Attention based model 是什么，它解决了什么问题？

机器学习　深度学习（Deep Learning）

关注者 1514　被浏览 5899

Attention based model 是什么，它解决了什么问题？

Attention 这个词似乎太宽泛了，搜索也没有找到想要的结果。

在没有这 attention 之前遇到了什么问题？
attention 具体是什么样的思想？

之前在Colah君的blog上好像也见他说过。
主要是在看NVIDIA的一篇blog时有点看不懂，故有此问。
Introduction to Neural Machine Translation with GPUs (part 3)

谢谢！

关注问题　写回答　添加评论　分享　邀请回答　…　收起

15 个回答　默认排序

Tao Lei

491 人赞同了该回答

我从NLP的角度来说吧。

NLP里面有一类典型的natural language generation问题：给定一段上下文(context)，生成一段与context相关的目标文本(target)。

▲ 491　▼　18 条评论　分享　收藏　感谢　收起

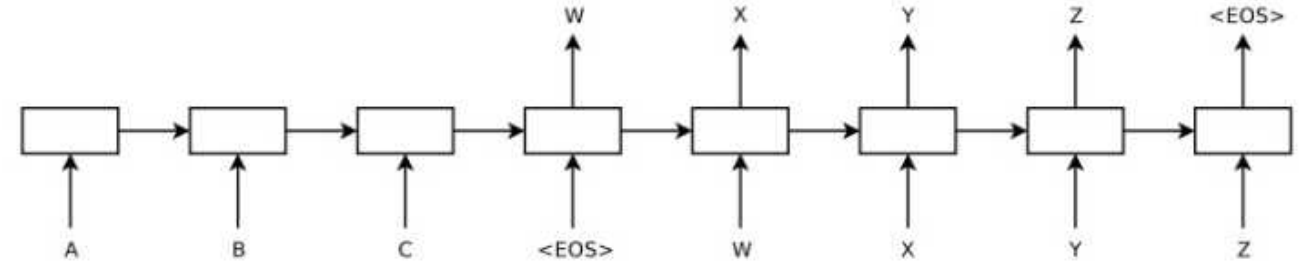
Attention based model 是什么，它解决了什么问题？

- 阅读理解：context是一段文章和一个选择题，需要输出答案。。

这类问题的核心是要对 条件概率 P(target|context) 进行建模。

在Deep learning火起来后，最常见的建模方式是用Recurrent Neural Networks (RNN) 将上下文"编码"，然后再"解码"成目标文本。

以机器翻译为例。Google最近的论文[1]中，用一个 RNN encoder读入context， 得到一个context vector（RNN的最后一个hidden state）；然后另一个RNN decoder以这个hidden state为起始state，依次生成target的每一个单词。如下图：



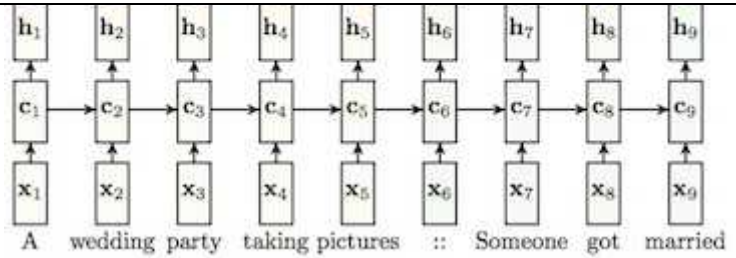
Sounds like a magic!! But,

这种做法的缺点是，无论之前的context有多长，包含多少信息量，最终都要被压缩成一个几百维的vector。这意味着context越大，最终的state vector会丢失越多的信息。正如楼主贴出blog中的Figure 1所显示，输入sentence长度增加后，最终decoder翻译的结果会显著变差。

事实上，因为context在输入时已知，一个模型完全可以在decode的过程中利用context的全部信息，而不仅仅是最后一个state。

Attention based model的核心思想

Attention based model 是什么，它解决了什么问题？



- 首先，在生成target side的states时（$h_7 \cdots h_9$），所有context vectors（$h_1 \cdots h_5$）都会被当做输入。
- 其次，并不是所有context都对下一个状态的生成产生影响。例如，当翻译英文文章的时候，我们要关注的是“当前翻译的那个部分”，而不是整篇文章。“Attention”的意思就是选择恰当的context并用它生成下一个状态。

