Attention, please! A survey of neural attention models in deep learning

Article *in* Artificial Intelligence Review · March 2022

DOI: 10.1007/s10462-022-10148-x

CITATIONS
63

READS
1,027

2 authors:

Alana Correia
University of Campinas
5 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Esther Colombini
University of Campinas
69 PUBLICATIONS **274** CITATIONS
SEE PROFILE

Artificial Intelligence Review
https://doi.org/10.1007/s10462-022-10148-x

Attention, please! A survey of neural attention models in deep learning

Alana de Santana Correia[1] · **Esther Luna Colombini**[1]

Abstract
In humans, Attention is a core property of all perceptual and cognitive operations. Given our limited ability to process competing sources, attention mechanisms select, modulate, and focus on the information most relevant to behavior. For decades, concepts and functions of attention have been studied in philosophy, psychology, neuroscience, and computing. For the last 6 years, this property has been widely explored in deep neural networks. Currently, the state-of-the-art in Deep Learning is represented by neural attention models in several application domains. This survey provides a comprehensive overview and analysis of developments in neural attention models. We systematically reviewed hundreds of architectures in the area, identifying and discussing those in which attention has shown a significant impact. We also developed and made public an automated methodology to facilitate the development of reviews in the area. By critically analyzing 650 works, we describe the primary uses of attention in convolutional, recurrent networks, and generative models, identifying common subgroups of uses and applications. Furthermore, we describe the impact of attention in different application domains and their impact on neural networks' interpretability. Finally, we list possible trends and opportunities for further research, hoping that this review will provide a succinct overview of the main attentional models in the area and guide researchers in developing future approaches that will drive further improvements.

Keywords Survey · Attention mechanism · Neural networks · Deep learning · Attention models

The Coordination for the Improvement of Higher Education Personnel (CAPES) partially supported this research. This work was carried out within the scope of PPI-Softex with support from the MCTI, through the Technical Cooperation Agreement [01245.013778/2020-21]

✉ Alana de Santana Correia
alana.correia@ic.unicamp.br

Esther Luna Colombini
esther@ic.unicamp.br

[1] Laboratory of Robotics and Cognitive Systems (LaRoCS), Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, Campinas, SP, Brazil

Published online: 03 March 2022

1 Introduction

Attention is a behavioral and cognitive process of focusing selectively on a discrete aspect of information, whether subjective or objective, while ignoring other perceptible information (Colombini et al. 2014), playing an essential role in human cognition and the survival of living beings in general. In animals of lower levels in the evolutionary scale, it provides perceptual resource allocation allowing these beings to respond correctly to the environment's stimuli to escape predators and capture preys efficiently. In human beings, attention acts on practically all mental processes, from reactive responses to unexpected stimuli in the environment—guaranteeing our survival in the presence of danger—to complex mental processes, such as planning, reasoning, and emotions. Attention is necessary because, at any moment, the environment presents much more perceptual information than can be effectively processed, the memory contains more competing traits than can be remembered, and the choices, tasks, or motor responses available are much greater than can be dealt with (Chun et al. 2011).

At early sensorial processing stages, data is separated between sight, hearing, touch, smell, and taste. At this level, attention selects and modulates processing within each of the five modalities and directly impacts processing in the relevant cortical regions. For example, attention to visual stimuli increases discrimination and activates the relevant topographic areas in the retinotopic visual cortex (Tootell et al. 1998), allowing observers to detect contrasting stimuli or make more precise discriminations. In hearing, attention allows listeners to detect weaker sounds or differences in extremely subtle tones but essential for recognizing emotions and feelings (Woldorff et al. 1993). Similar effects of attention operate on the somatosensory cortex (Johansen-Berg and Lloyd 2000), olfactory cortex (Zelano et al. 2005), and gustatory cortex (Veldhuizen et al. 2007). In addition to sensory perception, our cognitive control is intrinsically attentional. Our brain has severe cognitive limitations—the number of items that can be kept in working memory, the number of choices that can be selected, and the number of responses that can be generated at any time are limited. Hence, evolution has favored selective attention concepts as the brain has to prioritize.

Long before contemporary psychologists entered the discussion on attention, James (1890) offered us a precise definition that has been, at least, partially corroborated more than a century later by neurophysiological studies. According to James, "Attention implies withdrawal from some things in order to deal effectively with others... Millions of items of the outward order are present to my senses which never properly enter into my experience. Why? Because they have no interest for me. My experience is what I agree to attend to. Only those items which I notice shape my mind—without selective interest, experience is an utter chaos." Despite being a subjective definition that dates back to empirical studies carried out in the 19th century, Jame's definition demonstrates the selective role of attention in the choice of perceptual information and its importance for human cognition. Since then, mainly the selective role of attention has been studied by researchers from different areas and has been fundamental for creating artificial attentional systems.

For the past decades, the concept of attention has permeated most aspects of research in perception and cognition, being considered as a property of multiple and different perceptual and cognitive operations (Colombini et al. 2014). Thus, to the extent that these mechanisms are specialized and decentralized, attention reflects this organization. These mechanisms are in wide communication, and the executive control processes help set priorities for the system. Selection mechanisms operate throughout the brain and are involved

in almost every stage, from sensory processing to decision making and awareness. Attention has become a broad term to define how the brain controls its information processing, and its effects can be measured through conscious introspection, electrophysiology, and brain imaging. Attention has been studied from different perspectives for a long time.

1.1 Pre-deep learning models of attention

Computational attention systems based on psychophysical models, supported by neurobiological evidence, have existed for at least three decades (Frintrop et al. 2010a). Treisman's Feature Integration Theory (FIT) (Treisman and Gelade 1980), Wolfe's Guides Search (Wolfe et al. 1989), Triadic architecture (Rensink 2000), Broadbent's Model (Broadbent 2013), Norman Attentional Model (Norman 1968; Kahneman 1973), Closed-loop Attention Model (Van der Velde et al. 2004), SeLective Attention Model (Phaf et al. 1990), among several other models, introduced the theoretical basis of computational attention systems. These models have essential attentional components gradually introduced into deep neural networks, such as feature integration, winner-take-all (WTA) mechanisms, feature selection, bottom-up, and top-down flows.

Initially, attention was mainly studied with visual experiments where a subject looks at a scene that changes in time (Frintrop et al. 2010b). In these models, the attentional system was restricted only to the selective attention component in visual search tasks, focusing on the extraction of multiple features through a sensor. Therefore, most of the attentional computational models occurred in computer vision to select important image regions. Koch and Ullman (1987) introduced the area's first visual attention architecture based on FIT (Treisman and Gelade 1980). The idea behind it is that several features are computed in parallel, and their conspicuities are collected on a salience map. WTA determines the most prominent region on the map, which is finally routed to the central representation. From then on, only the region of interest proceeds to more specific processing. Neuromorphic Vision Toolkit (NVT), derived from the Koch–Ullman (Itti et al. 1998) model, was the basis for developing research in computational visual attention for several years. Navalpakkam and Itti introduce a derivative of NVT which can deal with top-down cues (Navalpakkam and Itti 2006). The idea is to learn the target's feature values from a training image in which a binary mask indicates the target. The attention system of Hamker (2005, 2006) calculates various features and contrast maps and turns them into perceptual maps. With target information influencing processing, they combine detection units to determine whether a region on the perceptual map is a candidate for eye movement. VOCUS (Frintrop 2006) introduced a way to combine bottom-up and top-down attention, overcoming the limitations of the time. Several other models have emerged in the literature, each with peculiarities according to the task. Many computational attention systems focus on the computation of mainly three features: intensity, orientation, and color. These models employed neural networks or filter models that use classical linear filters to compute features.

Computational attention systems were used successfully before Deep Learning (DL) in object recognition (Salah et al. 2002), image compression (Ouerhani 2003), image matching (Walther 2006), image segmentation (Ouerhani 2003), object tracking (Walther et al. 2004), active vision (Clark and Ferrier 1988), human–robot interaction (Breazeal and Scassellati 1999), object manipulation in robotics (Rotenstein et al. 2007), robotic navigation (Clark and Ferrier 1992), and SLAM (Frintrop and Jensfelt 2008). In mid-1997, Scheier and Egner (1997) presented a mobile robot that uses attention for navigation. Still, in the 90s, Baluja and Pomerleau (1997) used an attention system to navigate

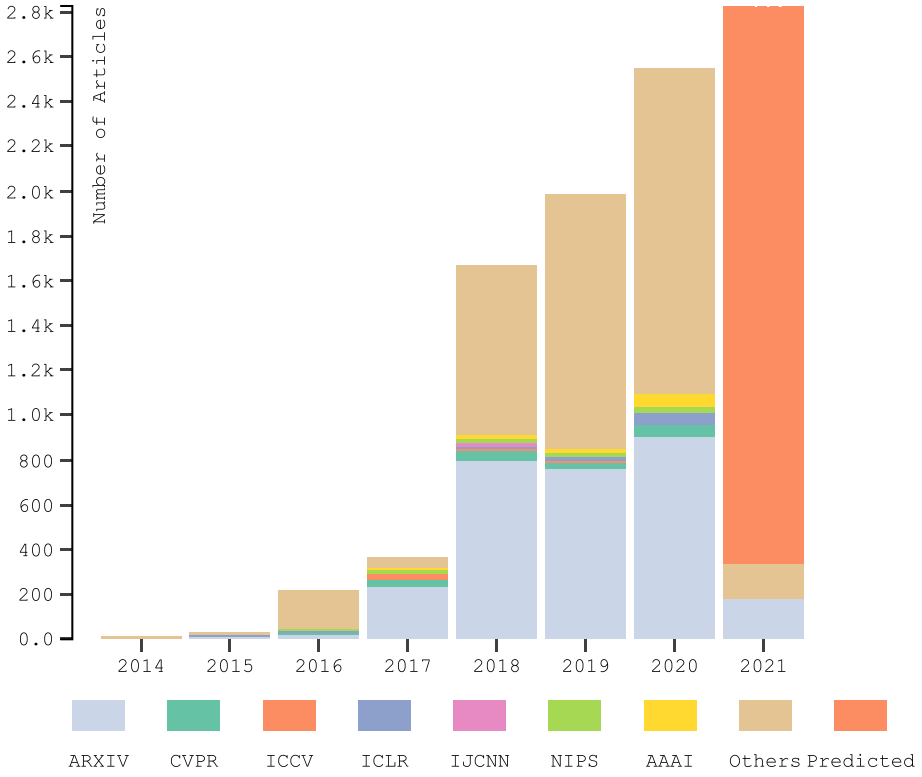

Fig. 1 Works published by year between 01/01/2014 to 15/02/2021. The main sources collected are ArXiv, CVPR, ICCV, ICLR, IJCNN, NIPS, and AAAI. The other category refers mainly to the following publishing vehicles: ICML, ACL, ACM, EMNLP, ICRA, ICPR, ACCV, CORR, ECCV, ICASSP, ICLR, IEEE ACCESS, Neurocomputing, and several other magazines

an autonomous car, which followed relevant regions of a projection map. Walther (2006) combined an attentional system with an object recognizer based on SIFT features and demonstrated that the attentional front-end enhanced the recognition results. Salah et al. (2002) combined attention with neural networks in an Observable Markov model for handwritten digit recognition and face recognition. Ouerhani (2003) proposed the focused image compression, which determines the number of bits to be allocated for encoding regions of an image according to their salience. High saliency regions have a high quality of reconstruction concerning the rest of the image.

1.2 Deep learning models of attention: the beginning

By 2014, the deep learning (DL) community noticed attention as a fundamental concept for advancing deep neural networks. These current researches are supported by many physicological theories and some previous computational attention systems that precede the DL era. As shown in Fig. 1, the number of published works in neural attention models grows each year significantly in the leading repositories. Currently, the state-of-the-art in the DL field uses attention. In neural networks, attention mechanisms dynamically manage the

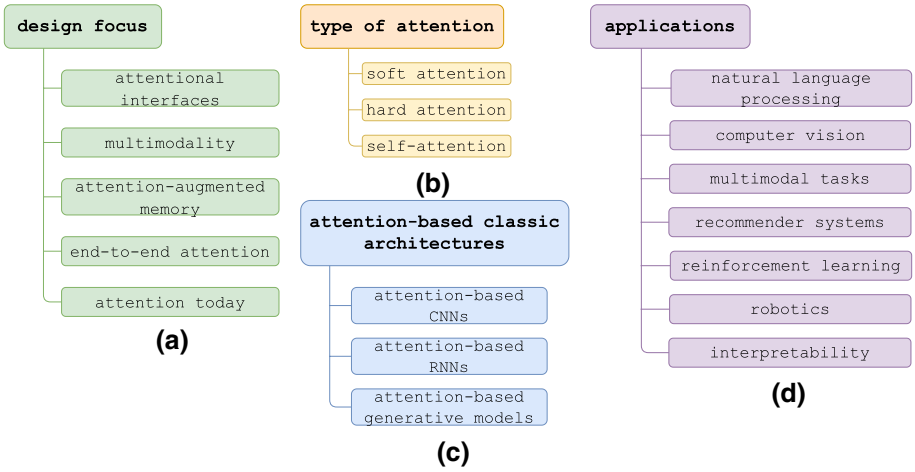

Fig. 2 Proposed perspectives to group neural attention models based on **a** design focus, **b** type of attention, **c** attention-based classic architectures, and **d** applications

flow of information, the features, and the resources available, improving learning. These mechanisms filter out irrelevant stimuli for the task and help the network to deal with long-time dependencies simply. Many neural attentional models are simple, scalable, flexible, and promising results in many application domains (Gregor et al. 2015; Vaswani et al. 2017; Weston et al. 2014). Given the current research extent, interesting questions related to neural attention models arise in the literature: how these mechanisms help improve neural networks' performance, which classes of problems benefit from this approach, and how these benefits arise.

