An Attentive Survey of Attention Models

Sneha Chaudhari[*], **Gungor Polatkan**, **Rohan Ramanath**, **Varun Mithal**
AI@LinkedIn
{snchaudhari, gpolatkan, rramanat, vamithal}@linkedin.com

Abstract

Attention Model has now become an important concept in neural networks that has been researched within diverse application domains. This survey provides a structured and comprehensive overview of the developments in modeling attention. In particular, we propose a taxonomy which groups existing techniques into coherent categories. We review the different neural architectures in which attention has been incorporated, and also show how attention improves interpretability of neural models. Finally, we discuss some applications in which modeling attention has a significant impact. We hope this survey will provide a succinct introduction to attention models and guide practitioners while developing approaches for their applications.

论文主要工作

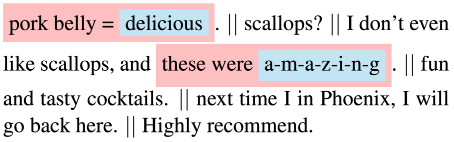
1 Introduction

Attention Model(AM), first introduced for Machine Translation [Bahdanau *et al.*, 2014] has now become a predominant concept in neural network literature. Attention has become enormously popular within the Artificial Intelligence(AI) community as an essential component of neural architectures for a remarkably large number of applications in Natural Language Processing, Statistical Learning, Speech and Computer Vision.

The intuition behind attention can be best explained using human biological systems. For example, our visual processing system tends to focus selectively on parts of the image, while ignoring other irrelevant information in a manner that can assist in perception [Xu *et al.*, 2015]. Similarly, in several problems involving language, speech or vision, some parts of the input can be more relevant compared to others. For instance, in translation and summarization tasks, only certain words in the input sequence may be relevant for predicting the next word. Likewise, in an image captioning problem, some regions of the input image may be more relevant for generating the next word in the caption. AM incorporates this notion of relevance by allowing the model to dynamically *pay attention to* only certain parts of the input that help in performing the task at hand effectively. An example of sentiment classification of Yelp reviews using AM is shown in Figure 1 [Yang *et al.*, 2016]. In this example, the AM learns that out of five sentences, the first and third sentences are more relevant. Furthermore, the words *delicious* and *amazing* within those sentences are more meaningful to determine the sentiment of the review.

有选择地聚焦在图像的某些部分上，同时以有助于感知的方式忽略其他无关信息

通过允许模型动态关注输入的某些部分来帮助有效地形成手头的任务，从而将这种相关性概念结合起来



pork belly = delicious . || scallops? || I don't even like scallops, and these were a-m-a-z-i-n-g . || fun and tasty cocktails. || next time I in Phoenix, I will go back here. || Highly recommend.

Figure 1: Example of attention modeling in sentiment classification of Yelp reviews. Figure from [Yang *et al.*, 2016]
.

The rapid advancement in modeling attention in neural networks is primarily due to three reasons. First, these models are now the state-of-the-art [Young *et al.*, 2018] for multiple tasks such as Machine Translation, Question Answering, Sentiment Analysis, Part-of-Speech tagging, Constituency Parsing and Dialogue Systems. Second, they offer several other advantages beyond improving performance on the main task. They have been extensively used for improving interpretability of neural networks, which are otherwise considered as black-box models. This is a notable benefit mainly because of growing interest in the fairness, accountability, and transparency of Machine Learning models in applications that influence human lives. Third, they help overcome some challenges with Recurrent Neural Networks(RNNs) such as performance degradation with increase in length of the input (Section 2) and the computational inefficiencies resulting from sequential processing of input (Section 4.3). Therefore, in this work we aim to provide a brief, yet comprehensive survey on attention modeling.

除了提高主要任务的性能外，它们还具有其他一些优点。它们已被广泛用于改善神经网络的可互换性，否则神经网络被视为黑盒模型。这是一个显着的好处，主要是因为人们对影响人类生活的应用中机器学习模型的公平性，责任性和透明度越来越感兴趣。

它们有助于克服回归神经网络（RNN）带来的一些挑战，例如性能下降，输入长度增加（第2节）以及输入顺序处理导致的计算效率低下

Organization: We briefly explain the AM proposed by Bahdanau *et al.* [2014] in Section 2 and describe our taxonomy in Section 3. We then discuss key neural architectures using AM and how attention is facilitating the interpretability of neural networks in Section 4 and 5 respectively. Finally, we present applications where attention has been widely applied in Section 6 and conclude the paper in Section 7.

