Stacked Attention Network

by Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

arXiv:1511.02274v2 [cs.LG] 26 Jan 2016

Slides by Jakub Hajic

7 Mar 2016

Stacked Attention Network

Basic idea:

- obtain a region-based representation of the image
  - (params: VGG16 cut before dense layers, with 14x14 regions and region representation dimension 512)
- obtain a vector representation of the query
  - Embedding + LSTM or CNN
- combine the two to obtain a weight for each image region (**‘attention’**)
- combine weighted sum of regions and query to create new query
- repeat process with new query (**‘stacked’**)

Stacked Attention Network

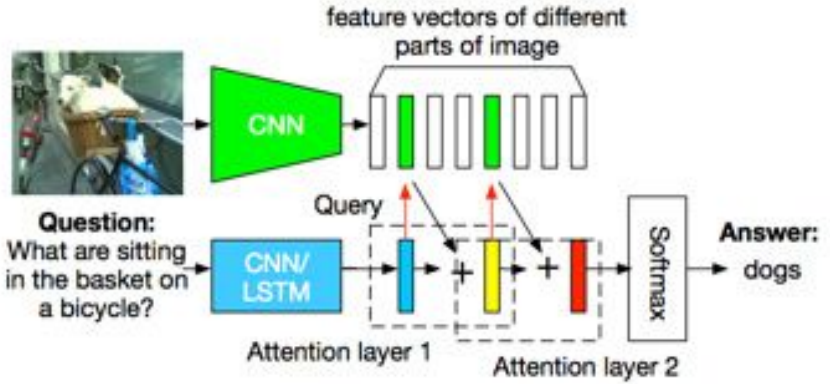


figure from Yang et al.