To the best of our knowledge, most surveys available in the literature do not address all of these questions or are more specific to some domain. Wang and Tax (2016) propose a review on recurrent networks and applications in computer vision, Hu (2019), and Galassi et al. (2020) offer surveys on attention in natural language processing (NLP). Lee et al. (2019b) present a review on attention in graph neural networks, and Chaudhari et al. (2019) presented a more general, yet short, review. Differently, in this paper, our goal is to show a complete overview of the field over four different intuitive perspectives, as shown in Fig. 2. In particular, we group the existing works by their design focus (e.g., attentional interfaces, multimodality, attention-augmented memory, end-to-end attention, and attention today) to introduce the main strategies that apply attention in neural networks. This perspective helps the reader understand the area's main developments in chronological order, from the first attention model created to current developments. Our second perspective groups methods by training strategies, differentiable and attentional focus characteristics (e.g., soft attention, hard attention, and self-attention). The third perspective groups methods by DL architectures (e.g., attention-based CNNs, attention-based RNNs, and attention-based generative models), showing to the reader the following topics: (1) how attentional mechanisms have been implemented in the classic Deep Learning architectures, (2) what are the gains obtained by these mechanisms in different parts of the models, and finally (3) research insights that deserve to be investigated in classic architectures. We use the fourth and final perspective to provide a comprehensive survey of the field arranged by application and problem sets. By doing so, we aim to help the reader understand which problems have

already been addressed. Perhaps most importantly, it reveals critical applications and problems where attention models have not yet been applied.

We argue that these perspectives allow readers to learn about different neural attention methods from different perspectives. In general, the discussion on each perspective can be considered separately, with exclusive sections to discuss each one. Additionally, we summarize the challenges that have yet to be addressed in the field and provide promising directions for future work.

1.3 Contributions

To assess the breadth of attention applications in deep neural networks, we present a systemic review of the field in this survey. Throughout our review, we critically analyzed 650 papers while addressing quantitatively 6567 papers to extract different metrics to discover overall trends and plot some figures.

As the main contributions of our work, we highlight:

1. It is the first paper in the attention in DL literature created from such a comprehensive systematic review;
2. A replicable research methodology. We provide, in the Appendix, the detailed process conducted to collect our data, and we make available the scripts to collect the papers and create the graphs we use;
3. An in-depth overview of the field. We critically analyzed 650 papers and extracted different metrics from 6567, employing various visualization techniques to highlight overall trends in the area;
4. We present a comprehensive analysis of attentional models in deep learning from four different perspectives that are relevant when selecting an attention model. To the best of our knowledge, this is the first literature review article to feature a paper focused on the aspects we have chosen to analyze;
5. We present the main neural architectures that employ attention mechanisms, describing how they have contributed to the NN field, and we chronologically detail and highlight the main developments from 2014 to the present, presenting a complete overview for the reader;
6. We categorize attentional mechanisms concerning training strategies, differentiable and attentional focus characteristics;
7. We introduce how attentional modules or interfaces have been used in classic DL architectures extending the Neural Network Zoo diagrams. From this perspective, we highlight the benefits of attention when plugged into different parts of classic architectures. Over hundreds of papers, we have identified patterns related to the model's attention mechanism's location, the task being performed, and the desired improvements to the problem. Such patterns are challenging to identify due to the mechanisms variety and lack of standardization in the models. However, for each classic model type, we were able to create different groups to categorize each usage coherently;
8. We present the main application areas where attention is being used. We highlight which problems attentional mechanisms try to minimize in each area, and perhaps most importantly, which areas are less explored;
9. Finally, we present a broad description of application domains, trends, and research opportunities.

1.4 Organization of the survey

This survey is structured as follows. In Sect. 2 we present the field overview reporting the main events from 2014 to the present. Section 3 contains a description of attention main mechanisms. In Sect. 4 we analyze how attentional modules are used in classic DL architectures. Section 5 explains the main classes of problems and applications of attention. Finally, in Sect. 6 we discuss limitations, open challenges, current trends, and future directions in the area, concluding our work in Sect. 7 with directions for further improvements.

2 Overview

Historically, research in computational attention systems has existed since the 1980s. Only in mid-2014, the Neural Attentional Networks (NANs) emerged in natural language processing (NLP), where attention provided significant advances, bringing promising results through scalable and straightforward networks. Attention allowed us to move towards the complex tasks of conversational machine comprehension, sentiment analysis, machine translation, question-answering, and transfer learning, previously challenging. Subsequently, NANs appeared in other fields equally important for artificial intelligence, such as computer vision (CV), reinforcement learning (RL), and robotics. There are currently numerous attentional architectures, but few of them have a significantly higher impact, as shown in Fig. 3. In this image, we depict the most relevant group of works organized according to citation levels and innovations where RNNSearch (Bahdanau et al. 2015), Transformer (Vaswani et al. 2017), Memory Networks (Weston et al. 2014), "show, attend and tell" (Xu et al. 2015), and RAM (Mnih et al. 2014) stand out as key developments.

The *bottleneck problem* in the classic encoder–decoder framework worked as the initial motivation for attention research in Deep Learning. In this framework, the encoder encodes a source sentence into a fixed-length vector from which a decoder generates the translation. The main issue is that a neural network needs to compress all the necessary information from a source sentence into a fixed-length vector. Cho et al. (2014a) showed that the performance of the classic encoder–decoder deteriorates rapidly as the size of the input sentence increases. To minimize this bottleneck, Bahdanau et al. (2015) proposed RNNSearch, an extension to the encoder–decoder model that learns to align and translate together. RNNSearch generates a translated word at each time-step, looking for a set of positions in the source sentence with the most relevant words. The model predicts a target word based on the context vectors associated with those source positions and all previously generated target words. The main advantage is that RNNSearch does not encode an entire input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors, choosing a subset of these vectors adaptively while generating the translation. The attention mechanism allows extra information to be propagated through the network, eliminating the fixed-size context vector's information bottleneck. This approach demonstrated that the attentive model outperforms classic encoder–decoder frameworks for long sentences for the first time.

RNNSearch was instrumental in introducing the first attention mechanism, soft attention (Sect. 3). This mechanism has the main characteristic of smoothly selecting the network's most relevant elements. Based on RNNSearch, there have been numerous attempts to augment neural networks with new properties. Two research directions stand out as

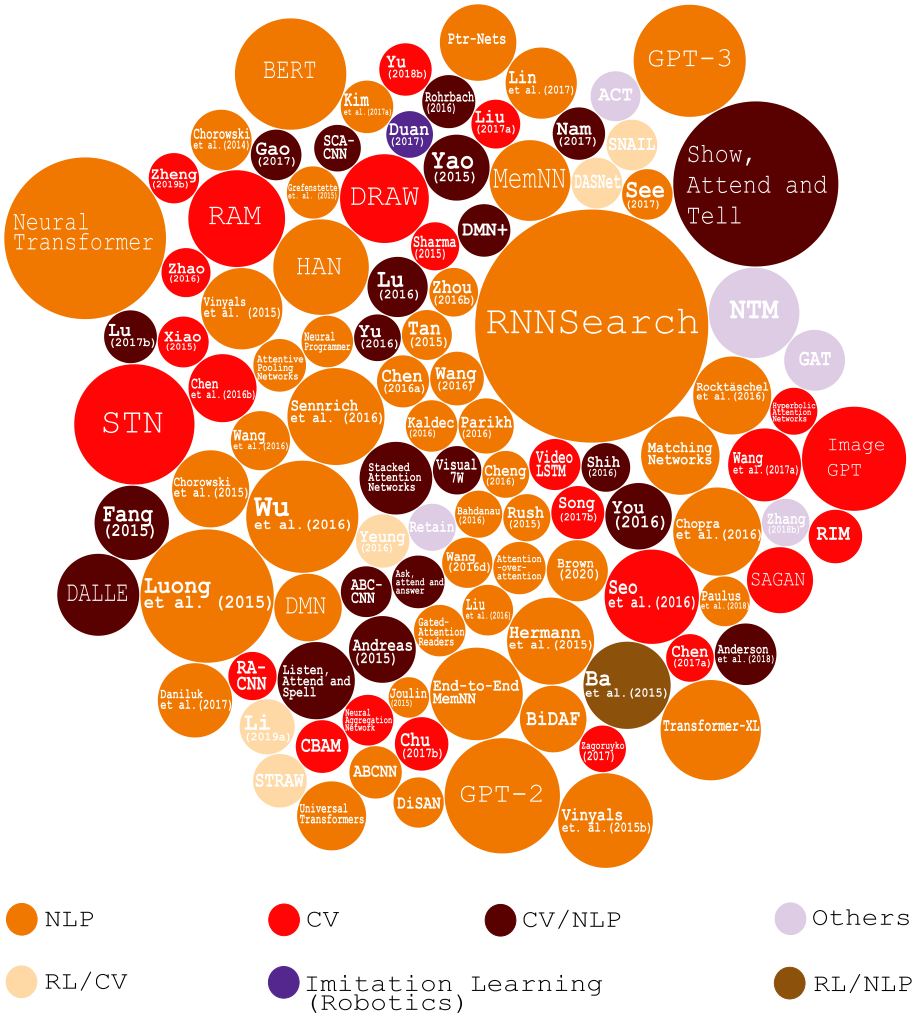

Fig. 3 Main neural attention networks (NAN). Each circle corresponds to an architecture. The radius of the circles is defined based on the impact of the NAN on the field. The impact was defined by the citation number and the architecture innovation level. The greater the radius of the circle, the more significant the impact of architecture, and vice versa. Architectures labels are color-coded as follows: orange—natural language processing, red—computer vision, dark brown—computer vision and natural language processing, dark yellow—reinforcement learning and computer vision, light yellow—reinforcement learning and natural language processing, blue—imitation learning and robotics, and purple—others. (Color figure online)

particularly interesting—attentional interfaces and end-to-end attention. Attentional interfaces treat attention as a module or set of elective modules, easily plugged into classic Deep Learning neural networks, just like RNNSearch. So far, this is the most explored research direction in the area, mainly for simplicity, general use, and the good results of generalization that the attentional interfaces bring. End-to-end attention is a younger research direction, where the attention block covers the entire neural network. High and low-level attentional layers act recursively or cascaded at all network abstraction levels to produce

the desired output in these models. End-to-end attention models introduce a new class of neural networks in Deep Learning. End-to-end attention research makes sense since no isolated attention center exists in the human brain, and its mechanisms are used in different cognitive processes. In this sense, this section goal is introduce a complete overview of the area, presenting the main strategies that have been proposed in the literature for applying attention. We categorize the main strategies into attentional interfaces (Sect. 2.1), multimodality (Sect. 2.2), attention-augmented memory (Sect. 2.3), end-to-end attention models (Sect. 2.4), and attention today (Sect. 2.5). In each section, we also present the events in chronological order to help understand the main developments from the first attention model created in 2014 to the present.

2.1 Attentional interfaces

RNNSearch is the basis for research on attentional interfaces, widely employed in several other applications. Classic soft attention RNNSearch is used in voice recognition (Chan et al. 2016) allowing one RNN to process the audio. At the same time, another examines it, focusing on the relevant parts as it generates a description. In text analysis (Vinyals et al. 2015b), it allows a model to look at the words as it generates an analysis tree. For conversational modeling (Vinyals and Le 2015), it allows the model to focus on the last parts of the conversation as it generates its response.

Despite RNNSearch success, there are similar versions computationally cheaper. Luong et al. (2015) proposed global and local attention for machine translation. Global attention is similar to RNNSearch's original soft attention but cheaper. Local attention, on the other hand, is the first mechanism in the area to make hard attention (Sect. 3) differentiable. To do so, the model first searches for a single position aligned to the current target word. Then, a window centered around the position is used to calculate a context vector.

There are also essential extensions to deal with other information bottlenecks in addition to the classic encoder–decoder problem, such as capturing insights about hierarchy in texts or documents. For this, Bidirectional Attention Flow (BiDAF) (Seo et al. 2017) proposes a multi-stage hierarchical process to question-answering. It uses the bidirectional attention flow to build a multi-stage hierarchical network with context paragraph representations at different granularity levels. The attention layer does not summarize the context paragraph in a fixed-length vector. Instead, attention is calculated for each step, and the vector assisted at each step, along with representations of previous layers, can flow to the subsequent modeling layer. This reduces the loss of information caused by the early summary. At each stage of time, attention is only a function of the query and the paragraph of the context in the current stage and does not depend directly on the previous stage's attention. The hypothesis is that this simplification leads to a work division between the attention layer and the modeling layer, forcing the attention layer to focus on learning attention between the query and the context.

Yang et al. (2016c) proposed the Hierarchical Attention Network (HAN) to capture two essential insights about document structure. Documents have a hierarchical structure: words form sentences, sentences form a document. Humans, likewise, construct a document representation by first building representations of sentences and then aggregating them into a document representation. Different words and sentences in a document are differentially informative. Moreover, the importance of words and sentences is highly context-dependent, i.e., the same word or sentence may have different importance in different contexts. To include sensitivity to this fact, HAN consists of two levels of attention

mechanisms—one at the word level and one at the sentence level—that let the model pay more or less attention to individual words and sentences when constructing the document's representation. Similarly, Xiong et al. (2017) created a coattentive encoder that captures the interactions between the question and the document with a dynamic pointing decoder that alternates between estimating the start and end of the answer span.

To learn approximate solutions to computationally intractable problems (i.e., convex hulls, computing Delaunay triangulations, and the planar Travelling Salesman Problem), Pointer Networks (Ptr-Net) (Vinyals et al. 2015a) modifies the RNNSearch's attentional mechanism to represent variable-length dictionaries. Instead of using attention to blend hidden units of an encoder to a context vector at each decoder step, it uses attention as a pointer to select a member of the input sequence as the output. This approach was fundamental to inspire current methods that use attention as a switch between layers (Choi et al. 2017). For example, See et al. (2017) used a hybrid between classic sequence-to-sequence attentional models and a Ptr-Net (Vinyals et al. 2015a) to abstractive text summarization. The hybrid pointer-generator (See et al. 2017) copies words from the source text via pointing, which aids an accurate reproduction of information while retaining the ability to produce novel words through the generator. Finally, it uses a mechanism to keep track of what has been summarized, which discourages repetition.