文章目录结构

Related surveys: There have been a few surveys on attention focusing on Computer Vision [Wang and Tax, 2016], and graphs [Lee *et al.*, 2018]. Another similar work is by Galassi *et al.* [2019], but we further incorporate an accessible taxonomy, key architectures and applications, and interpretability aspect of AM. We hope that our contributions will not only foster broader understanding of AM but also help AI developers & engineers to determine the right approach for their application domain.

[*] Corresponding Author

2 Attention Model

A sequence-to-sequence model consists of an encoder-decoder architecture [Cho *et al.*, 2014b] as shown in Figure 2(a). The encoder is an RNN that takes an input sequence of tokens $\{x_1, x_2, ..., x_T\}$, where T is the length of input sequence, and encodes it into fixed length vectors $\{h_1, h_2, ..., h_T\}$. The decoder is also an RNN which then takes a single fixed length vector h_T as its input and generates an output sequence $\{y_1, y_2, ..., y_{T'}\}$ token by token, where T' is the length of output sequence. At each position t, h_t and s_t denote the hidden states of the encoder and decoder respectively.

Encoder 压缩过程的信息丢失

编码器必须将所有输入信息压缩成传递给解码器的单个固定长度矢量hT。

使用单个固定长度矢量来压缩长而详细的输入序列可能会导致信息丢失

Decoder 解码时：没有对齐功能

它无法模拟输入和输出序列之间的对齐，这是结构化输出任务（如翻译或摘要）的一个重要方面

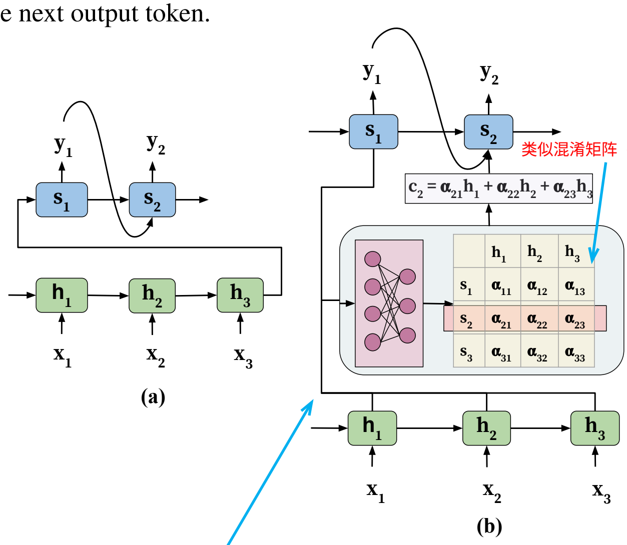
Challenges of traditional encoder-decoder: There are two well known challenges with this traditional encoder-decoder framework. First, the encoder has to compress all the input information into a single fixed length vector h_T that is passed to the decoder. Using a single fixed length vector to compress long and detailed input sequences may lead to loss of information [Cho *et al.*, 2014a]. Second, it is unable to model alignment between input and output sequences, which is an essential aspect of structured output tasks such as translation or summarization [Young *et al.*, 2018]. Intuitively, in sequence-to-sequence tasks, each output token is expected to be more influenced by some specific parts of the input sequence. However, decoder lacks any mechanism to selectively focus on relevant input tokens while generating each output token.

核心思想：

通过降低解码器访问整个编码输入序列{h1，h2，. . . ，hT}来减轻这些挑战。

在输入序列上引起注意权重α，以优先考虑存在相关信息的位置集合，以生成下一个输出Token

Key idea: AM aims at mitigating these challenges by allowing the decoder to access the entire encoded input sequence $\{h_1, h_2, ..., h_T\}$. The central idea is to induce attention weights α over the input sequence to prioritize the set of positions where relevant information is present for generating the next output token.



类似混淆矩阵 {Hi--Sj}

Figure 2: Encoder-decoder architecture: (a) traditional (b) with attention model

Attention的使用

Usage of attention: The corresponding encoder-decoder architecture with attention is shown in Figure 2(b). The attention block in the architecture is responsible for automatically learning the attention weights α_{ij}, which capture the relevance between h_i (the encoder hidden state, which we refer to as candidate state) and s_j (the decoder hidden state, which we refer to as query state). These attention weights are then used for building a context vector c, which is passed as an input to the decoder. At each decoding position j, the context vector c_j is a weighted sum of all hidden states of the encoder and their corresponding attention weights, i.e. $c_j = \sum_{i=1}^{T} \alpha_{ij} h_i$. This additional context vector is the mechanism by which decoder can access the entire input sequence and also focus on the relevant positions in the input sequence.

