_0 q + b^1)$	Pavlopoulos et al., 2017
location-based	$f(q, K) = f(q)$	Luong et al., 2015

Similarity to **q**

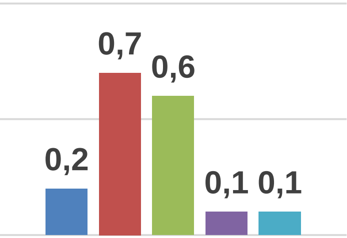
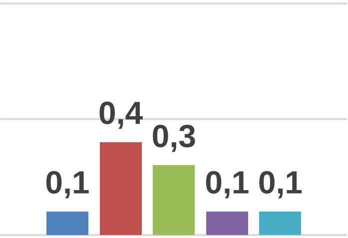
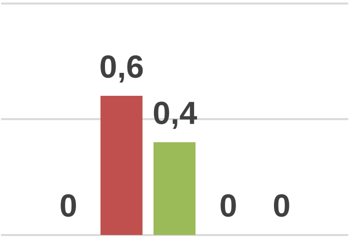
Similarity to a learned model $\mathbf{w}_{imp}$

Andrea Galassi

PhD ~~candidate~~ survivor

Distribution Functions

- From energy scores to weights

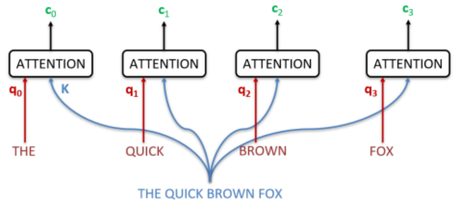
Properties		Sparsity Speeds up the computation	Locality Selection Windows Gaussians
Logistic sigmoid	Softmax	Sparsemax	Hard/Local Attention
			
Kim and Kim, 2018		Martins & Astudillo, 2016	Gregor et al., 2015; Luong et al., 2015; Xu et al., 2015; Yang et al., 2018

Andrea Galassi

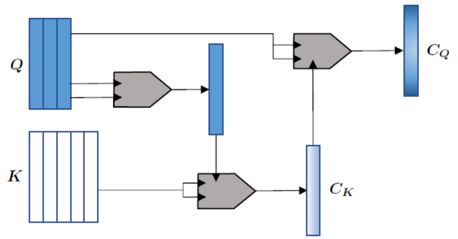
PhD ~~candidate~~ survivor

Other topics

- Seq-to-seq models



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



- Multi-output attention

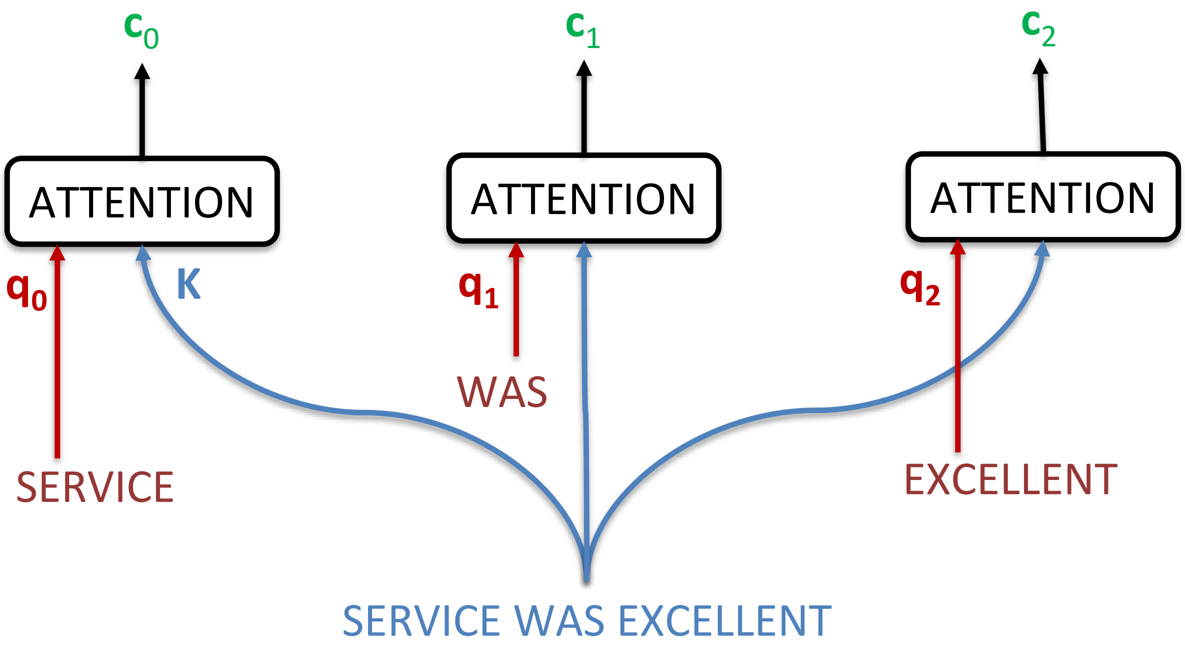
- Exploiting knowledge: supervised attention