End-To-End Memory Network

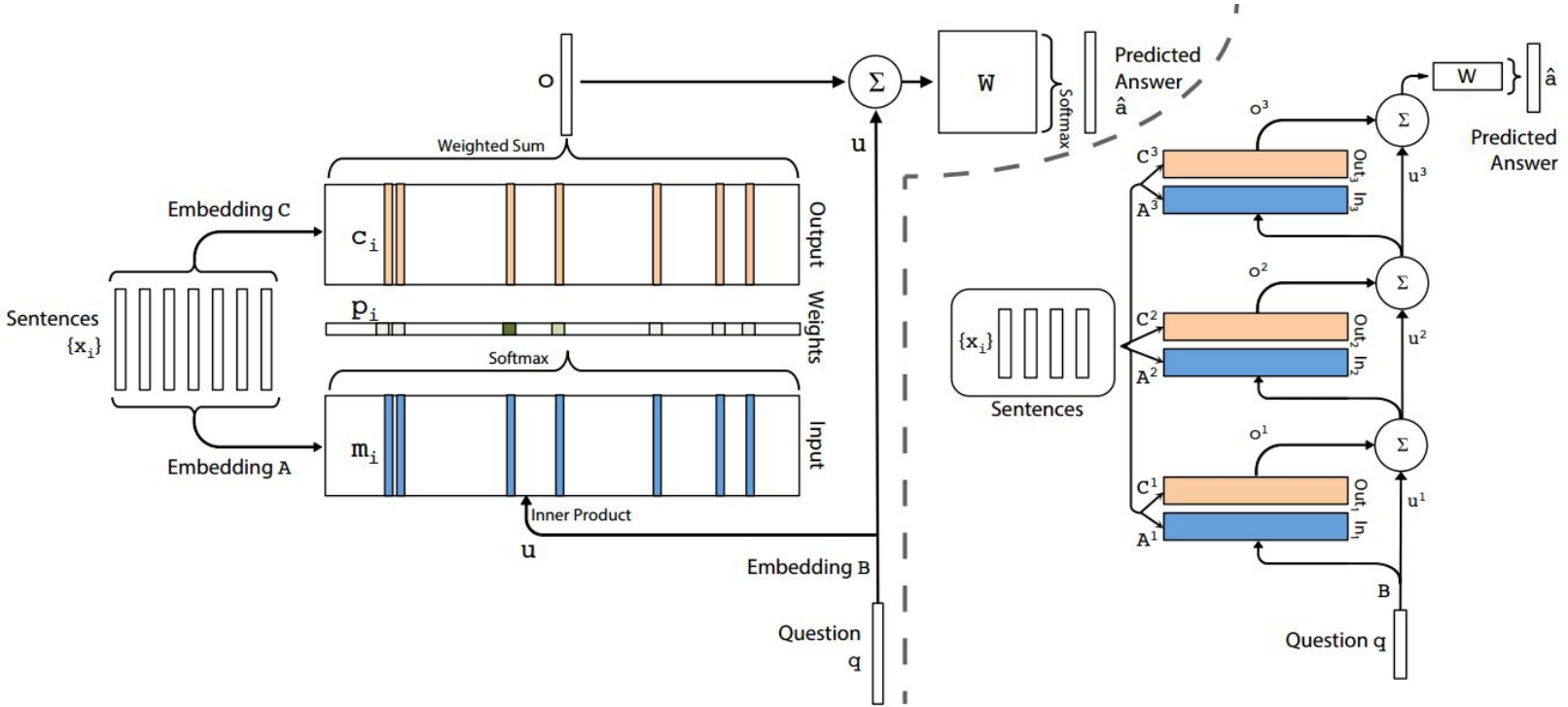
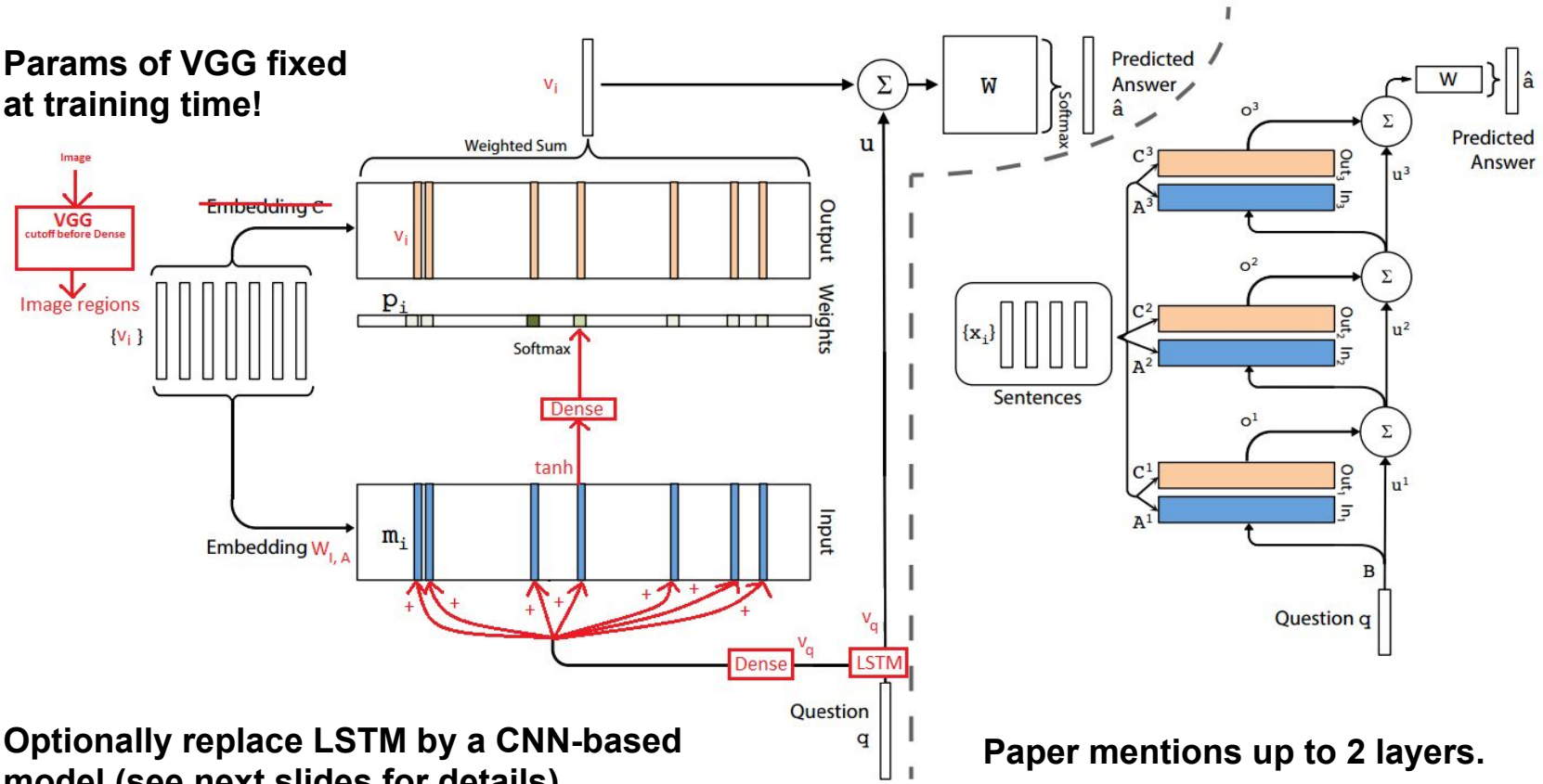


figure from Sukhbaatar et al.: End-To-End Memory Networks, arXiv:1503.08895v5 [cs.NE] 24 Nov 2015