该前馈网络学习特定注意力量αij作为两个状态的函数，hi（候选状态）和sj-1（查询状态），其被神经网络作为输入。

Learning attention weights: The attention weights are learned by incorporating an additional feed forward neural network within the architecture. This feed forward network learns a particular attention weight α_{ij} as a function of two states, h_i (candidate state) and s_{j-1} (query state) which are taken as input by the neural network. Further, this feed forward network is jointly trained with encoder-decoder components of the architecture.

3 Taxonomy of Attention

Attention 分类

We consider attention in four broad categories and elucidate the different types of attention within each category as shown in Table 1[1]. We would like to emphasize that these categories are not mutually exclusive. Attention can be applied as a combination of multiple categories eg. a multi-level, self and soft attention combination has been used by Yang *et al.* [2016]. Hence, one can think of these categories as dimensions along which attention can be considered while employing it for an application of interest. To make this concept comprehensible, we provide a list of key technical papers and specify the multiple types of attention used within the proposed approaches in Table 2.

分类之间不互斥

Category	Type
Number of Sequences	distinctive, co-attention, self
Number of Abstraction Levels	single-level, multi-level
Number of Positions	soft/global, hard, local
Number of Representations	multi-representational, multi-dimensional

Table 1: Categories and types of attention within each category.

方面<1>根据序列的数量

3.1 Number of sequences

独特注意力：单输入

候选和查询状态分别属于两个不同的输入和输出序列时使用

用处：摘要，图像字幕，语音识别，

Thus far we have only considered the case which involves a single input and corresponding output sequence. This type of attention, which we refer to as **distinctive**, is used when candidate and query states belong to two distinct input and output sequences respectively. Most attention models employed for translation [Bahdanau *et al.*, 2014], summarization [Rush *et al.*, 2015], image captioning [Xu *et al.*, 2015] and speech recognition [Chan *et al.*, 2016] fall within the distinctive type of attention.

共同关注：多输入：多个输入序列同时共同学习他们的注意力权重，捕获这些输入之间的相互作用。

A **co-attention** model operates on multiple input sequences at the same time and jointly learns their attention weights, to capture interactions between these inputs. Lu *et al.* [2016] use a co-attention model for visual question answering. The authors argue that in addition to modeling visual attention on the input image, it is also important to model question attention because all words in the text of question

[1] Given the space constraints, we cannot cite all relevant papers, so we aim to cover a representative sample that outlines the scope of the field.

（1）Attention Block：自动学习注意力权重αij（αij捕捉了Hi与Si之间的 相关度）
（2）使用所有注意力权重组合成全局上下文向量C --》作为Decoder输入
（3）结合之前全局上下文+该附加上下文向量--》解码器可以访问整个输入序列并且还关注输入序列中的相关位置的机制

Reference	Application	Category			
		Number of Sequences	Number of Abstraction Levels	Number of Representations	Number of Positions
Bahdanau *et al.* [2014]	Machine Translation	distinctive	single-level	-	soft
Xu *et al.* [2015]	Image Captioning	distinctive	single-level	-	hard
Luong *et al.* [2015]	Machine Translation	distinctive	single-level	-	local
Yang *et al.* [2016]	Document Classification	self	multi-level	-	soft
Chan *et al.* [2016]	Speech Recognition	distinctive	single-level	-	soft
Lu *et al.* [2016]	Visual Question Answering	co-attention	multi-level	-	soft
Wang *et al.* [2017]	Sentiment Classification	co-attention	multi-level	-	soft
Ying *et al.* [2018]	Recommender Systems	self	multi-level	-	soft
Shen *et al.* [2018]	Language Understanding	self	single-level	multi-dimensional	soft
Kiela *et al.* [2018]	Sentence Representation	self	single-level	multi-representational	soft

Table 2: Summary of key papers for technical approaches in AMs. '-' means not applicable.

此外，基于注意力的图像表示用于引导问题注意，反之亦然。
这实质上有助于同时检测问题中的关键短语以及与答案相关的图像的相应区域。

are not equally important to the answer of the question. Further, attention based image representation is used to guide the question attention and vice versa, which essentially helps to simultaneously detect key phrases in the question and corresponding regions of images relevant to the answer.

自我Attention：输入序列，输出不是序列的任务：（比如文本分类）

In contrast, for tasks such as text classification and recommendation, input is a sequence but the output is not a sequence. In this scenario, attention can be used for learning relevant tokens in the input sequence for every token in the *same* input sequence. In other words, the query and candidate states belong to the same sequence for this type of attention. For this purpose, **self attention**, also known as inner attention has been proposed by Yang *et al.* [2016].