Attentional interfaces are also widely used to manage the internal memory of recurrent networks using two-away attention mechanisms and history-of-words strategies. FusionNet (Huang et al. 2018c) presents important research with a history-of-word strategy to characterize attention information from the lowest word-embedding level up to the highest semantic level representation. This approach considers that data input is gradually transformed into a more abstract representation, forming each word's history in human mental flow. FusionNet employs a fully-aware multi-level attention mechanism and an attention score function. It takes advantage of the history-of-word to capture efficiently long-range information using Bidirectional long short-term memory (BiLSTMs). Differently, Rocktäschel et al. (2016) introduce two-away attention, a new form to manage Long short-term memory (LSTM) cells for recognizing textual entailment (RTE) task. The mechanism allows the model to attend over past output vectors, solving the LSTM's cell state bottleneck. The LSTM with attention does not need to capture the premise's whole semantics in the LSTM cell state. Instead, attention generates output vectors while reading the premise and accumulating a representation in the cell state that informs the second LSTM which of the premises' output vectors to attend to determine the RTE class.

Due to its robustness to environment changes, in computer vision, attentional interfaces are mainly inspired by human saccadic movements. The human visual attention mechanism can explore local differences in an image while highlighting the relevant parts. One person focuses attention on parts of the image, glimpsing to quickly scan the entire image to find the main areas during the recognition process. In this process, the different regions' internal relationship guides the eyes' movement to find the next area to focus. Ignoring the irrelevant parts easies learning in the presence of disorder. Another advantage of glimpse and visual attention is its robustness. Our eyes can see an object in a real-world scene but ignore irrelevant parts. Convolutional neural networks (CNNs) are extremely different. CNNs are rigid, and the number of parameters grows linearly with the size of the image. Also, for the network to capture long-distance dependencies between pixels, the architecture needs to have many layers, compromising the model's convergence. Besides, the network treats all pixels in the same way. This process does not resemble the human visual system that contains visual attention mechanisms and a glimpse structure that provides unmatched performance in object recognition.

The Recurrent Attention Model (RAM) (Mnih et al. 2014) and Spatial Transformer Network (STN) (Jaderberg et al. 2015) are pioneering architectures with attentional interfaces based on human visual attention. RAM (Mnih et al. 2014) extracts information from an image or video by adaptively selecting a sequence of regions, glimpses, only processing the selected areas at high resolution. The model contains a recurrent neural network (RNN) that processes different parts of the images (or video frames) at each time t, building a dynamic internal representation of the scene via reinforcement learning training. In RAM, attention brings an essential contribution to neural networks: a sensorimotor mechanism that enables the model to move its glimpse to regions in the image crucial to learning. In contrast, classic neural networks receive a massive amount of data and do not explore them in small portions sequentially. Humans do not learn to receive a large stack of data simultaneously; instead, they experience the world gradually. Besides, RAM has a reduced parameter set, and its architecture is independent of the input image size, which does not occur in convolutional neural networks. This approach is generic and can use static images, videos, or a perceptual module of an agent that interacts with the environment.

To make visual models more invariant to transformations—a natural feature in the visual human system—the STN was created by Jaderberg et al. (2015). Suppose the input is transformed in this model. In that case, the model must generate the correct classification label, even if it is distorted in unusual ways. STN works as an attentional module attachable—with few modifications—to any neural network to actively spatially transform feature maps. STN learns transformation during the training process. Unlike pooling layers, where receptive fields are fixed and local, a Spatial Transformer is a dynamic mechanism that can spatially transform an image, or feature map, producing the appropriate transformation for each input sample. The transformation is performed across the map and may include changes in scale, cut, rotations, and non-rigid body deformations. This approach allows the network to select the most relevant image regions (attention) and transform them into a desired canonical position by simplifying recognition in the following layers.

Following the RAM approach, the Deep Recurrent Attentive Writer (DRAW) (Gregor et al. 2015) represents a change to a more natural way of constructing the image in which parts of a scene are created independently of the others. This process is how human beings draw a scene by recreating a visual scene sequentially, refining all parts of the drawing for several iterations, and reevaluating their work after each modification. Although natural to humans, most approaches to automatic image generation aim to generate complete scenes at once. This means that all pixels are conditioned in a single latent distribution, making it challenging to scale large image approaches. DRAW belongs to the family of variational autoencoders. It has an encoder that compresses the images presented during training and a decoder that reconstructs the images. Unlike other generative models, DRAW iteratively constructs the scenes by accumulating modifications emitted by the decoder, each observed by the encoder. DRAW uses RAM attention mechanisms to attend to parts of the scene while ignoring others selectively. This mechanism's main challenge is to learn where to look, which is usually addressed by reinforcement learning techniques. However, at DRAW, the attention mechanism is differentiable, making it possible to use backpropagation.

2.2 Multimodality

The first attention interfaces' use in DL were limited to NLP and CV domains to solve isolated tasks. Currently, attentional interfaces are studied in multimodal learning. Sensory

multimodality in neural networks is a historical problem widely discussed by the scientific community (Ramachandram and Taylor 2017; Gao et al. 2020b). Multimodal data improves the robustness of perception through complementarity and redundancy. The human brain continually deals with multimodal data and integrates it into a coherent representation of the world. However, employing different sensors present a series of challenges computationally, such as incomplete or spurious data, different properties (i.e. dimensionality or range of values), and the need for data alignment association. The integration of multiple sensors depends on a reasoning structure over the data to build a common representation, which does not exist in classical neural networks. Attentional interfaces adapted for multimodal perception are an efficient alternative for reasoning about misaligned data from different sensory sources.

The first widespread use of attention for multimodality occurs with the attentional interface between a convolutional neural network and an LSTM in image captioning (Xu et al. 2015). In this model, a CNN processes the image, extracting high-level features, whereas the LSTM consumes the features to produce descriptive words, one by one. The attention mechanism guides the LSTM to relevant image information for each word's generation, equivalent to the human visual attention mechanism. The visualization of attention weights in multimodal tasks improved the understanding of how architecture works. This approach derived from countless other works with attentional interfaces that deal with video-text data (Yao et al. 2015; Wu et al. 2018a; Fakoor et al. 2016), image-text data (Tian et al. 2018a; Pu et al. 2018), monocular/RGB-D images (Liu et al. 2017c; Zhang et al. 2018c, d), RADAR (Zhang et al. 2018c), remote sensing data (Zhang et al. 2019h; Fang et al. 2019a; Wang et al. 2018d; Mei et al. 2019), audio-video (Hori et al. 2017a; Zhang et al. 2019m), and diverse sensors (Zadeh et al. 2018a, b; Santoro et al. 2018), as shown in Fig. 4.

Despite the wide variety of sensors, most works are focused on sensory alignment or sensory fusion of visual-textual data. For alignment, Zhang et al. (2019p) used an adaptive attention mechanism to learn to emphasize different visual and textual sources for dialogue systems for fashion retail. An adaptive attention scheme automatically decided the evidence source for tracking dialogue states based on visual and textual context. Pu et al. (2018) introduced two mechanisms to video-text features alignment. The first mechanism adaptively and sequentially focuses on different layers of a CNN's features instead of focusing on a specific layer. At the same time, spatial-temporal attention focuses on essential regions in the selected layer. This approach helps minimize the loss of semantic meaning and spatial-temporal importance of the feature vectors to feed the RNN decoder. Dual Attention Networks (Nam et al. 2017) presented attention mechanisms to capture the fine-grained interplay between images and textual information. The mechanism allows visual and textual attention to guide each other during collaborative inference. Finally, Abolghasemi et al. (2019) demonstrated an approach for augmenting a deep visuomotor policy trained through demonstrations with Task Focused Visual Attention (TFA). Attention receives as input a manipulation task specified in natural language text, an image with the environment, and returns as output the area with an object that the robot needs to manipulate. The attention mechanism uses language encoding to align correspondent regions in the image, eliminating background and distractions. The main contribution of TFA relies on introducing multimodal tasks using reinforcement learning. TFA shows that attention can help models generalize better and learn more robust policies, which is still challenging.

Some approaches leverage the power of multimodal alignment to generate challenging samples during training. In this line, Liu et al. (2019e) proposed a cross-modal attention-guided erasing approach for referring expressions. Previous attention models focus on only the most dominant features of both modalities and neglect textual-visual correspondences

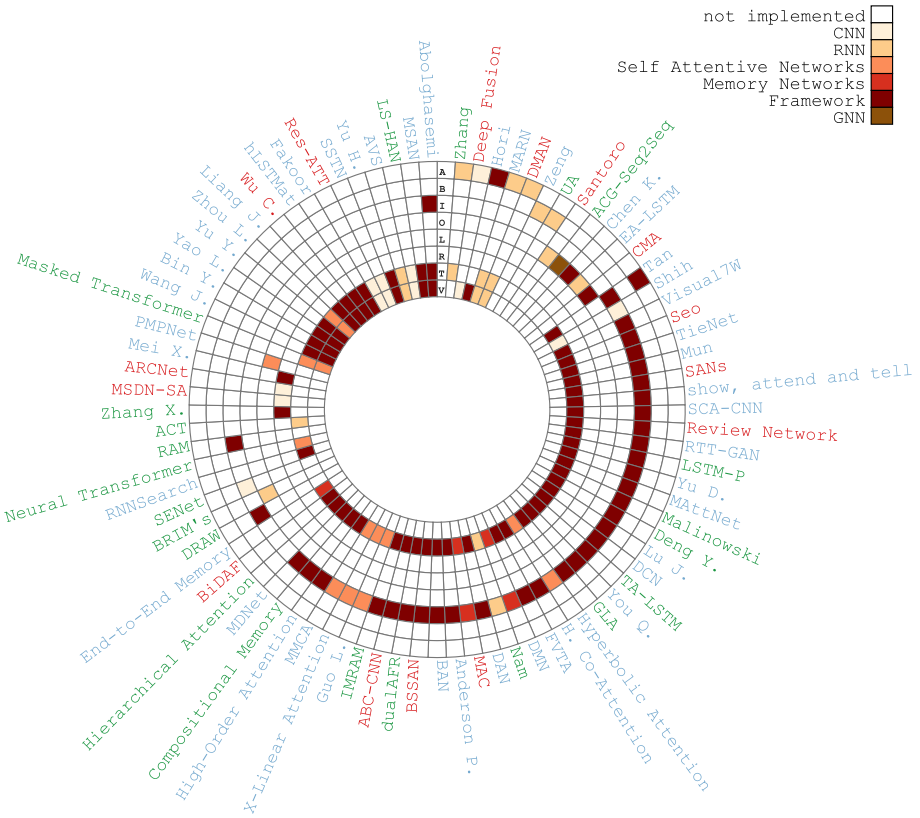

Fig. 4 A diagram showing sensory modalities of neural attention models. Radial segments correspond to attention architectures, and each track corresponds to a modality. Modalities are: (A) audio, (B) biomedical signals, (I) image, (O) other sensors, (L) LiDAR, (R) remote sensing data, (T) text, and (V) video. The following coloring convention is used for the individual segments: white (the modality is not implemented), light yellow (CNN), light orange (RNN), orange (Self-attentive networks), red (Memory networks), dark red (framework), and brown (GNN). This diagram emphasizes multimodal architectures so that only the most representative single modality (i.e., text or image) architectures are shown. Most multimodal architectures use the image/text or video/text modalities. (Color figure online)

between images and referring expressions. To tackle this issue, cross-modal attention discards the most dominant information from either textual or visual domains to generate difficult training samples and drive the model to discover complementary textual-visual correspondences.

In multimodal learning, most methods use the alignment between modalities for collaborative inference but do not yet explore different inter-and cross-modal sensory fusion stages to generate a single representation from different sensors, as like humans. Multimodal sensory fusion via attention is still an underdeveloped area. However, exists some main approaches. An important research development by Wu et al. (2018a) presented a new attention-based hierarchical fusion to explore the complementary multimodal features. It progressively fuses temporal, motion, audio, and semantic label features for video representation. The model consists of three attention layers. First, the low-level attention layer deals with temporal, motion, and audio features inside and across modalities. Second,

high-level attention selectively focuses on semantic label features. Finally, the sequential attention layer incorporates hidden information generated by encoded low-level attention and high-level attention. Hori et al. (2017a) extended simple attention multimodal fusion. Unlike the simple multimodal fusion method, the feature-level attention weights can change according to the decoder state and the context vectors. Hence, enabling the decoder network to pay attention to different features or modalities when predicting each subsequent word in the description. Finally, Memory Fusion Network (Zadeh et al. 2018a) presented the Delta-memory attention module for multi-view sequential learning. In the model, an LSTM for each modality encodes the modality-specific dynamics and interactions. Delta-memory attention discovers both cross-modality and temporal interactions in different memory dimensions of LSTMs and a Multi-view Gated Memory (unifying memory) stores the cross-modality interactions over time.

Following a different research approach, Li et al. (2019k) introduced, to the image captioning task, the Long Short-Term Memory with Pointing (LSTM-P) inspired by humans pointing behavior (Matthews et al. 2012), and Pointer Networks (Vinyals et al. 2015a). The pointer behavior is a psychology research area that focuses on understanding the ontogenic origins that lead us to point to things in the world when we do not fully know them. Some researchers argue that this is an attempt to promote individual or social group attention to an unknown focus to enhance learning. (Matthews et al. 2012). In the LSTM-P, the pointing mechanism encapsulates dynamic contextual information (current input word and LSTM cell output) to deal with the image captioning scenario's novel and rare objects. The model has CNN as encoder, LSTM as decoder, the attentional mechanism as a pointer, and a pre-trained object learners module to recognize objects. First, the representation of the image extracted by the encoder is injected into the decoder to trigger the word generation step by step. Meanwhile, the object learners module generates the probability distribution of the input image classes. This distribution is concatenated with the current hidden state of the LSTM in a copy layer which produces the probability that the distribution will be copied directly to the output. The attentional mechanism learns to switch between copying the copy layer content to the output or activating the decoder to produce the output naturally. This strategy has two fundamental contributions to minimize two significant challenges of neural networks. The first is a pioneering approach to demonstrate the use of attention to plug in external knowledge. The second is to facilitate learning in unbalanced or few samples datasets. In LSTM-P, the attention mechanism can point out directly to object learners distributions when objects are scarce in the dataset, and the encoder has difficulty recognizing them.