在大部分的论文中，Attention是一个权重vector（通常是softmax的输出），其维度等于context的长度。越大的权重代表对应位置的context越重要。不同论文 [2,3,4,7] 对attention权重的计算方式不同，但其核心抛不开上述两点。

如果将生成每个target word前的attention vector拼接起来，可以观察到类似word alignment的有趣矩阵。下图来自最近的sentence summarization paper [3]：

Attention based model 是什么，它解决了什么问题？

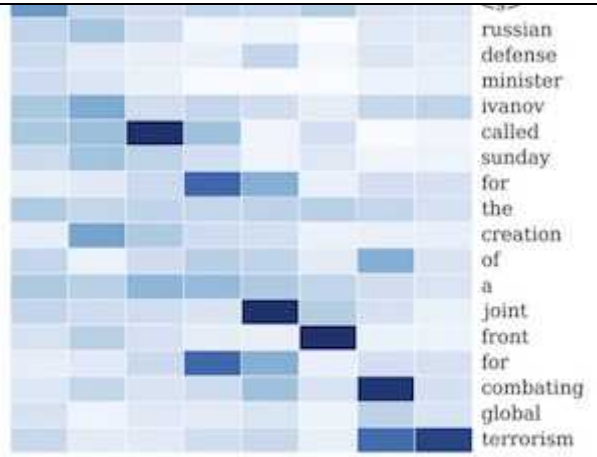


Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.

此外Attention的用途不仅限于text generation / language modeling。还可以做QA (question answering)，以及language inference（判断前后两个句子是否“蕴含”、“矛盾”或者“无关” [2]）等等。

如果希望了解Attention的Technical detail，可以参考最近Google Deepmind和Alexander Rush等人的论文。

补充：
感谢 @hdsh tesr 的评论和提出的问题。在这里说明一下：

1. Google research团队并不是
   类似的idea在2013年就出现过

Attention based model 是什么，它解决了什么问题？

由于题主提问的是Attention based model，答主只是选择Google的工作作为例子和Background，并未深入探讨Neural MT。对此可能造成的误解表示抱歉。

抛开Google在机器翻译界的影响力不谈。值得一提的是，Google的该工作是第一个用纯end-to-end neural network system，达到甚至超过传统MT系统的。此前，机器翻译系统一直是highly engineering甚至是dirty的。

2. Attention严格意义上讲是一种idea，而不是某一个model的实现。用到该思路的论文 [2,3,4,7] 的实现方式也可以完全不同。例如，在Alexander Rush最近的Summarization Paper中 [3]，就完全没有使用RNN对context编码，而是直接将context中的单词作一次linear projection，并smooth一下。其计算attention vector的方式也没有用到neural activations比如tanh， 而是直接bi-linear scoring + softmax normalization。相比之下，[2][4]和[7]的实现更为"deep learning"，甚至attention vector本身也是recurrent的。

[1] Sequence to Sequence Learning using Neural Networks
[2] Reasoning about Neural Attention
[3] A Neural Attention Model for Abstractive Sentence Summarization
[4] Neural Machine Translation by Jointly Learning to Align and Translate
[5] Recurrent Continuous Translation Models
[6] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
[7] Teaching Machines to Read and Comprehend

编辑于 2015-10-27

Attention based model 是什么，它解决了什么问题？

94 人赞同了该回答

最近刚好在看Attention相关的paper，并整理在了博客Attention中，跑来回答一下自己的理解。

在引入Attention(注意力)之前，图像识别或语言翻译都是直接把完整的图像或语句直接塞到一个输入，然后给出输出。
而且图像还经常缩放成固定大小，引起信息丢失。
而人在看东西的时候，目光沿感兴趣的地方移动，甚至仔细盯着部分细节看，然后再得到结论。
Attention就是在网络中加入关注区域的移动、缩放机制，连续部分信息的序列化输入。
关注区域的移动、缩放采用强化学习来实现。

Attention在图像领域中，物体识别、主题生成上效果都有提高。

Attention可以分成hard与soft两种模型:

- hard: Attention每次移动到一个固定大小的区域
- soft: Attention每次是所有区域的一个加权和

相关博客与论文有:
Survey on Advanced Attention-based Models
Recurrent Models of Visual Attention (2014.06.24)
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015.02.10)
DRAW: A Recurrent Neural Network For Image Generation (2015.05.20)
Teaching Machines to Read and Comprehend (2015.06.04)
Learning Wake-Sleep Recurrent Attention Models (2015.09.22)
Action Recognition using Visual Attention (2015.10.12)
Recursive Recurrent Nets with Attention Modeling for OCR in the Wild (2016.03.09)

编辑于 2016-05-24

Attention based model 是什么，它解决了什么问题？

过拟合

33 人赞同了该回答

Attention实质上是一种content-based addressing的机制，即从网络中某些状态集合中选取与给定状态较为相似的状态，进而做后续的信息抽取。

假设给定一个向量 u ，我们希望从另一个向量集合 $v_1, v_2, \ldots, v_n$ 中抽取代表向量，来计算向量与该集合的相似度。假设已经给定向量与向量的相似度计算函数 $f(u, v)$ （可以是一个小型神经网络）。

做法1：对每个 v_i 计算相似度 $a_i = f(u, v_i)$ ，然后取最大 a_i 所对应的 v_i 作为代表向量，最后计算相似度。最后这一步可以使用同一 f ，也可以单独再训另一个神经网络。这样做的缺点是，如果向量集合中有很多向量的相似度 a_i 是近似的，那么取 max 这种只能留一个向量的操作事实上会丢掉大量重要信息。

做法2：取向量集合的平均向量作为代表向量，然后计算最终的相似度。这样做的缺点是，如果只有一个向量与给定向量显著相关，其他几乎不相关，那么由于取平均操作，会导致真正具有代表性的向量被其他“噪声”向量所淹没。这点在向量集合很大的时候尤为明显。

可见,我们既想在有多个相似向量的情形下尽量在最终的代表向量中保留这些向量的信息，又想在只有一个显著相关向量的情形下直接提取该向量做代表向量，避免其被噪声淹没。那么解决方案只有：加权平均。

我们可以利用相似度作为权值对向量做加权平均，这样向量越相似，在最终的代表向量中其权重越高，保留的信息越多，反之亦然。由于相似度本身可能scale千差万别，直接加权平均会导致代表向量的scale的暴增或者暴减，我们利用softmax函数将这些相似度归一化，作为最终的权值。这样得到的代表向量就是原向量集合的“软性”max，因为所有向量信息都有保留，但是最相关的向量又恰好保留的最多，与硬性的max这种留

Attention based model 是什么，它解决了什么问题？

Content-based Addressing机制可以在很多地方用到。之前做过一个应用，在计算query与广告相关性时，广告可能有多重内容表示（title, content, keywords等等），这样实质上就需要建模一个向量（query的embedding）与一个向量集合的相似度（广告的所有内容embedding），利用上述机制训练出的模型相较直接用max的模型AUC提升至少2个点以上。 另外在机器翻译中，Bahdanau等人提出将encoder的状态序列保留，然后在decoder阶段的每一步都从状态序列中抽取与decoder的上一步状态相似的代表向量，来做为当前这一步状态的输入信息，从而达到alignment的功效。事实上在Neural Turing Machine一文中，Alex Graves等人清晰地阐述了将网络中部分状态保留在memory中，并且利用content-based addressing读取memory得到代表向量，然后参与到网络后续状态演化过程的想法。上述的两个例子，以及后续的Attention相关的论文都可以看做这种机制的一种实例。

编辑于 2017-04-01

▲ 33 ▼ 5 条评论 分享 收藏 感谢 收起

罗浩.ZJU
控制科学与工程博士在读

25 人赞同了该回答

可以看下我写的几篇博客，第一篇是基础，第二篇是一个具体的应用，第三篇也是个简单的应用介绍，按照这个顺序看就行，希望有所帮助

Attention model - 迷川浩浩的博客 - 博客频道 - CSDN.NET

基于Visual attention的图片主题生成 - 迷川浩浩的博客 - 博客频道 - CSDN.NET

基于attention的video描述 - 迷川浩浩的博客 - 博客频道 - CSDN.NET

其实我觉得Attention model就是比以前的网络多训练了一个有关于Attention的东西，这个东西需要用到递归过程来更新，所以一般是和RNN结合起来的，当然空间CNN也行