Attention 可以用作学习输入序列每个TOKEN 与其他TOKEN的相关度--》Self Att

<方面2>根据抽象的级别 （E.g:字，词，语句...）

3.2 Number of abstraction levels

In the most general case, attention weights are computed only for the original input sequence. This type of attention can be termed as **single-level**. On the other hand, attention can be applied on multiple levels of abstraction of the input sequence in a *sequential* manner. The output (context vector) of the lower abstraction level becomes the query state for the higher abstraction level. Additionally, models that use **multi-level attention** can be further classified based on whether the weights are learned top-down [Zhao and Zhang, 2018] (from higher level of abstraction to lower level) or bottom-up [Yang *et al.*, 2016].

仅针对原始输入序列计算注意力权重

注意力可以按顺序方式应用于输入序列的多个抽象级别。 较低抽象级别的输出（上下文向量）成为较高抽象级别的查询状态

可以根据从上到下各个级别的抽象进一步分类

We illustrate a key example in this category which uses the attention model at two different levels of abstraction, i.e. at word level and sentence level, for the document classification task [Yang *et al.*, 2016]. This model is called a "Hierarchical Attention Model"(HAM) because it captures the natural hierarchical structure of documents i.e. document is made up of sentences and sentences are made up of words. The multi-level attention allows the HAM to extract words that are important in a sentence and sentences that are important in a document as follows. It first builds an attention based representation of sentences with first level attention applied on sequence of word embedding vectors. Then it aggregates these sentence representations using a second level attention to form a representation of document. This final representation of the document is used as a feature vector for the classification task.

E,G：根据语句级和字符级进行文章分类（HAM层次注意力）。

它捕获文档的自然分层结构，即文档由句子组成，句子由单词组成。

层次步骤：

最后的表示 作为特征输入到任务中

Note that the co-attention work [Lu *et al.*, 2016] described in Section 3.1 also belongs to multi-level category where it co-attends to the image and question at three levels: word level, phrase level and question level. This combination of co-attention and multi-level attention is depicted in Figure 3.

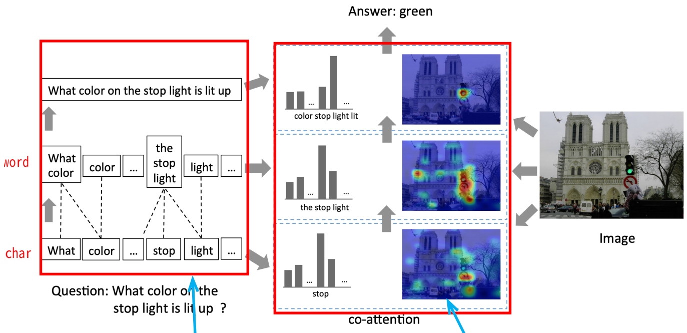


Figure 3: The AM proposed by Lu *et al.* [2016] for Visual Question Answering task which is a combination of co-attention (visual and text) and multi-level (word level, phrase level and question level) attention.

Zhao and Zhang [2018] propose to use "attention-via-attention", which also uses multi-level attention (with characters on the lower level and words on the higher level) but learns the attention weights in top-down fashion.

<方面3>位置：差异来自计算注意力函数的输入序列的位置。

3.3 Number of positions

In the third category, the differences arise from positions of the input sequence where attention function is calculated. The attention introduced by Bahdanau *et al.* [2014] is also known as **soft attention**. As the name suggests, it uses a weighted average of all hidden states of the input sequence to build the context vector. The usage of the soft weighing method makes the neural network amenable to efficient learning through backpropagation, but also results in quadratic computational cost.

它使用输入序列的所有隐藏状态的加权平均来构建上下文向量。

使神经网络适合通过反向传播进行有效学习，但也导致二次计算成本

Hard Att:根据输入序列中随机采样的隐藏状态计算上下文向量，这是使用由注意权重参数化的多因素分布来完成的

缺点：由于计算成本降低，模型是有益的，但是在输入的每个位置做出一个艰难的决策会使得结果框架不可微并且难以优化

因此，文献中已经提出了用于强化学习的各种学习方法和政策梯度方法来克服这种限制

Xu *et al.* [2015] propose a **hard attention** model in which the context vector is computed from stochastically sampled hidden states in the input sequence. This is accomplished using a multinoulli distribution parameterized by the attention weights. The hard attention model is beneficial due to decreased computational cost, but making a hard decision at every position of the input renders the resulting framework non-differentiable and difficult to optimize. As a result, variational learning methods and policy gradient methods in reinforcement learning have been proposed in the literature to overcome this limitation.