2.3 Attention-augmented memory

Attentional interfaces also allow the neural network iteration with other cognitive elements (i.e., memories, working memory). Memory control and logic flow are essential for learning. However, they are elements that do not exist in classical architectures. The memory of classic Recurrent Neural Networks (RNNs), encoded by hidden states and weights, is usually minimal and is not sufficient to remember facts from the past accurately. Most Deep Learning models do not have a simple way to read and write data to an external memory component. The Neural Turing Machine (NTM) (Graves et al. 2014) and Memory Networks (MemNN) (Weston et al. 2014)—a new class of neural networks—introduced the possibility for a neural network dealing with addressable memory. NTM is a differentiable approach that can be trained with gradient descent algorithms, producing a practical

learning program mechanism. NTM memory is a short-term storage space for information with its rules-based manipulation. Computationally, these rules are simple programs, where data are those programs' arguments. Therefore, an NTM resembles a working memory designed to solve tasks that require rules, where variables are quickly linked to memory slots. NTMs use an attentive process to read and write elements to memory selectively. This attentional mechanism makes the network learn to use working memory instead of implementing a fixed set of symbolic data rules.

Memory Networks (Weston et al. 2014) are a relatively new framework of models designed to alleviate the problem of learning long-term dependencies in sequential data by providing an explicit memory representation for each token in the sequence. Instead of forgetting the past, Memory Networks explicitly consider the input history, with a dedicated vector representation for each history element, effectively removing the chance to forget. The limit on memory size becomes a hyper-parameter to tune, rather than an intrinsic limitation of the model itself. This model was used in question-answering tasks where the long-term memory effectively acts as a (dynamic) knowledge base, and the output is a textual response. Large-scale question-answer tests were performed, and the reasoning power of memory networks that answer questions that require an in-depth analysis of verb intent was demonstrated. Mainly due to the success of MemNN, networks with external memory are a growing research direction in DL, with several branches under development as shown in Fig. 5.

Some important branches of MemMM deserve to be highlighted. End-to-end Memory Networks (Sukhbaatar et al. 2015) is the first version of MemNN applicable to realistic, trainable end-to-end scenarios, which requires low supervision during training. Oh et al. (2019) extends MemNN to suit the task of semi-supervised segmentation of video objects. Frames with object masks are placed in memory, and a frame to be segmented acts as a query. The memory is updated with the new masks provided and faces challenges such as changes, occlusions, and accumulations of errors without online learning. The algorithm acts as an attentional space-time system calculating when and where to meet each query pixel to decide whether the pixel belongs to a foreground object or not. Kumar et al. (2016) propose the first network with episodic memory—a type of memory extremely relevant to humans—to iterate over representations emitted by the input module updating its internal state through an attentional interface. Lu et al. (2020), an episodic memory with a key-value retrieval mechanism chooses which parts of the input to focus on thorough attention. The module then produces a summary representation of the memory, taking into account the query and the stored memory. Finally, the latest research has invested in Graph Memory Networks (GMN), which are memories in GNNs (Wu et al. 2020), to better handle unstructured data using key-value structured memories (Miller et al. 2016; Ahmadi et al. 2020; Moon et al. 2019).

2.4 End-to-end attention models

In mid-2017, research aiming at end-to-end attention models appeared in the area. The Neural Transformer (NT) (Vaswani et al. 2017) and Graph Attention Networks (GAT) (Veličković et al. 2018)—purely attentional architectures—demonstrated to the scientific community that attention is a key element for the future development in Deep Learning. The Transformer's goal is to use self-attention (Sect. 3) to minimize traditional recurrent neural networks' difficulties. The Neural Transformer is the first neural architecture that uses only attentional modules and fully-connected neural networks to process

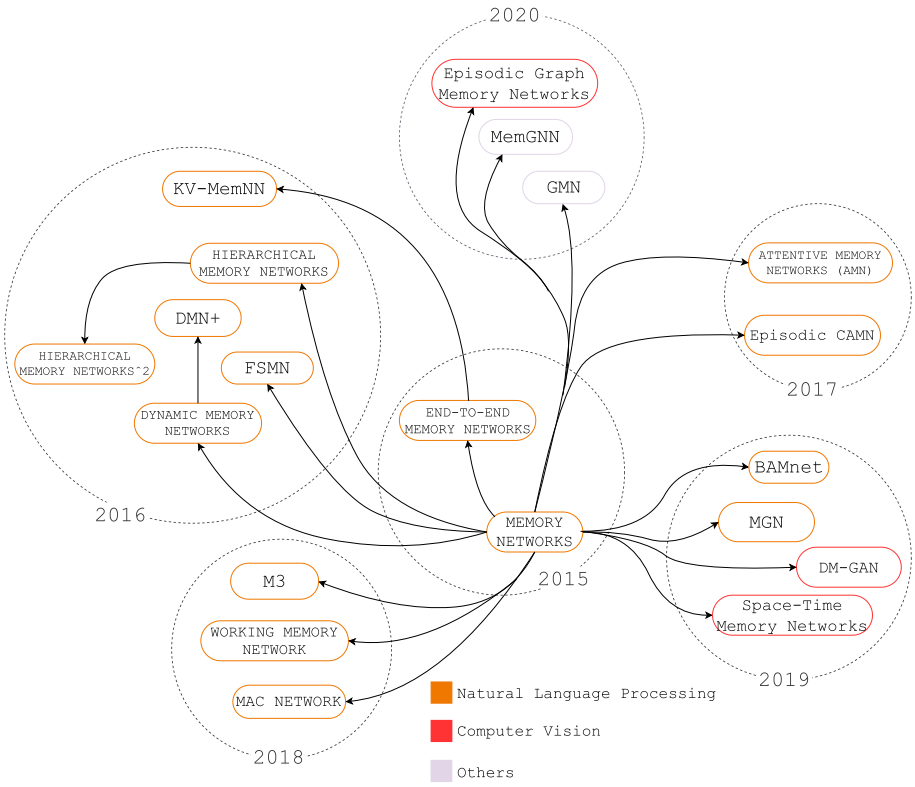

Fig. 5 Memory-based neural networks (MemNN). Architectures labels are color-coded as follows: orange—natural language processing, red—computer vision, purple—others. The end-to-End Memory networks is the first end-to-end differentiable version of MemNN. GMN (Ahmadi et al. 2020) and MemGNN (Ahmadi et al. 2020) are the first graph networks with memory. DMN (Xiong et al. 2016), MemGNN (Ahmadi et al. 2020), Episodic graph memory networks (Lu et al. 2020), Episodic CAMN (Abdulnabi et al. 2017), are the first instances of the episodic memory framework. (Color figure online)

sequential data successfully. It dispenses recurrences and convolutions, capturing the relationship between the sequence elements regardless of their distance. Attention allows the Transformer to be simple, parallelizable, and low training cost (Vaswani et al. 2017). GATs are an end-to-end attention version of GNNs (Wu et al. 2020). They have stacks of attentional layers that help the model focus on the unstructured data's most relevant parts to make decisions. The main purpose of attention is to avoid noisy parts of the graph by improving the signal-to-noise ratio (SNR) while also reducing the structure's complexity. Furthermore, they provide a more interpretable structure for solving the problem. For example, when analyzing the attention of a model under different components in a graph, it is possible to identify the main factors contributing to achieving a particular response condition.

There is a growing interest in NT and GATs, and some extensions have been proposed (Wang et al. 2019d, a; Abu-El-Haija et al. 2018; Li et al. 2019e), with numerous Transformer-based architectures as shown Fig. 6. These architectures and all that use self-attention belong to a new category of neural networks, called Self-Attentive Neural

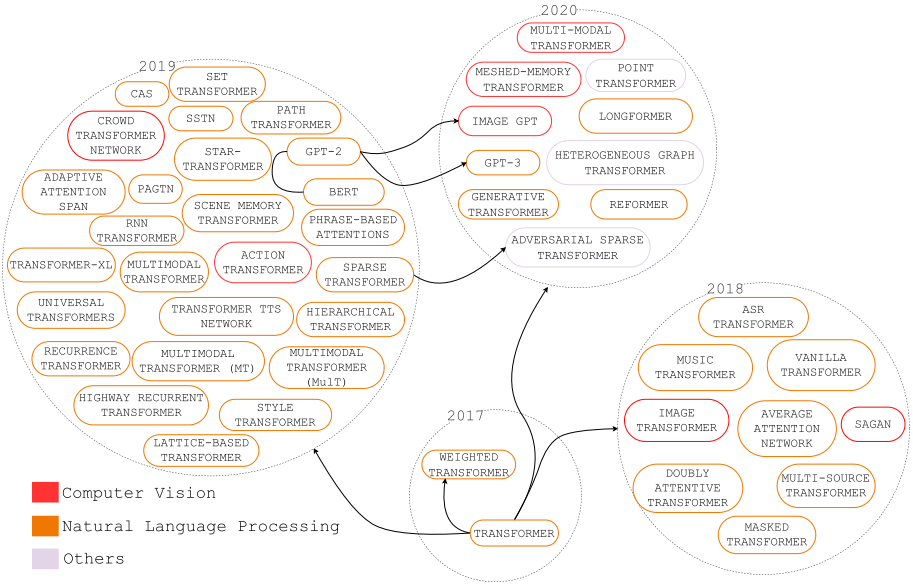

Fig. 6 Transformer-based neural networks. Architectures labels are color-coded as follows: orange—natural language processing, red—computer vision, purple—others. (Color figure online)

Networks. They aim to explore self-attention in various tasks and improve the following drawbacks: (1) a Large number of parameters and training iterations to converge; (2) High memory cost per layer and quadratic growth of memory according to sequence length; (3) Auto-regressive model; (4) Low parallelization in the decoder layers. Specifically, Star-transformer (Guo et al. 2019a) proposes a lightweight alternative to reduce the model's complexity with a star-shaped topology. To reduce the cost of memory, Music Transformer (Huang et al. 2018a) introduces relative self-attention and factored self-attention. Lee et al. (2019a) also features an attention mechanism that reduces self-attention from quadratic to linear, allowing scaling for high inputs and data sets.

Some approaches adapt the Transformer to new applications and areas. In natural language processing, several new architectures have emerged, mainly in multimodal learning. The Multi-source Transformer (Libovický et al. 2018) explores four different strategies for combining input into the multi-head attention decoder layer for multimodal translation. Style Transformer (Dai et al. 2019a), Hierarchical Transformer (Liu and Lapata 2019), HighWay Recurrent Transformer (Chiang et al. 2020), Lattice-Based Transformer (Xiao et al. 2019), Transformer TTS Network (Li et al. 2019f), Phrase-Based Attention (Nguyen and Joty 2018) are some important architectures in style transfer, document summarization and machine translation. Transfer Learning in NLP is one of Transformer's major contribution areas. BERT (Devlin et al. 2019), GPT-2 (Radford et al. 2019), and GPT-3 (Brown et al. 2020) based NT architecture to solve the problem of Transfer Learning in NLP because current techniques restrict the power of pre-trained representations. In computer vision, the generation of images is one of the Transformer's great news. Image Transformer (Parmar et al. 2018), SAGAN (Zhang et al. 2019a), and Image GPT (Chen et al. 2020) uses self-attention mechanism to attend the local neighborhoods. The size of the images that the model can process in practice significantly increases, despite maintaining significantly larger receptive fields per layer than the typical convolutional neural networks.

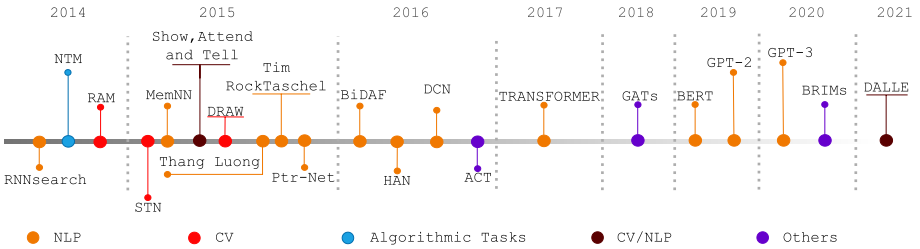

Fig. 7 Key developments in attention in DL timeline. RNNSearch presented the first attention mechanism. NTM and Memory networks introduced memory and dynamic flow control. RAM and DRAW learned to combine multi-glimpse, visual attention, and sequential processing. STN introduced a module to increase the robustness of CNNs to variations in spatial transformations. Show, attend and tell created attention for multimodality. The Ptr-Net used attention as a pointer. BiDAF, HAN, and Dynamic Coattetion Network (DCN) presented attentional techniques to align data with different hierarchical levels. ACT introduced the computation time topic. Transformer (Vaswani et al. 2017) was the first self-attentive neural network with an end-to-end attention approach. GATs introduced attention in GNNs. BERT (Devlin et al. 2019), GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020), and DALL·E are the state-of-the-art in language models and text-to-image generation. Finally, BRIMs (Mittal et al. 2020) learned to combine bottom-up and top-down signals

Recently, at the beginning of 2021, OpenAi introduced the scientific community to DALL·E,[1] the Newest language model based on Transformer and GPT-3, capable of generating images from texts extending the knowledge of GPT-3 for viewing with only 12 billions of parameters.

2.5 Attention today

Currently, hybrid models that employ the main key developments in attention's use in Deep Learning (Fig. 7) have aroused the scientific community's interest. Mainly, hybrid models based on Transformer, GATs, and Memory Networks have emerged for multimodal learning and several other application domains. Hyperbolic Attention Networks (HAN) (Gülçehre et al. 2019), Hyperbolic Graph Attention Networks (GHN) (Zhang et al. 2019l), Temporal Graph Networks (TGN) (Rossi et al. 2020) and Memory-based Graph Networks (MGN) (Ahmadi et al. 2020) are some of the most promising developments. Hyperbolic networks are a new class of architecture that combine the benefits of self-attention, memory, graphs, and hyperbolic geometry in activating neural networks to reason with high capacity over embeddings produced by deep neural networks. Since 2019 these networks have stood out as a new research branch because they represent state-of-the-art generalization on neural machine translation, learning on graphs, and visual question answering tasks while keeping the neural representations compact. Since 2019, GATs have also received much attention due to their ability to learn complex relationships or interactions in a wide spectrum of problems ranging from biology, particle physics, social networks to recommendation systems. To improve the representation of nodes and expand the capacity of GATs to deal with data of a dynamic nature (i.e. evolving features or connectivity

[1] https://github.com/lucidrains/DALLE-pytorch.

over time), architectures that combine memory modules and the temporal dimension, like MGNs and TGNs, were proposed.