▲ 491 ▼ 18 条评论 分享 收藏 感谢 收起

Attention based model 是什么，它解决了什么问题？

比如我们做文本翻译的时候，cat and dog，翻译成“猫和狗”，显然cat 对于“猫”的贡献要远超过其他几个单词，所以Attention model 就是来训练这个权重，随着过程的进行，每次特征向量的权重都在变化，比如cat用完了，下一次网络可能就会更加关心and 和 dog这两个单词，所以下一个单词翻译成“和” “狗”的概率会增加，使得最后的结果优于传统的RNN或者CNN网络

编辑于 2016-12-19

Ying Zhang
无话可说

15 人赞同了该回答

语音上的attention model思路大体和nlp上的attention一致，@Tao Lei已经解释的很清楚。事实上就我所知的第一篇语音识别中的attention model的思路就来自于Dzmitry Bahdanau去年发表的文章 arxiv.org/pdf/1409.0473.... Speech attention发表在去年nips workshop上，题目是 End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results(arxiv.org/abs/1412.1602)，模型表现虽然略逊于SOTA, 但是这种结果还是令人振奋的。今年Dzmitry Bahdanau改良了speech attention model(之前的speech attention model仍需要借助HMM或者CTC等其他模型才可以得到phoneme/character label), 使得模型可以变得end-to-end，从音频序列直接输出得到character label，使得模型更加灵活。我暂时还没有仔细拜读这篇文章，有兴趣同学可以查阅arxiv.org/abs/1508.0439...。不得不提的是，Dzmitry出生于91年，google citation已然300+，只能说人比人气死人。:-(

编辑于 2015-11-03

阳阳

Attention based model 是什么，它解决了什么问题？

18 人赞同了该回答

图像视频方面的“Attention”很容易直观理解，就是所谓的“注意力”。你看一幅图并不是每个像素点都看的，而只是关注一个区域，有点类似ROI(Region of Interest)。之前无论是crop还是slide window都消耗了大量的计算，而且patch尺寸是不能动态调整的，各种bounding对复杂情况又难以找全所有ROI。Attention有种自动寻找与结果相关的ROI的意思。

Attention有hard attention和soft attention之分。而关于attention如何确定，传统的NN训练是无能为力的。这里要用到强化学习(reinforcement learning)。RL大法固然好，但有一个问题是方差大。这就是说，如果你要建一个大网络，简直就是个灾难。所以大家想要找一个可微attention模型（soft attention），即attention项和作为结果的loss function都是输入的可微函数，这样梯度信息保留下来，就可以用BP愉快地训练啦。

编辑于 2016-05-24

知乎用户
肥宅

5 人赞同了该回答

楼上好多人提到了attention model 在NLP和图片生成领域的应用，其实attention model 还有一个非常重要的应用场景，就是 active vision.

在deep mind 的NIPS paper “Recurrent Models of Visual Attention ”中提出了一个非常有意思的模型。相比较于直接处理整张图片，我们其实只需要在每一个 time step 获取一个小patch就足够完成某个task(比如 digit recognition)。Agent 每次只需要选取一个小patch,并且将此patch的图片信息与location信息encode在一起，生成一个小模块称之为glimpse。这个glimpse又是RNN的当前输入，根据输出的hidden state 可以来做

Attention based model 是什么，它解决了什么问题？

如果场景更复杂一点呢？毕竟识别简单的数字并不能显示attention model的强大之处，有人将这个模型与active vision结合在一起，发表在ECCV 2016，还是一篇oral，Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion。此时agent就不是在2D空间里预测下一个patch的坐标了，而是选择下一个action。Agent每走一步，选择下一个action，并且识别object的category。这么做的意义就是在很多情况下，因为occlusion的问题，仅仅从一张图片无法识别出这个object，此时我们就需要通过新的action到达一个新的viewpoint，从这个viewpoint得到更多的信息，来帮助agent识别出此object。这篇论文的实验有两个，object recognition 和scene recognition, 其实问题本质上都一样，都可以把这些过程简化成一个视觉信息为输入的POMDP问题。