GLobal：类似于soft AM

Local AM:介于 Hard，sotf之间

关键思想：是首先检测输入序列中的注意点或位置，并在该位置周围选择一个窗口以创建局部软注意模型。

输入序列中的位置可以设置（单调对齐）或通过预测函数（预测对齐）学习。

优点：在窗口内提供软注意力和硬注意力之间的参数权衡，计算效率和可区分性

Luong *et al.* [2015] propose two attention models, namely **local** and **global**, in context of machine translation task. The global attention model is similar to the soft attention model. The local attention model, on the other hand, is intermediate between soft and hard attention. The key idea is to first detect an attention point or position within the input sequence and pick a window around that position to create a local soft attention model. The position within input sequence can either be set (monotonic alignment) or learned by a predictive function (predictive alignment). Consequently, the advantage of local attention is to provide a parametric trade-off between soft and hard attention, computational efficiency and differentiability within the window.

<4> 表示的数量：多表示AM+多维度AM—》组合成更好的表示/维度表示。

3.4 Number of representations

在某些场景中，使用输入的一个特征表示可能不足以完成下游任务，使用AM对多个特征表示组合成更好表示

注意力可以用于为这些不同的表示分配重要权重，这可以确定最相关的方面，忽略输入中的噪声和冗余。

它可以确定下游应用的输入的多个表示的相关性。 最终表示是这些多个表示及其注意权重的加权组合

优点：通过检查权重来直接评估哪些嵌入对于哪些特定下游任务是优选的。

Generally a single feature representation of the input sequence is used in most applications. However, in some scenarios, using one feature representation of input may not suffice for the downstream task. In this case, one approach is to capture different aspects of the input through multiple feature representations. Attention can be used to assign importance weights to these different representations which can determine the most relevant aspects, disregarding noise and redundancies in the input. We refer to this model as **multi-representational AM**, as it can determine the relevance of multiple representations of the input for downstream application. The final representation is a weighted combination of these multiple representations and their attention weights. One benefit of attention here is to directly evaluate which embeddings are preferred for which specific downstream tasks, by inspecting the weights.

E.g：AM+多种表示--》更好的表示

(1) AM+多种词嵌入---》更好语句表示

Kiela *et al.* [2018] learns attention weights over different word embeddings of the same input sentence to improve sentence representations. Similarly, Maharjan *et al.* [2018] use attention to dynamically weigh different feature representations of books capturing lexical, syntactic, visual and genre information.

引入权重以确定输入嵌入向量的每个维度的相关性

计算向量的每个特征的分数可以选择能够在任何给定上下文中最好地描述Token的特定含义的特征。

Based on similar intuition, in **multi-dimensional** attention, weights are induced for determining the relevance of each dimension of the input embedding vector. The intuition is that computing a score for each feature of the vector can select the features that can best describe the token's specific meaning in any given context. This is especially useful for natural language applications where word embeddings suffer from the polysemy problem. Examples of this approach are shown in Lin *et al.* [2017] for more effective sentence embedding representation and in Shen *et al.* [2018] for language understanding problem.

4 Network Architectures with Attention

In this section we describe three salient neural architectures used in conjunction with attention: (1) the encoder-decoder framework, (2) memory networks which extend attention beyond a single input sequence, and (3) architectures which circumvent the sequential processing component of recurrent models with the use of attention.

4.1 Encoder-Decoder

The earliest use of attention was as part of RNN based encoder-decoder framework to encode long input sentences [Bahdanau *et al.*, 2014]. Consequently, attention has been most widely used with this architecture.

An interesting fact is that AM can take any input representation and reduce it to a single fixed length context vector to be used in the decoding step. Thus, it allows one to decouple the input representation from the output. One could exploit this benefit to introduce hybrid encoder-decoders, the most popular being Convolutional Neural Network(CNN) as an encoder, and RNN or Long Short Term Memory (LSTM) as the decoder. This type of architecture is particularly useful for many multi-modal tasks such as image and video captioning, visual question answering and speech recognition.

并非所有输入和输出都是顺序数据的问题都可以通过上述公式解决，是另一类神经模型：

两个不同点：

(1) 输出是离散的，并指向输入序列中的位置（因此名称指针网络）

(2) 输出每一步的目标类数取决于输入的长度（因此也是变量）。

解决：使用传统的编码器 - 解码器框架无法实现这一点，其中输出字典是先验已知的

However, not all problems where both input and output are sequential data can be solved with the aforementioned formulation (e.g. sorting or travelling salesman problem). *Pointer networks* [Vinyals *et al.*, 2015] are another class of neural models with the following two differences, (1) the output is discrete and points to positions in the input sequence (hence the name pointer network), and (2) the number of target classes at each step of the output depends on the length of the input (and hence variable). This cannot be achieved using the traditional encoder-decoder framework where the output dictionary is known apriori (eg. in case of natural language modeling). The authors achieve this using attention weights to model the probability of choosing the i^{th} input symbol as the selected symbol at each output position. This approach can be applied to discrete optimization problems such as travelling salesperson problem, and sorting.