At the end of 2020, two research branches still little explored in the literature were strengthened: (1) explicit combination of bottom-up and top-down stimuli in bidirectional recurrent neural networks and (2) adaptive computation time. Classic recurrent neural networks perform recurring iteration within a particular level of representation instead of using a top-down iteration, in which higher levels act at lower levels. However, Mittal et al. (2020) revisited the bidirectional recurrent layers with attentional mechanisms (BRIMs) to explicitly route the flow of bottom-up and top-down information, promoting selection iteration between the two levels of stimuli. The approach separates the hidden state into several modules so that upward iterations between bottom-up and top-down signals can be appropriately focused. The layer structure has concurrent modules so that each hierarchical layer can send information both in the bottom-up and top-down directions.

BRIMs showed that modularity, sparsity, hierarchy, and top-down with the bottom-up flow are essential things that significantly improve out-of-distribution (OOD) generalization, a challenging problem in classic deep learning models and reinforcement learning techniques. When the distribution of the test set changes in a minimal aspect, the classic models fail significantly, including some mainly attentional neural networks, such as Neural Transformer. In contrast, BRIM performs state-of-the-art results in OOD generalization. BRIM mainly uses self-attention as a link among equal LSTM modules, generating a very sparse and modular framework with only a small portion of modules actives at time t. Improving outcomes in OOD generalization is fundamental to producing the general AI because phenomena in the real world hardly belong to the same distribution. Based on BRIM and recurrent independent mechanisms research conducted by Goyal et al. (2019), significant discussions have been created in the literature to face OOD problems (Hendrycks et al. 2020) (Arjovsky 2020). Nonetheless, for neuroscientists, modularity, hierarchy, and sparsity biological neural networks structures have been discussed as fundamental elements for intelligence in the last 20 years. In reverse engineering neocortex experiments, neuroscientists showed a very modular and hierarchical structure with repetitive circuits organized by macro-columns composed mainly of pyramidal neurons (Hole and Ahmad 2021).

Attention to select data dynamically is widespread in the literature. However, there are still scarce mechanisms to control computations steps/time or activate/deactivate neural structures. In this line, the adaptive computation time (ACT) is an interesting little-explored topic in the literature that began to expand only in 2020 despite initial studies emerging in 2017. ACT applies to different neural networks (e.g., RNNs, CNNs, LSTMs, Transformers). The general idea is that complex data might require more computation to produce a final result. In contrast, some unimportant or straightforward data might require less. The attention mechanism dynamically decides how long to process network training data. The seminal approach by Graves (2016) made minor modifications to an RNN, allowing the network to perform a variable number of state transitions and a variable number of outputs at each stage of the input. The resulting output is a weighted sum of the intermediate outputs, i.e., soft attention. A halting unit decides when the network should stop or continue. To limit computation time, attention adds a time penalty to the cost function by preventing the network from processing data for unnecessary amounts of time. This approach has recently been updated and expanded to other architectures. Spatially Adaptive Computation Time (SACT) (Figurnov et al. 2017) adapts ACT to adjust the per-position amount of computation to each spatial position of the block in convolutional layers,

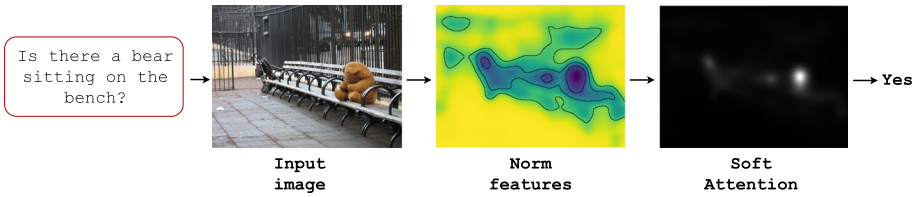

Fig. 8 An intuitive example of soft attention. Visual QA architecture outputs an answer given an image and a textual question as input. It uses a soft attention mechanism that weighted visual features for the task for further processing. The premise is that the norm of the visual features correlates with their relevance. Besides, those feature vectors with high magnitudes correspond to image regions that contain relevant semantic content

learning to focus computing on the regions of interest and to stop when the features maps are "good enough". Finally, Differentiable Adaptive Computation Time (DACT) (Eyzaguirre and Soto 2020) introduced the first differentiable end-to-end approach to computation time on recurring networks.

3 Attention mechanisms

Concerning training strategies, differentiable characteristics, and attentional focus characteristics, the mechanisms can be categorized into soft attention (global attention), hard attention (local attention), and self-attention (intra-attention).

Soft Attention. Soft attention assigns a weight of 0 to 1 for each input element. It decides how much attention should be focused on each element, considering the interdependence between the input of the deep neural network's mechanism and target. It uses softmax functions in the attention layers to calculate weights so that the entire attentional model is deterministic and differentiable. Soft attention can act in the spatial and temporal context. The spatial context operates mainly to extract the features or the weighting of the most relevant features. For the temporal context, it works by adjusting the weights of all samples in sliding time windows, as samples at different times have different contributions. Despite being deterministic and differentiable, soft mechanisms have a high computational cost for large inputs. Figure 8 shows an intuitive example of a soft attention mechanism.

Hard Attention. Hard attention determines whether a part of the mechanism's input should be considered or not, reflecting the interdependence between the input of the mechanism and the target of the deep neural network. The weight assigned to an input part is either 0 or 1. Hence, as input elements are either seen, the objective is non-differentiable. The process involves making a sequence of selections on which part to attend. In the temporal context, for example, the model attends to a part of the input to obtain information, decidinng where to attend in the next step based on the known information. A neural network can make a selection based on this information. However, as there is no ground truth to indicate the correct selection policy, the hard-attention type mechanisms are represented by stochastic processes. As the model is not differentiable, reinforcement learning techniques are necessary to train models with hard attention. Inference time and computational costs are reduced compared to soft mechanisms once the entire input is not being stored or processed. Figure 9 shows an intuitive example of a hard attention mechanism.

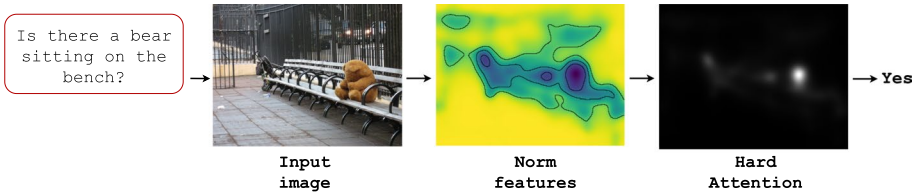

Fig. 9 An intuitive example of hard attention. Given an image and a textual question as input, the Visual QA architecture outputs an answer. It uses a hard attention mechanism that selects only the important visual features for the task for further processing

Fig. 10 Self-attention examples. **a** Self-attention in sentences, **b** self-attention in images. The first image shows five representative query locations with color-coded dots with the corresponding color-coded arrows summarizing the most-attended regions. (Color figure online)

Self-Attention. Self-attention quantifies the interdependence between the input elements of the mechanism. This mechanism allows the inputs to interact with each other "self" and determine what they should pay more attention to. The self-attention layer's main advantages compared to soft and hard mechanisms are parallel computing ability for a long input. This mechanism layer checks the attention with all the same input elements using simple and easily parallelizable matrix calculations. Figure 10 shows an intuitive example of a self-attention mechanism.

4 Attention-based classic deep learning architectures

This section introduces how attentional modules or interfaces have been used in classic DL architectures. Throughout it, we highlight the benefits of attention when plugged into different parts of classic architectures. We have identified pattern uses of attention related to the mechanism's location in the model, the task being performed, and the desired improvements to the problem. Then, for each classic model type, we create different groups to categorize each usage coherently. Specifically, we present the uses of attention in convolutional (Sect. 4.1), recurrent networks (Sect. 4.2) and generative models (Sect. 4.3).

4.1 Attention-based convolutional neural networks (CNNs)

Attention emerges in CNNs to filter information and allocate resources to the neural network efficiently. There are numerous ways to use attention on CNNs, which makes it very difficult to summarize how this occurs and the impacts of each use. We divided the uses of attention into six distinct groups (Fig. 11): (1) DCN attention pool—attention replaces the

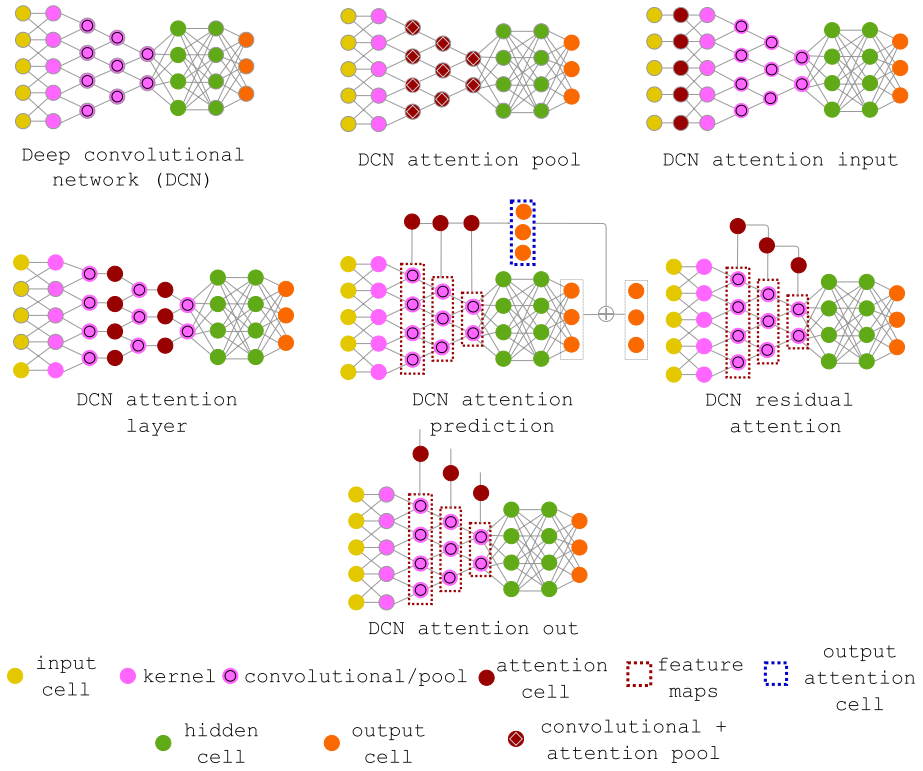

Fig. 11 Attention-based convolutional neural networks. DCN attention pool group uses an attention pool, instead of regular pooling, as a strategy to determine the importance of each individual in a given feature map window. The premise is that only a few of these windows are significant and must be selected concerning a particular objective. DCN attention input group uses structures similar to the human's visual attention. DCN attention layer group collects important stimuli of high (semantic level) and low level (saliency) for subsequent layers of architecture. DCN attention prediction group uses attention in the final stages of prediction, sometimes as an ensemble element. DCN residual attention group uses attention as a residual module between any convolutional layers to mitigate the vanishing problem, capturing only the relevant stimuli from each feature map. DCN attention out-group can represent the category of recurrent attention processes

classic CNN pooling mechanism; (2) DCN attention input—the attentional modules are filter masks for the input data. This mask assigns low weights to regions irrelevant to neural network processing and high weights to relevant areas; (3) DCN attention layer—attention is between the convolutional layers; (4) DCN attention prediction—attentional mechanisms assist the model directly in the prediction process; (5) DCN residual attention—extracts information from the feature maps and presents a residual input connection to the next layer; (6) DCN attention out—attention captures important stimuli of feature maps for other architectures, or other instances of the same architecture. To maintain consistency with the Deep Neural Network's area, we extend The Neural Network Zoo schematics[2] to accommodate attention elements.

[2] https://www.asimovinstitute.org/neural-network-zoo/.

DCN attention input mainly uses attention to filter input data—a structure similar to the multi-glimpse mechanism and visual attention of human beings. Multi-glimpse refers to the ability to quickly scan the entire image and find the main areas relevant to the recognition process, while visual attention focuses on a critical area by extracting key features to understand the scene. When a person focuses on one part of the image, the different regions' internal relationship is captured, guiding eye movement to find the next relevant area-ignoring the irrelevant parts easy learning in the presence of disorder. For this reason, human vision has an incomparable performance in object recognition. The main contribution of attention at the CNNs' input is robustness. If our eyes see an object in a real-world scene, parts far from the object are ignored. Therefore, the distant background of the fixed object does not interfere in recognition. However, CNNs treat all parts of the image equally. The irrelevant regions confuse the classification and make it sensitive to visual disturbances, including background, changes in camera views, and lighting conditions. Attention in CNNs' input contributes to increasing robustness in several ways: (1) It makes architectures more scalable, in which the number of parameters does not vary linearly with the size of the input image; (2) Eliminates distractors; (3) Minimizes the effects of changing camera lighting, scale, and views. (4) It allows the extension of models for more complex tasks, i.e., fine-grained classification or segmentation. (5) Simplifies CNN encoding. (6) Facilitates learning by including relevant priorities for architecture.

The main approaches use soft, hard attention mechanisms and saliency maps to focus on relevant regions of the input image. Zhao et al. (2016) used visual attention-based image processing to generate the focused image. Then, the focused image is input into CNN to be classified. The information entropy guides reinforcement learning agents to achieve a better image classification policy according to the classification. Wang et al. (2016c) used attention to create representations rich in motion information for action recognition. The attention extracts saliency maps using both motion and appearance information to calculate the objectness scores. For a video, attention processes frame by frame to generate a saliency-aware map for each frame. The classic pipeline uses only CNN sequence features as input for LSTMs, failing to capture adjacent frames' motion information. The saliency-aware maps capture only regions with relevant movements making CNN encoding simple and representative for the task. Liu et al. (2019a) used attention as input of a CNN to provide important priors in counting crowded tasks. An attention map generator first provides two priors for the system: candidate crowd regions and crowd regions' congestion degree. The priors guide subsequent CNNs to pay more attention to those regions with crowds and improve their capacity to be resistant to noise. Specifically, the congestion degree prior provides fine-grained density estimation for a system.