SAN compared to MemNN

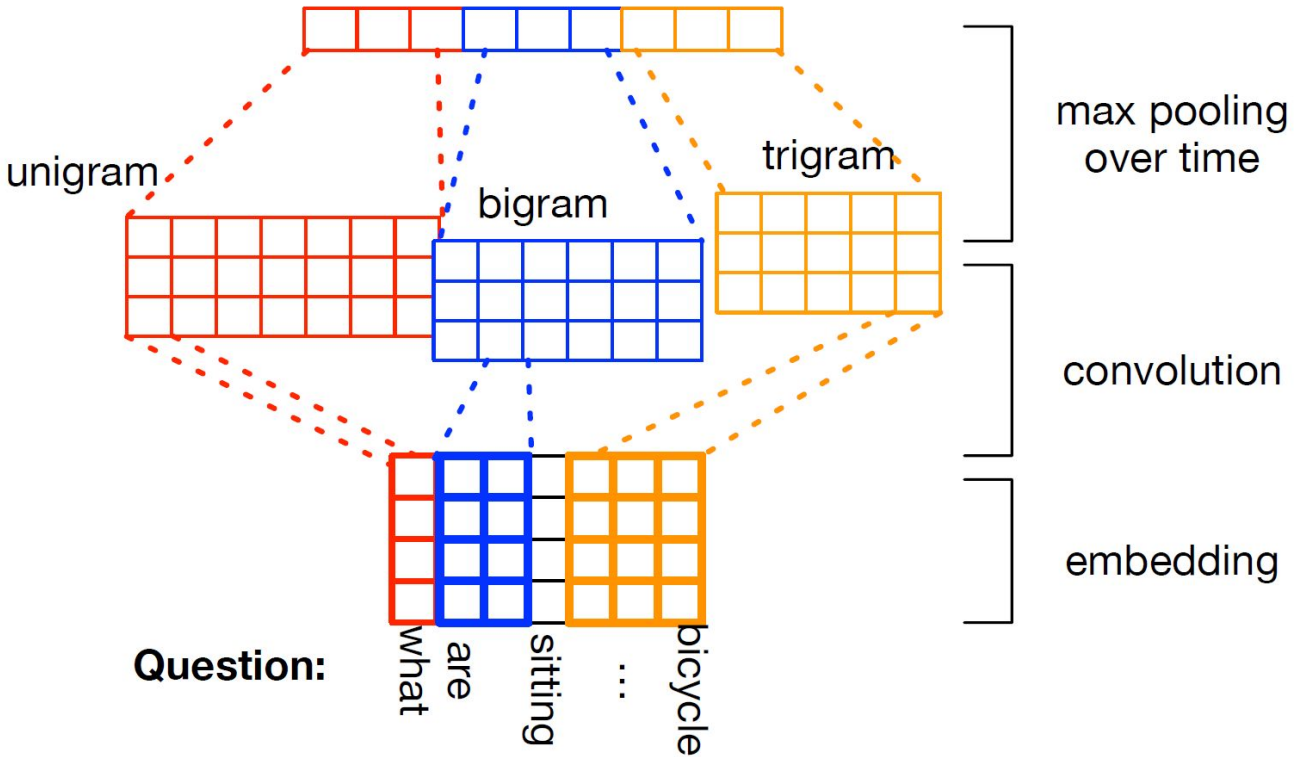
Params of VGG fixed at training time!



Optionally replace LSTM by a CNN-based model (see next slides for details)

Paper mentions up to 2 layers.

SAN: CNN-based model



SAN: CNN-based model

Basic Idea:

- Embed query
- Apply a unigram, bigram and trigram convolution operation
  - given by $h_{c,t} = \tanh(W_c x_{t:t+c-1} + b_c).$, where *c* equals 1, 2, or 3 for uni-, bi- or trigrams
- Run max-pooling over *t* for every *c = 1, 2, 3* to obtain h'_1, h'_2, h'_3
- Concatenate results to form query representation $h = [h'_1, h'_2, h'_3]$

SAN: success case

figure from Yang et al. From left: Original image, first attention layer, second attention layer
Direct quote from paper: “The bright part of the image is the detected attention”