4.2 Memory Networks

问答和聊天机器人等应用程序需要能够从事实数据库中学习信息。

网络的输入是知识数据库和查询，其中一些事实与查询的相关性高于其他事实

Applications like question answering and chat bots require the ability to learn from information in a database of facts. The input to the network is a knowledge database and a query, where some facts are more relevant to the query than others.

通过使用一组内存block来存储事实数据库，并使用注意力来模拟内存中每个事实的相关性来回答查询，从而实现这一

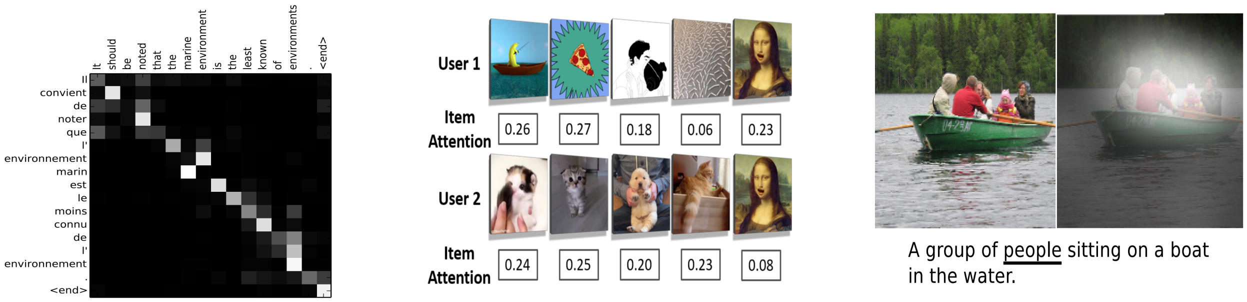
使用注意力还通过使目标连续并通过反向传播实现端到端训练来提供计算优势。

端到端存储器网络可以被认为是AM的概括，其中不是仅仅针对单个序列建模关注，而是通过大型序列数据库对其进行建模（事实）

End-to-End Memory Networks [Sukhbaatar *et al.*, 2015] achieve this by using an array of memory blocks to store the database of facts, and using attention to model relevance of each fact in the memory for answering the query. Using attention also provides computational advantage by making the objective continuous and enabling end-to-end training via backpropagation. End-to-End Memory Networks can be considered as a generalization of AM, wherein instead of modeling attention only over a single sequence they model it over a large database of sequences (facts).

4.3 Networks without RNNs

Recurrent architectures rely on sequential processing of input at the encoding step which results in computational inefficiency, as the processing cannot be parallelized [Vaswani *et al.*, 2017]. To address this, the authors propose *Transformer* architecture where the encoders and decoders are composed



(a) Alignment of French and English sentences in MT [Bahdanau *et al.*, 2014]

(b) Weights of items in user's history for recommendation [He *et al.*, 2018]

(c) Relevant image regions for image captioning [Xu *et al.*, 2015]

Figure 4: Examples of visualization of attention weights.

of a stack of identical layers with two sub-layers: position-wise Feed Forward Network(FFN) layer and multi-head self attention layer.

Position-wise FFN: The input is sequential which demands the model to make use of the temporal aspect of the input, yet components that capture this positional information (i.e. RNNs / CNNs) are not used. To account for this, the encoder phase in the Transformer generates content embedding as well as position encoding for each token of the input sequence using position-wise FFN.

Multi-Head Self-Attention: The self attention is used within each sub-layer to relate tokens and their positions within the same input sequence. Further, attention is known as multi-head, because several attention layers are stacked in parallel, with different linear transformations of the same input. This helps the model to capture various aspects of the input and improves its expressiveness.

Transformer architecture achieves significant parallel processing, shorter training time and higher accuracy for translation without any recurrent component, which is a notable benefit. However, the position encoding only weakly incorporates position information and might not work for problems that are more sensitive to positional variation. Shen *et al.* [2018] use temporal convolutions to encode positional information along with the self-attention mechanism of the transformer.

Additionally, there are more straightforward methods to break the sequential processing of input. Raffel and Ellis [2015] propose *Feed Forward Attention* models where they use AM to collapse the temporal dimension of data and use FFNs instead of RNNs to solve sequential data problems. In this scenario, AM is employed to produce a fixed length context vector from the variable length input sequence, which can be fed as an input to FFN.

5 Attention for Interpretability

There is a huge interest in the interpretability of AI models-driven by both performance as well as transparency and fairness of models[2]. However, neural networks, particularly deep learning architectures have been criticized for their lack of interpretability [Guidotti *et al.*, 2018].