Andrea Galassi

PhD ~~candidate~~ survivor

Seq-to-seq

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

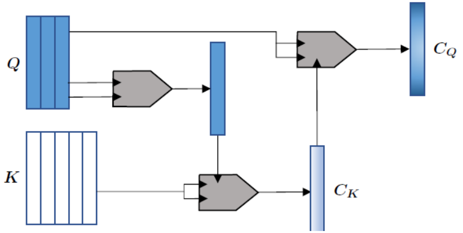


Andrea Galassi

Multi-input attention: Co-attention

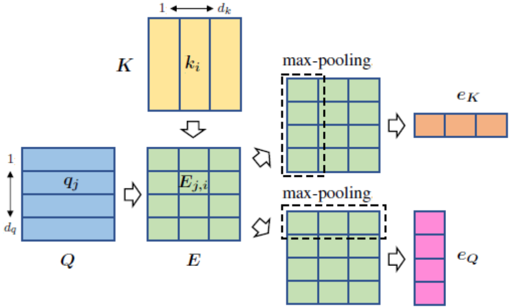
- If **q** is matrix? Two matrices of data: **K** and **Q**
- Attention on both
- Interactions between the two sets

- Coarse Grained:
  - Embedding of the other set



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

- Fine Grained:
  - Co-attention matrix G: Energy score for each pair



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

Andrea Galassi

PhD ~~candidate~~ survivor

Multi-output attention

- More than one relevance distribution
  - 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)
- Possible to enforce different attention distributions through regularization

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

Andrea Galassi

PhD ~~candidate~~ survivor

Supervised Attention

- Pre training, to model some knowledge
  - 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)

Andrea Galassi

PhD ~~candidate~~ survivor

Conclusion

- Attention is nowadays a key component in neural architectures
- Improves neural architectures, allowing also their explanation, without increasing costs
- Popular trend in NLP and CV, but not only
  - 40+ works EMNLP18
  - 40+ works AAAI18
  - 30+ works IJCAI18

- Future: Could it be used to understand deep networks?

Andrea Galassi

PhD ~~candidate~~ survivor

This and much more on

Attention, please!
A Critical Review of
Neural Attention Models in NLP

Galassi A., Lippi M., Torroni P., 2019

https://arxiv.org/abs/1902.02181

Andrea Galassi

PhD candidate survivor

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)
- **Neural turing machines (Graves et al., 2014)**
- **Effective approaches to attention-based neural machine translation (Luong et al., 2015)** **<= ArXiv version!**
- **Attention is all you need (Vaswani et al., 2017)**
- Iterative alternating neural attention for machine reading (Sordoni et al., 2016)
- Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM (Ma et al., 2018)
- **Neural machine translation by jointly learning to align and translate (Bahdanau et al., 2015)**
- 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)

Andrea Galassi

PhD ~~candidate~~ survivor

References

- 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)

Andrea Galassi