In classic CNNs, the size of the receptive fields is relatively small. Most of them extract features locally with convolutional operations, which fail to capture long-range dependencies between pixels throughout the image. However, larger receptive fields allow for better use of training inputs, and much more context information is available at the expense of instability or even convergence in training. Also, traditional CNNs treat channel features equally. This naive treatment lacks the flexibility to deal with low and high-frequency information. Some frequencies may contain more relevant information for a task than others. However, equal treatment by the network makes it difficult to converge the models. To mitigate such problems, most literature approaches use attention between convolutional layers (i.e., DCN attention layer and DCN residual attention), as shown in Fig. 12. Between layers, attention acts mainly for feature recalibration, capturing long-term dependencies, internalizing, and correctly using past experiences.

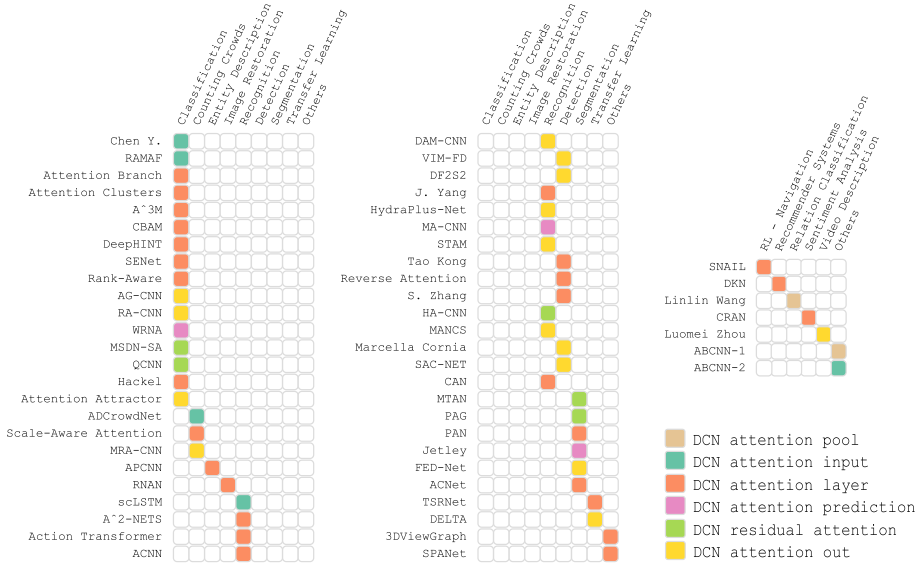

Fig. 12 Attention-based convolutional neural networks main architectures by task and attention use

The pioneering approach to adopting attention between convolutional layers is the Squeeze-and-Excitation Networks (Hu et al. 2020a) created in 2016 and winner of the ILSVRC in 2017. It is also the first architecture to model channel interdependencies to recalibrate filter responses in two steps, squeeze and excitation, i.e., SE blocks. To explore local dependencies, the squeeze module encodes spatial information into a channel descriptor. The output is a collection of local descriptors with expressive characteristics for the entire image. To make use of the information aggregated by the squeeze operation, excitation captures channel-wise dependencies by learning a non-linear and non-mutually exclusive relationship between channels, ensuring that multiple channels can be emphasized. In this sense, SE blocks intrinsically introduce attentional dynamics to boost feature discrimination between convolutional layers.

The inter-channel and intra-channel attention to capturing long-term dependencies and simultaneously taking advantage of high and low-level stimuli are widely explored in the literature. Zhang et al. (2019j) proposed residual local and non-local attention blocks consisting of trunk and mask branches. Their attention mechanism helps learn local and non-local information from the hierarchical features. It preserves low-level features while maintaining a representational quality of high-level features. The Cbam (Woo et al. 2018) infers attentional maps in two separate dimensions, channel and spatial, for adaptive feature refinement. The double attention block in Chen et al. (2018e) aggregates and propagates global informational features considering the entire spatio-temporal context of images and videos, allowing subsequent convolution layers to access resources from across space efficiently. In the first stage, attention gathers features from all space into a compact set employing groupings. In the second stage, it selects and adaptively distributes the resources for each architectural location. Following similar exploration proposals, several attentional modules can be easily plugged into classic CNNs (Fukui et al. 2019a; Han et al. 2019; Ji et al. 2017; Yang et al. 2017b).

Hackel et al. (2018) explored attention to preserving sparsity in convolutional operations. Convolutions with kernels greater than 1×1 generate fill-in, reducing feature maps' sparse nature. Generally, the change in data sparsity has little influence on the network output. However, memory consumption and execution time considerably increase when it occurs in many layers. To guarantee low memory consumption, attention acts as a $k - selection$ filter, which has two different versions of selection: (1) it acts on the output of the convolution, preferring the largest k positive responses similar to a rectified linear unit; (2) it chooses the k highest absolute values, expressing a preference for responses of great magnitude. The parameter k controls the level of sparse data and, consequently, computational resources during training and inference. Results point out that training with attentional control of data sparsity can reduce in more than 200% the forward pass runtime in one layer.

Simple Neural Attentive Meta-Learner (SNAIL) (Mishra et al. 2018)—a pioneering class of meta-learner based attention architectures—proposed combining temporal convolutions with soft attention to previous aggregate information and dynamically point to past experiences. This approach demonstrates that attention acts as a complement to the disadvantages of convolution. Attention allows precise access in an infinitely large context. At the same time, convolutions provide high-bandwidth access at the expense of a finite context. By merging convolutional layers with attentional layers, SNAIL can effectively unrestrictly access the number of previous experiences. The model can learn a more efficient representation of features. As additional benefits, SNAIL architectures become simpler to train than classic RNNs.

The DCN attention out-group uses attention to share relevant feature maps with other architectures or even with instances of the current architecture. Usually, the main objective is to facilitate the fusion of features, multimodality, and external knowledge. In some cases, attention regularly works by turning classic CNNs into recurrent convolutional neural networks—a new trend in Deep Learning to deal with challenging image problems. Recurrent attention convolutional neural network (RA-CNN) (Fu et al. 2017) is a pioneering framework for recurrent convolutional networks. In their framework, attention proceeds along two dimensions, i.e., discriminative feature learning and sophisticated part localization. Given an input image, a classic CNN extracts feature maps. Then, the attention proposed network maps convolutional features to a feature vector that could be matched with the category entries. Hence, attention estimates the focus region for the next CNN instance, i.e., the next finer scale. Once it locates the focus region, the system cuts and enlarges the region to a finer scale with higher resolution to extract more refined features. Thus, each CNN in the stack generates a prediction so that the stack's deepest layers produce more accurate predictions.

For fusion of features, Chen et al. (2019b) presented Feature-fusion Encoder–Decoder Network (FED-net) to image segmentation. Their model uses attention to fuse features of different levels of an encoder. The attention module merges features from its current level with features from later levels at each encoder level. After the merger, the decoder performs convolutional upsampling with the information from each attention level, which contributes by modulating the most relevant stimuli for segmentation. Tian et al. (2018b) used feature pyramid-based attention to combining meaningful semantic features with semantically weak but visually strong features in a face detection task. Their goal is to learn more discriminative hierarchical features with enriched semantics and details at all levels to detect hard-to-detect faces, like tiny or partially occluded faces. Their attention mechanism can fuse different feature maps from top to bottom recursively by combining transposed

convolutions and element-wise multiplication maximizing mutual information between the lower and upper-level representations.

To plug external knowledge, Delta (Li et al. 2019i) framework presented an efficient strategy for transfer learning. Their attention system acts as a behavior regulator between the source model and the target model. Attention identifies the source model's completely transferable channels, preserving their responses and identifying the non-transferable channels to dynamically modulate their signals, increasing the target model's generalization capacity. Specifically, the attentional system characterizes the distance between the source/target model through the feature maps' outputs. It incorporates that distance to regularize the loss function. Optimization normally affects the weights of the neural network and assigns generalization capacity to the target model. Regularization modulated by attention on high and low semantic stimuli manages to take important steps to plug in external knowledge in the semantic problem.

The DCN attention prediction group uses attention directly in the prediction process. Various attentional systems capture features from different convolutional layers as input and generate a prediction as an output. Voting between different predictors generates the final prediction. Reusing activations of CNNs feature maps to find the most informative parts of the image at different depths makes prediction tasks more discriminative. Each attentional system learns to relate stimuli and part-based fine-grained features, which, although correlated, are not explored together in classical approaches. Nonetheless, Zheng et al. (2017) proposed a multi-attention mechanism to group channels, creating part classification sub-networks. The mechanism takes input feature maps from convolutional layers. It generates multiple single clusters spatially-correlated subtle patterns as a compact representation. The sub-network classifies an image by each part. The attention mechanism proposed in Rodríguez et al. (2018) uses a similar approach. However, instead of grouping features into clusters, the attentional system has the most relevant feature map regions selected by the attention heads. The output heads generate a hypothesis given the attended information, whereas confidence gates generate a confidence score for each attention head.

Finally, the DCN attention pool group replaces classic pooling strategies with attention-based pooling. The objective is to create a non-linear encoding to select only stimuli relevant to the task, given that classical strategies select only the most contrasting stimuli. To modulate the resulting stimuli, attentional pooling layers generally capture different relationships between feature maps or between different layers. For example, Wang et al. (2016b) created an attentional mechanism that captures pertinent relationships between convoluted context windows and the relation class embedding through a correlation matrix learned during training. The correlation matrix modulates the convolved windows, and finally, the mechanism selects only the most salient stimuli. A similar approach is also followed in (Yin et al. 2016) for modeling sentence pairs.

4.2 Attention-based recurrent neural networks (RNNs)

Attention in RNNs is mainly responsible for capturing long-distance dependencies. Currently, there are not many ways to use attention on RNNs. RNNSearch's mechanism for encoder–decoder frameworks inspires most approaches (Bahdanau et al. 2015). We divided the uses of attention into three distinct groups (Fig. 13): (1) Recurrent attention input—the first stage of attention to select elementary input stimulus, i.e., elementary features, (2) Recurrent memory attention—the first stage of attention to historical weight components,

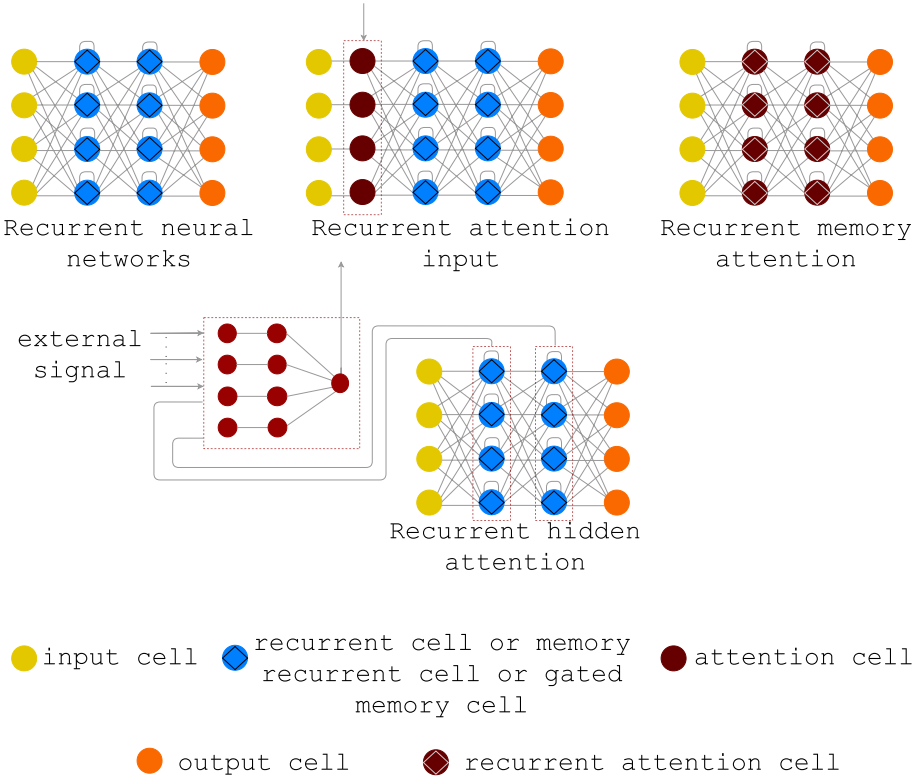

Fig. 13 Attention-based recurrent neural networks. The architecture is a classic recurrent network or a Bi-RNN when hidden layers are recurrent cells. When the hidden layer is a recurrent memory, the architecture is an LSTM or Bi-LSTM. Finally, the architecture is a GRU when the hidden layer is a gated memory cell. The recurrent attention input group uses attention to filter input data. Recurrent hidden attention groups automatically select relevant encoder hidden states across all time steps. Usually, this group implements attention in encoder–decoder frameworks. The recurrent memory attention group implements attention within the memory cell. There are not many architectures in this category as far as we know, but the main uses are related to filtering the input data and the weighting of different historical components for predicting the current time step

(3) Recurrent hidden attention—the second stage of attention to select categorical information to the decode stage.