Modeling attention is particularly interesting from the perspective of interpretability because it allows us to directly inspect the internal working of the deep learning architectures. The hypothesis is that the magnitude of attention weights highly correlates with how relevant a specific region of input is, for the prediction of output at each position in a sequence. This can be easily accomplished by visualizing the attention weights for a set of input and output pairs. Li *et al.* [2016] uphold attention as one of the important ways to explain inner workings of neural models.

As shown in Figure 4(a), Bahdanau *et al.* [2014] visualize attention weights which clearly show automatic alignment of sentences in French and English, despite the fact that subject-verb-noun locations differ from language to language. In particular, attention model shows non-monotonic alignment by correctly aligning *environnement marin* with *marine environment*. Figure 4(b) shows attention weights can help to recognize user's interests. User 1 seems to have a preference for "cartoon" videos , while user 2 prefers videos on "animals" [He *et al.*, 2018]. Finally, Xu *et al.* [2015] provide extensive list of visualizations of the relevant image regions (i.e. with high attention weights) which had a significant impact on the generated text in the image captioning task (example shown in Figure 4(c)).

We also summarize a few other interesting findings as follows. De-Arteaga *et al.* [2019] explore gender bias in occupation classification, and show how the words getting more attention during classification task are often gendered. Yang *et al.* [2016] note that the importance of words *good* and *bad* is context dependent for determining the sentiment of the review. The authors inspect the attention weight distribution of these words to find that they span from 0 to 1 which means the model captures diverse context and assign context-dependent weight to the words. Chan *et al.* [2016] note that in speech recognition, attention between character output and audio signal can correctly identify start position of the first character in audio signal and attention weights are similar for words with acoustic similarities. Finally, Kiela *et al.* [2018] find that the multi-representational attention assigns higher weights to GloVe, FastText word embeddings, particularly GloVe for low frequency words.

As another interesting application of attention, Lee *et al.* [2017] and Liu *et al.* [2018] provide a tool for visualizing attention weights of deep-neural networks. The goal is to interpret and perturb the attention weights so that one can simulate what-if scenarios and observe the changes in predictions interactively.

[2] https://fatconference.org

6 Applications

Attention models have become an active area of research because of their intuition, versatility and interpretability. Variants of attention models have been used to address unique characteristics of a diverse set of application domains eg. summarization, reading comprehension, language modeling, parsing etc. We discuss attention modeling in three application domains: (i) Natural Language Generation(NLG), (ii) Classification, and (iii) Recommender systems.

NLG tasks involve generating natural language text as the output. Some NLG applications that have benefited from incorporating an AM include Machine Translation (MT), Question Answering (QA) and Multimedia Description (MD).

MT uses algorithms to translate text or speech from one language to another. Modeling attention in neural techniques for MT allows for better alignment of sentences in different languages, which is a crucial problem in MT. The advantage of the attention model also becomes more apparent while translating longer sentences [Bahdanau *et al.*, 2014]. Several studies including [Britz *et al.*, 2017] and [Tang *et al.*, 2018] have shown performance improvements in MT using attention.

QA problems have made use of attention to (i) better understand questions by focusing on relevant parts of the question [Hermann *et al.*, 2015], (ii) store large amount of information using memory networks to help find answers [Sukhbaatar *et al.*, 2015], and (iii) improve performance in visual QA task by modeling multi-modality in input using co-attention [Lu *et al.*, 2016].

MD is the task of generating a natural language text description of a multimedia input sequence which can be speech, image and video [Cho *et al.*, 2015]. Similar to QA, here attention performs the function of finding relevant acoustic signals in speech input [Chorowski *et al.*, 2015] or relevant parts of the input image [Xu *et al.*, 2015] to predict the next word in caption. Further, Li *et al.* [2017] exploit the temporal and spatial structures of videos using multi-level attention for video captioning task. The lower abstraction level extracts specific regions within a frame and higher abstraction level focuses on small subset of frames selectively.

Document Classification: As mentioned earlier in Section 3, classification problems mainly make use of self attention to build more effective document representations. Yang *et al.* [2016] use a multi-level self attention, whereas Lin *et al.* [2017] propose a multi-dimensional and Kiela *et al.* [2018] propose a multi-representational self attention model.

Sentiment Analysis: Similarly, in the sentiment analysis task, self attention helps to focus on the words that are important for determining the sentiment of input. A couple of approaches for aspect based sentiment classification by Wang *et al.* [2016] and Ma *et al.* [2018] incorporate additional knowledge of aspect related concepts into the model and use attention to appropriately weigh the concepts apart from the content itself. Sentiment analysis application has also seen multiple architectures being used with attention such as memory networks [Tang *et al.*, 2016] and Transformer [Ambartsoumian and Popowich, 2018; Song *et al.*, 2019].