编辑于 2017-05-15

tramphero
PHD在读,运动和科研

15 人赞同了该回答

Attention based model 是什么，它解决了什么问题？

translation，ent-to-end speech recognition, image caption, question answering..相应的论文都可以在workshop的references里找到下载链接。。

发布于 2015-10-24

▲ 15 ▼ 1 条评论 分享 收藏 感谢

知乎用户
翦渣

2 人赞同了该回答

encoder-decoder的structure inference

编辑于 2015-10-25

▲ 2 ▼ 2 条评论 分享 收藏 感谢

严本本

3 人赞同了该回答

在这里，也推荐一下本人总结的attention 的文章，希望对你有帮助blog.csdn.net/bvl101011...

发布于 2017-11-07

▲ 3 ▼ 5 条评论 分享 收藏 感谢

NLP小学生
又红又专的人工智能工程师

4 人赞同了该回答

▲ 491 ▼ 18 条评论 分享 收藏 感谢 收起

Attention based model 是什么，它解决了什么问题？

Align and Translate 里提到了attention mechanism这个词。从题目也看到了，作者提出这一概念的时候并没有突出「attention]，这只是一种「jointly align and translate」的技术。原文第一次提到「attention」的句子是：『 Intuitively, this implements a mechanism of attention in the decoder. The decoder decides parts of the source sentence to pay attention to』。可以感觉到作者那时并没有意识到「attention mechanism」会火起来。

简单说，在机器翻译的时候，要知道译句的每个词对应是原句的哪个词，这叫「alignment」。最初神经机器翻译并没有这个对齐的过程，只是粗暴的序列到序列的映射。Bahdanau等人考虑自动训练对齐过程，方式就是翻译的时候为每个原词分配一个score，score越大我们就应该越注意这个词。在这篇论文里，attention就是为原句的每个词分配score的过程。详见对这篇论文的解读：论文阅读

。

发展到今天，attention已经广泛应用于各种深度学习任务。比如图像识别的时候，人类分辨一只猫要重点关注鼻子眼睛，那么attention based model也为图片里鼻子眼睛分配更多的score。

attention 具体就是为训练样本的各部分分配score的思想，来模拟人类的「注意」，人越注意某部分，某部分的score就越大。

编辑于 2017-10-25

▲ 4 ▼ 添加评论 分享 收藏 感谢

Haoskism
Das Ding an sich

2 人赞同了该回答

人的脑力是有限的，所以需要优先处理重要的信息，因此眼睛演化出视锥细胞与视杆细胞对应神经节细胞不同的机制，分别用来仔细
部分视觉信息在进入大脑之前就已

▲ 491 ▼ 18 条评论 分享 收藏 感谢 收起

Attention based model 是什么，它解决了什么问题？

觉机制最为基本和简单的能力，识别只是基础，而对不同层次信息的初步筛选，精度筛选然后处理的机制才是处理视觉信息（可见光）的最重要部分。
已知世界如此之小，已做之事如此之少，哎~
这个答案就是写着玩的。

发布于 2016-08-22

▲ 2 ▼ 1 条评论 分享 收藏 感谢

Qoboty
专注图像，语音领域优秀架构

Attention based model for 语音识别 对Attention在语音识别中的应用进行了总结和介绍

发布于 2017-08-11

▲ 0 ▼ 添加评论 分享 收藏 感谢

知乎用户
倍返しだ

两个问题：

1，attention如果是看上下文计算当前输入的权值，影响下一层决策的话。文本分类，一条条独立的句子被输入，上下文的句子跟当前句子毫无关联，attention到底在干嘛？难道所谓的上下文只是一个sequence中的上下文？

2，attention是如何被训练的？

发布于 2017-06-16

▲ 491 ▼ 18 条评论 分享 收藏 感谢 收起

Attention based model 是什么，它解决了什么问题？

caitouwh

顺着问一下（不是回答）

可以理解RNN 中的attention 就是 CNN中的filter吗？　RNN引入ＣＮＮ　机制可以提高结构性的分析性能？　就好比ＣＮＮ现在要在 filter 中引入RNN。

纯初学，请教。

发布于 2017-08-21