The recurrent attention input group main uses are item-wise hard, local-wise hard, item-wise soft, and local-wise soft selection. Item-wise hard selects discretely relevant input data for further processing, whereas location-wise hard discretely focuses only on the most relevant features for the task. Item-wise soft assigns a continuous weight to each input data given a sequence of items as input, and location-wise soft assigns a continuous weight between input features. Location-wise soft estimates high weights for features more correlated with the global context of the task. Hard selection for input elements are applied more frequently in computer vision approaches (Mnih et al. 2014; Edel and Lausch 2016). On the other hand, soft mechanisms are often applied in other fields, mainly in natural language processing. The soft selection normally weighs relevant parts of the series or input features, and the attention layer is a feed-forward network differentiable and with a low computational cost. Soft approaches are interesting to filter noise from time series and

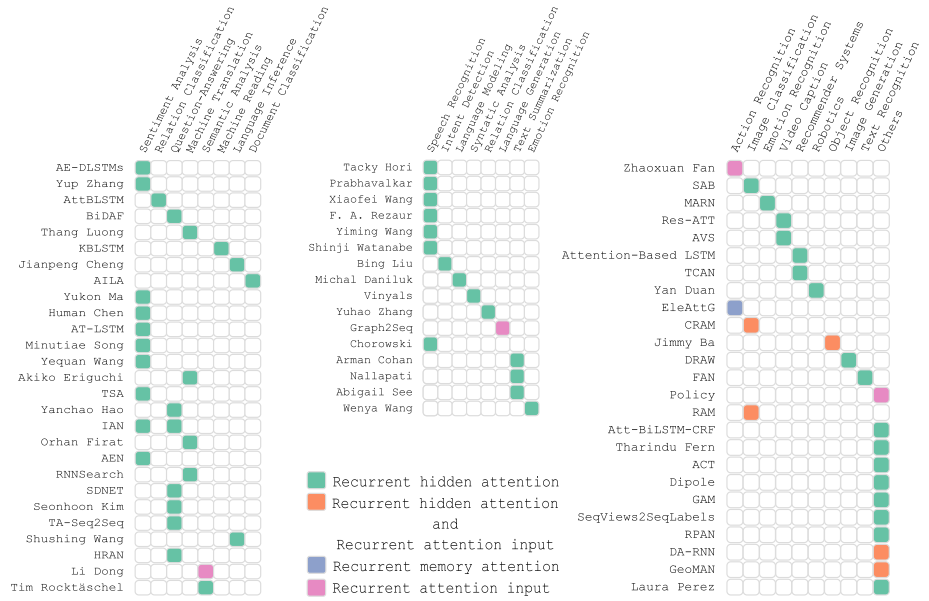

Fig. 14 Attention-based recurrent neural networks main architectures by task and attention use

to dynamically learn the correlation between input features and output (Qin et al. 2017; Liang et al. 2018b; Du et al. 2017). Besides, this approach is useful for addressing graph-to-sequence learning problems that learn a mapping between graph-structured inputs to sequence outputs, which current Seq2Seq and Tree2Seq may be inadequate to handle (Xu et al. 2018c).

Hard mechanisms take inspiration from how humans perform visual sequence recognition tasks, such as reading by continually moving the fovea to the next relevant object or character, recognizing the individual entity, and adding the knowledge to our internal representation. A deep recurrent neural network, at each step, processes a multi-resolution crop of the input image, called a glimpse. The network uses information from the glimpse to update its internal representation and outputs the next glimpse location. The Glimpse network captures salient information about the input image at a specific position and region size. The internal state is formed by the hidden units h_t of the recurrent neural network, which is updated over time by the core network. At each step, the location network estimates the next focus localization, and action networks depend on the task (e.g., for the classification task, the action network's outputs are a prediction for the class label.). Hard attention is not entirely differentiable and therefore uses reinforcement learning.

RAM (Mnih et al. 2014) was the first architecture to use a recurrent network implementing hard selection for image classification tasks. While this model has learned successful strategies in various image data sets, it only uses several static glimpse sizes. Capacity visual attention networks (CRAM) (Edel and Lausch 2016) uses an additional sub-network to dynamically change the glimpse size, with the assumption to increase the performance, and in Ba et al. (2015) explore modifications in RAM for real-world image tasks and multiple objects classification. CRAM is a similar RAM model except for two key differences: Firstly, a dynamically updated attention mechanism restrains the input region observed by the glimpse network and the next output region prediction from the emission network—a

network that incorporates the location and capacity as well as past information. More straightforwardly, the sub-network decides what the focus region's capacity should be at each time step. Secondly, the capacity sub-network outputs are successively added to the emission network's input, ultimately generating the information for the next focus region. Hence, allowing the emission network to combine the information from the location and the capacity networks.

Nearly all important works in the field belong to the recurrent hidden attention group, as shown in Fig. 14. In this category, the attention mechanism selects elements in the RNN's hidden layers for inter-alignment, contextual embedding, multiple-input processing, memory management, and capturing long-term dependencies, a typical problem with recurrent neural networks. Inter-alignment involves the encoder–decoder framework, and the attention module between these two networks is the most common approach. This mechanism builds a context vector dynamically from all previous decoder hidden states and the current encoder hidden state. Attention in inter-alignment helps minimize the bottleneck problem, with RNNSearch (Bahdanau et al. 2015) for machine translation tasks as its first representative. Further, several other architectures implemented the same approach in other tasks (Cheng et al. 2016; Yang et al. 2016c; Seo et al. 2017). For example, Yang et al. (2016c) extended the soft selection to the hierarchical attention structure. It allows computing soft attention at the word and sentence levels in the GRU networks encoder for document classification.

To create contextual embeddings and to manipulate multimodal inputs, co-attention is highly effective for text matching applications. Co-attention enables the learning of pairwise attention, i.e., learning to attend based on computing word-level affinity scores between two documents. Such a mechanism is designed for architectures comprised of queries and context, such as questions and answers and emotions analysis. Co-attention models can be fine-grained or coarse-grained. Fine-grained models consider each element of input concerning each element of the other input. Coarse-grained models calculate attention for each input, using an embedding of the other input as a query. Although efficient, co-attention suffers from information loss from the target and the context due to the anticipated summary. Attention flow emerges as an alternative to summary problems. Unlike co-attention, attention flow links and merges context and query information at each stage, allowing embeddings from previous layers to flow to subsequent modeling layers. The attention flow layer is not used to summarize the query and the context in vectors of unique features, reducing information loss. Attention is calculated in two directions, from the context to the query and from the query to the context. The output is the query-aware representations of context words. Attention flow allows a hierarchical process of multiple stages to represent the context at different granularity levels without an anticipated summary.

Hard attention mechanisms do not often occur on recurrent hidden attention networks. However, Ke et al. (2018) demonstrate that hard selection to retrieve past hidden states based on the current state mimics an effect similar to the brain's ability. Humans use a very sparse subset of past experiences and can access them directly and establish relevance with the present, unlike classic RNNs and self-attentive networks. Hard attention is an efficient mechanism for RNNs to recover sparse memories. It determines which memories will be selected on the forward pass, which will receive gradient updates. At time t, RNN receives a vector of hidden states h^{t-1}, a vector of cell states c^{t-1}, and an input x^t, and computes new cell states c^t and a provisional hidden state vector $\widetilde{h}^t$ that also serves as a provisional output. First, the provisional hidden state vector $\widetilde{h}^t$ is concatenated to each memory vector m_i in the memory M. MLP maps each vector to an attention weight a_i^t, representing memory relevance i in current moment t. With attention weights a_i^t sparse attention computes a

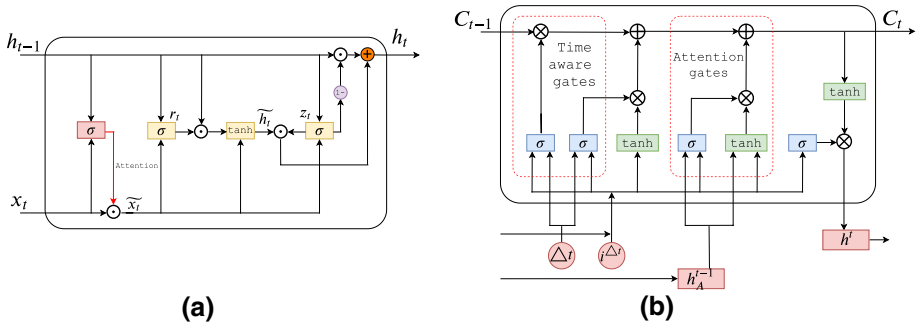

Fig. 15 Recurrent memory attention approaches. **a** Illustration of Element-wise-attention gate in GRU. Specifically, the input modulation is adaptable to the content and is performed in fine granularity, element-wise rather than input-wise. **b** Gate attention in LSTM. The model calculates attention scores to weigh the relevance of different parts of history. There are two additional gates for updating the current cell using the previous state h_A^{t-1}. The first, input attention gate layer, analyzes the current input $i^{\triangle t}$ and h_A^{t-1} to determine which values to be updated in current cell C^t. The second, the modulation attention gate, analyzes the current input $i^{\triangle t}$ and h_A^{t-1}. Then computes the set of candidate values that must be added when updating the current cell state C^t. The attention mechanisms in the memory cell help to more easily capture long-term dependencies and the problem of data scarcity

hard decision. The attention mechanism is differentiable but implements a hard selection to forget memories with no prominence over others. This is quite different from typical approaches as the mechanism does not allow the gradient to flow directly to a previous step in the training process. Instead, it propagates to some local timesteps as a type of local credit given to a memory.

Finally, recurrent memory attention groups implement attention within the memory cell. As far as our research goes, there are not many architectures in this category. (Zhang et al. 2018d) proposed an approach that modulates the input adaptively within the memory cell by assigning different levels of importance to each element/dimension of the input, as shown in Fig. 15a. (Perera and Zimmermann 2018) proposed mechanisms of attention within memory cells to improve the past encoding history in the cell's state vector since all parts of the data history are not equally relevant to the current prediction. As shown in Fig. 15b, the mechanism uses additional gates to update LSTM's current cell.

4.3 Attention-based generative models

Attention emerges in generative models essentially to augmented memory. Currently, there are not many ways to use attention on generative models. Since generative adversarial networks (GANs) are not a neural network architecture but a framework, we do not discuss the use of attention in GANs but autoencoders (AE). We divided the uses of attention into three distinct groups (Fig. 16): (1) Autoencoder input attention—attention provides spatial masks corresponding to all the parts for a given input, while a component autoencoder (e.g., AE, variational autoencoders (VAE), sparse autoencoders (SAE)) independently models each of the parts indicated by the masks. (2) Autoencoder memory attention—attention module acts as a layer between the encoder–decoder to augmented memory. (3) Autoencoder attention encoder–decoder—a fully attentive architecture acts on the encoder, decoder, or both.

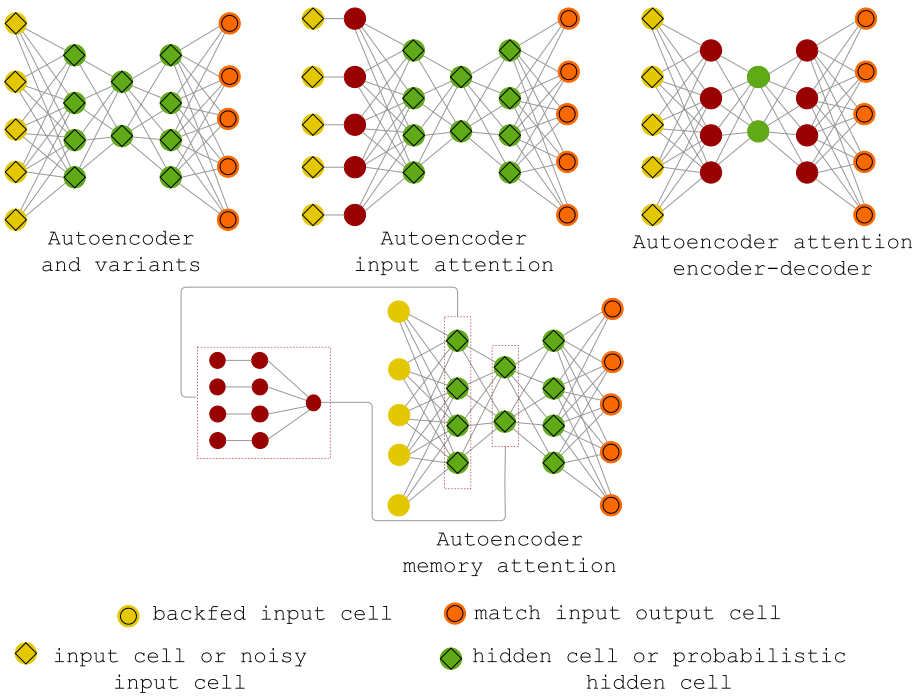

Fig. 16 Attention-based generative models. The Autoencoder input attention group uses attention to facilitate the decomposition of scenes in abstract building blocks. The input components extracted by the masks share significant properties and help to imagine new scenarios. This approach is very efficient for the network to learn to decompose challenging scenarios between semantically significant components. Autoencoder memory attention group handles a memory buffer that allows read/writes operations and is persistent over time. Such models generally handle input and output to the memory buffer using write/read operations guided by the attention system. The use of attention and memory history in autoencoders helps to increase the generalizability of architecture. Autoencoder attention encoder–decoder using a model (usually self-attentive model) to increase the ability to generalize

MONet (Burgess et al. 2019) is one of the few architectures to implement attention at the VAE input. A VAE is a neural network with an encoder parameterized by ϕ and a decoder parameterized by θ. The encoder parameterizes a distribution over the component latent z_k, conditioned on both the input data x and an attention mask m_k. The mask indicates which regions of the input the VAE should focus on representing via its latent posterior distribution, $q\phi(z_k|x, m_k)$. During training, the VAE's decoder likelihood term in the loss $p_\theta(x|z_k)$ is weighted according to the mask, such that it is unconstrained outside of the masked regions. In Li et al. (2016a), the authors use soft attention with learned memory contents to augment models to have more parameters in the autoencoder. In Bartunov and Vetrov (2017), Generative Matching Networks use attention to access the exemplar memory, with the address weights computed based on a learned similarity function between an observation at the address and a function of the latent state of the generative model. In Rezende et al. (2016), external memory and attention work as a way of implementing one-shot generalization by treating the exemplars conditioned on as memory entries accessed through a soft attention mechanism at each step of the incremental generative process similar to DRAW (Gregor et al. 2015). Although most approaches use soft attention to address the