Recommender Systems: AMs have also been extensively used in recommender systems for user profiling, i.e., assigning attention weights to interacted items of a user to capture long and short term interests in a more effective manner. This is intuitive because all interactions of a user are not relevant for the recommendation of an item and user's interests are transient as well as varied in the long and short time span. Multiple papers use self attention mechanism for finding the most relevant items in user's history to improve item recommendations either with collaborative filtering framework [He *et al.*, 2018; Shuai Yu, 2019], or within an encoder-decoder architecture for sequential recommendations [Kang and McAuley, 2018; Zhou *et al.*, 2018].

Recently attention has been used in novel ways which has opened new avenues for research. Some interesting directions include smoother incorporation of external knowledge bases, pre-training embeddings and multi-task learning, unsupervised representational learning, sparsity learning and prototypical learning i.e. sample selection.

7 Conclusion

In this survey we have discussed different ways in which attention has been formulated in the literature, and have attempted to provide an overview of various techniques by discussing a taxonomy of attention, key neural network architectures using attention, and application domains that have seen significant impact. We discussed how the incorporation of attention in neural networks has led to significant gains in performance, provided greater insight into neural network's inner working by facilitating interpretability, and also improved computational efficiency by eliminating sequential processing of input. We hope that this survey will provide a better understanding of the different directions in which research has been done on this topic, and how techniques developed in one area can be applied to other domains.

References

Artaches Ambartsoumian and Fred Popowich. Self-attention: A better building block for sentiment analysis neural network classifiers. *arXiv preprint arXiv:1812.07860*, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. In *EMNLP*, pages 1442–1451. ACL, September 2017.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*, pages 4960–4964. IEEE, 2016.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. ACL.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, Doha, Qatar, October 2014. ACL.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *NIPS*, pages 577–585, Cambridge, MA, USA, 2015. MIT Press.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *arXiv preprint arXiv:1901.09451*, 2019.

Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention, please! a critical review of neural attention models in natural language processing. *arXiv preprint arXiv:1902.02181*, 2019.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. Nais: Neural attentive item similarity model for recommendation. *IEEE TKDE*, 30(12):2354–2366, 2018.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *ICDM*, pages 197–206. IEEE, 2018.

Douwe Kiela, Changhan Wang, and Kyunghyun Cho. Dynamic meta-embeddings for improved sentence representations. In *EMNLP*, pages 1466–1477, 2018.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. Interactive visualization and manipulation of attention-based neural machine translation. In *EMNLP*, pages 121–126. ACL, 2017.

John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. Attention models in graphs: A survey. *arXiv preprint arXiv:1807.07984*, 2018.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

Xuelong Li, Bin Zhao, Xiaoqiang Lu, et al. Mam-rnn: Multi-level attention model based rnn for video captioning. In *IJCAI*, pages 2208–2214, 2017.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

Shusen Liu, Tao Li, Zhimin Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Visual interrogation of attention-based models for natural language inference and machine comprehension. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2018.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, Lisbon, Portugal, September 2015. ACL.

Yukun Ma, Haiyun Peng, and Erik Cambria. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI*, 2018.

Suraj Maharjan, Manuel Montes, Fabio A González, and Thamar Solorio. A genre-aware attention model to improve the likability prediction of books. In *EMNLP*, pages 3381–3391, 2018.

Colin Raffel and Daniel PW Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, Lisbon, Portugal, September 2015. ACL.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, 2018.

Min Yang Baocheng Li Qiang Qu Jialie Shen Shuai Yu, Yongbo Wang. Nairs: A neural attentive interpretable recommendation system. *The Web Conference(WWW)*, 2019.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*, 2019.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. In *EMNLP*, pages 214–224, Austin, Texas, November 2016. ACL.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *NIPS*, pages 2692–2700, Cambridge, MA, USA, 2015. MIT Press.

Feng Wang and David MJ Tax. Survey on the attention based rnn model and its applications in computer vision. *arXiv preprint arXiv:1601.06823*, 2016.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *EMNLP*, pages 606–615. ACL, 2016.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, 2016.

Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. Sequential recommender system based on hierarchical attention network. In *IJCAI*, pages 3926–3932. AAAI Press, 2018.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence magazine*, 13(3):55–75, 2018.

Shenjian Zhao and Zhihua Zhang. Attention-via-attention neural machine translation. In *AAAI*, 2018.

Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiusi Chen, and Jun Gao. Atrank: An attention-based user behavior modeling framework for recommendation. In *AAAI*, 2018.