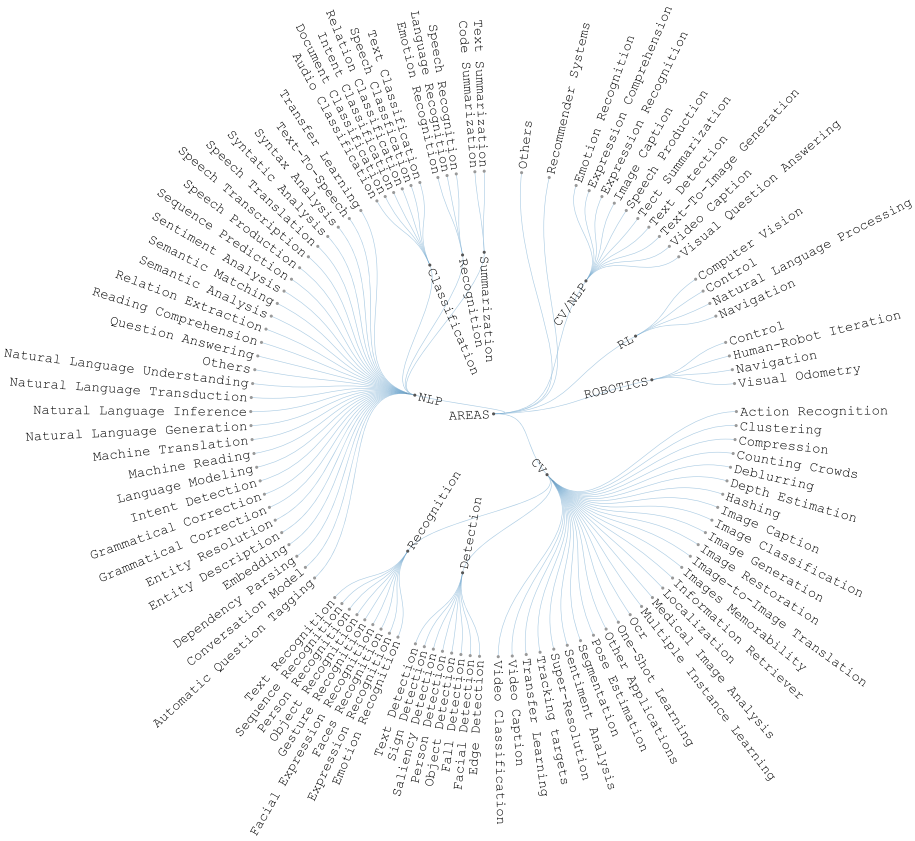

Fig. 17 Diagram showing the main existing applications of neural attention networks. The main areas are Natural language processing (NLP), Computer Vision (CV), multimodal tasks (mainly with images and texts—CV/NLP), reinforcement learning (RL), robotics, recommendation systems, and others (e.i., graph embeddings, interpretability.)

memory, in Bornschein et al. (2017) the authors use a stochastic, hard attention approach, which allows using variational inference about it in a context of few-shot learning.

In Escolano et al. (2018), self-attentive networks increase the autoencoder ability to generalize. The advantage of using this model instead of other alternatives, such as recurrent or convolutional encoders, is that this model is based only on self-attention and traditional attention over the whole representation created by the encoder. This approach allows us to easily employ different components of the networks (encoder and decoder) as modules that, during inference, can be used with other parts of the network without the need for previous step information.

5 Applications

In a few years, neural attention networks have been used in numerous domains due to versatility, interpretability, and significance of results. These networks have been explored mainly in computer vision, natural language processing, and multi-modal tasks, as shown in Fig. 17. For some applications, these models transformed the area entirely (i.e., question-answering, machine translation, document representations/ embeddings, graph embeddings), mainly due to significant performance impacts on the task in question. In others, they helped learn better representations and deal with temporal dependencies over long distances. This section explores a list of application domains and subareas, mainly discussing their main models and how they benefit from attention. Our main objective in this section is to show the reader what problems the attention is trying to solve in the application domain. For the main models, we show how good in terms of metrics the solution is compared to other alternatives in the literature. We present which areas have little or no contribution from attentional mechanisms, presenting good research opportunities. We also present the most representative instances within each area and list them with reference approaches in a wide range of applications.

5.1 Natural language processing (NLP)

In the NLP domain, attention plays a vital role in many sub-areas, as shown in Fig. 17. There are several state-of-the-art approaches, mainly in language modeling, machine translation, natural language inference, question answering, sentiment analysis, semantic analysis, speech recognition, and text summarization. Table 1 groups works developed in each of these areas. Several applications have been facing an increasing expansion, with few representative works, such as emotion recognition, speech classification, sequence prediction, semantic matching, and grammatical correction, as shown in Table 1.

For machine translation (MT), question answering (QA), and automatic speech recognition (ASR), attention works mainly to minimize three classic problems faced by classical neural networks used in NLP: (1) long temporal dependencies: The classic neural networks, even recurrent models such as LSTMS, have memory limitations in remembering earlier information; (2) long sequences: The classic neural networks have problems working with long sequences mainly due to vanishing gradient occasionally by deep layers, and (3) noisy reduction: Mostly, translating one word or answering one question is unnecessary to look at all data. It is necessary to look at a tiny piece of data; this piece has the crucial information for the model at time t; at this moment, the rest of the information is noisy. Classic models have no mechanisms to manipulate the data in a flexible form, difficulting the task. For this, mainly attentional interfaces between classic encoder–decoder frameworks have been used to align words easily in different domains. Alignment is a powerful strategy of attentional interfaces. It enables a decoder or encoder to search for necessary information in a given time step t dynamically, allowing the information flow to be requested on demand for a given need. For example, in ASR tasks, attention aligns acoustic frames extracting information from anchor words to recognize the main speaker while ignoring background noise and interfering speech. Hence, only information on the desired speech is used for the decoder. It provides a straightforward way to align each output symbol with different input frames with selective noise decoding.

Table 1 Summary main state-of-art approaches in natural language processing

Application	References
Natural language processing	
Classification	Choi et al. (2019b), Yang et al. (2016c), Kong et al. (2018a), Liu and Guo (2019), Guo et al. (2019a), Norouzian et al. (2019), Li et al. (2019h), Verga et al. (2018), Guo et al. (2019b), Wang et al. (2016b), Zhou et al. (2016b), Lin et al. (2016) and Zhang et al. (2017d)
Conversation model	Zhou et al. (2018a) and Zhang et al. (2019b)
Code summarization	Allamanis et al. (2016)
Dependency parsing	Dozat and Manning (2017) and Strubell et al. (2018)
Embedding	Lin et al. (2017), Schick and Schütze (2019) and Zhu et al. (2018c)
Entity resolution	Das et al. (2017) and Ganea and Hofmann (2017)
Entity description	Ji et al. (2017)
Language modeling	Dehghani et al. (2019), Joulin and Mikolov (2015) , Ke et al. (2018), Cheng et al. (2016), Vinyals et al. (2016), Dai et al. (2019b), Sukhbaatar et al. (2019), Baevski and Auli (2019), Daniluk et al. (2017)
Language inference	Kim et al. (2017b), Parikh et al. (2016), Shen et al. (2018a), Cheng et al. (2016), Shen et al. (2018c), Guo et al. (2019a), Wang and Jiang (2016) and Liu et al. (2016)
Language understanding	Kim et al. (2018c)
Language transduction	Grefenstette et al. (2015)
Language generation	Xu et al. (2018c)
Machine translation	Kim et al. (2017b), Vaswani et al. (2017), Gülçehre et al. (2019), Dehghani et al. (2019), Bahdanau et al. (2015), Luong et al. (2015), Shaw et al. (2018), Raffel et al. (2017), Cho et al. (2014b), Sennrich et al. (2016), Zenkel et al. (2019), Yang et al. (2019a), Yang et al. (2019b), Hao et al. (2019), Hieber et al. (2020), Zhang et al. (2018g), Bastings et al. (2017), Zhou et al. (2016a), Wu et al. (2016), Eriguchi et al. (2016), Li et al. (2019c), Firat et al. (2016), Deng et al. (2018) and Zhang et al. (2016)
Machine reading	Yang and Mitchell (2017)
Intent detection	Liu and Lane (2016)
Grammatical correction	Ahmadi (2017)
Question tagging	Sun et al. (2018)
Question answering	Kim et al. (2017b), Neelakantan et al. (2016), Xing et al. (2018), Dehghani et al. (2019), Weston et al. (2014), Kumar et al. (2016), Vinyals and Le (2015), Yang et al. (2016a), Tay et al. (2018)and Seo et al. (2017), Yu et al. (2018a)and Wang et al. (2018f)and Zhou et al. (2018d)and Santos et al. (2016)and Zhong et al. (2019)and Shao et al. (2017)and Zhu et al. (2018a)and Tan et al. (2015)and Dhingra et al. (2017)and Hao et al. (2017)and Kadlec et al. (2016)and Sukhbaatar et al. (2015)and Munkhdalai and Yu (2017)and Kim et al. (2019)and Bauer et al. (2018), Ran et al. (2019)and Hermann et al. (2015)and Huang et al. (2018c)and Xing et al. (2017), Sordoni et al. (2016) and Wang et al. (2016a)

Table 1 (continued)

Application	References
Reading comprehension	Shen et al. (2018c), Wang et al. (2018f), Cui et al. (2017), Liu et al. (2019d) and Cui et al. (2016)
Recognition	Chorowski et al. (2015), Raffel et al. (2017), Prabhavalkar et al. (2018), Irie et al. (2019), Lüscher et al. (2019), Dong et al. (2019), Salazar et al. (2019), Wang et al. (2019e), Zhou et al. (2018c), Watanabe et al. (2017), Bahdanau et al. (2016), Chorowski et al. (2014), rahman Chowdhury et al. (2018), Wang et al. (2019f), Zeyer et al. (2018), Kim et al. (2017a), Okabe et al. (2018), Hori et al. (2017b), Mirsamadi et al. (2017), Kumar et al. (2019), Huang et al. (2019), Majumder et al. (2019), Zhang et al. (2018i), Neumann and Vu (2017), Wang et al. (2017b) and Huang et al. (2018d)
Relation extraction	Zhang et al. (2019e)
Syntax analysis	Kuncoro et al. (2017)
Speech production	Tachibana et al. (2018)
Sentiment analysis	Wang et al. (2016d), Shen et al. (2018a), Cheng et al. (2016), Ma et al. (2017), Shuang et al. (2019), Baziotis et al. (2017), Xue and Li (2018), Feng et al. (2019), Song et al. (2019b), Shin et al. (2017), Song et al. (2019a), Du et al. (2019), Chen et al. (2017c), Ma et al. (2018c), Liu and Zhang (2017), Chen et al. (2016a), and Zeng et al. (2019a)
Sequence prodiction	Mensch and Blondel (2018)
Semantic matching	Zhang et al. (2018f)
Semantic analysis	Rocktäschel et al. (2016), She et al. (2018), Zhang et al. (2018a), Shen et al. (2018b), Tan et al. (2018), Dong and Lapata (2016) and Wu et al. (2018c)
Speech translation	Sperber et al. (2019)
Speech transcription	Chan et al. (2016)
Text summarization	Chopra et al. (2016), Paulus et al. (2018), Nallapati et al. (2016), See et al. (2017), Raffel et al. (2017), Rush et al. (2015)and Fan et al. (2018a)and Rekabdar et al. (2019) and Cohan et al. (2018)
Text-to-speech	Yasuda et al. (2019) and Zhang et al. (2019d) and Li et al. (2019f)
Transfer learning	Devlin et al. (2019) and Alt et al. (2018)

In MT, automatic alignment translates long sentences more efficiently. RNN-Search (Bahdanau et al. 2015) and Neural Transformer (Vaswani et al. 2017) brought impressive results to the area. In RNNSearch (Fig. 18a), the attentional interfaces guided by the decoder's previous state dynamically searches for important source words for the next time step. It consists of an encoder followed by a decoder. The encoder is a bidirectional RNN (BiRNN) (Schuster and Paliwal 1997) that consists of forward and backward RNN's. The forward RNN reads the input sequence in order and calculates the forward hidden state sequence. The backward RNN reads the sequence in the reverse order, resulting in the backward hidden states sequence. The decoder has an RNN and an attention

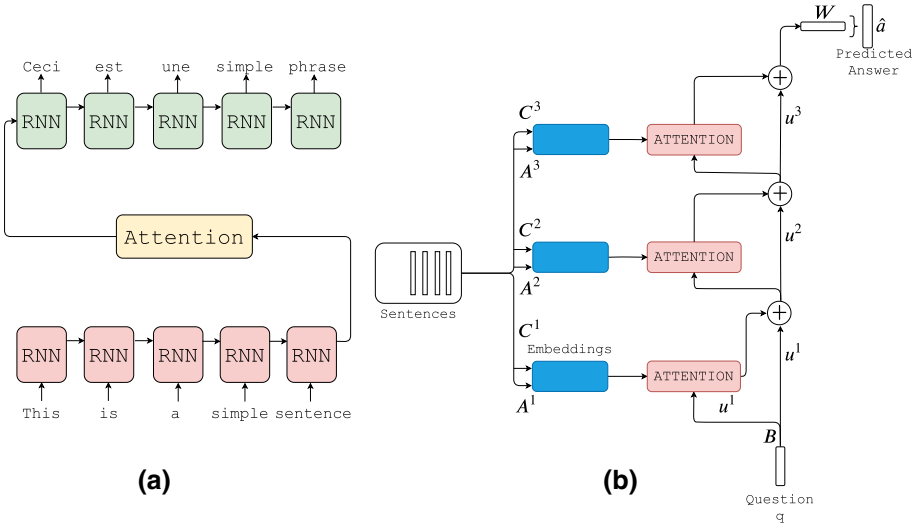

Fig. 18 Illustration of RNNSearch (Bahdanau et al. 2015) for machine translation, and End-to-End Memory Networks (Sukhbaatar et al. 2015) for question answering. a The RNNSearch architecture. The attention guided by the decoder's previous state dynamically searches for important source words for the next time step. b The End-to-End Memory Networks. The architecture consists of external memory and several stacked attention modules. To generate a response, the model makes several hops in memory using only attentional layers

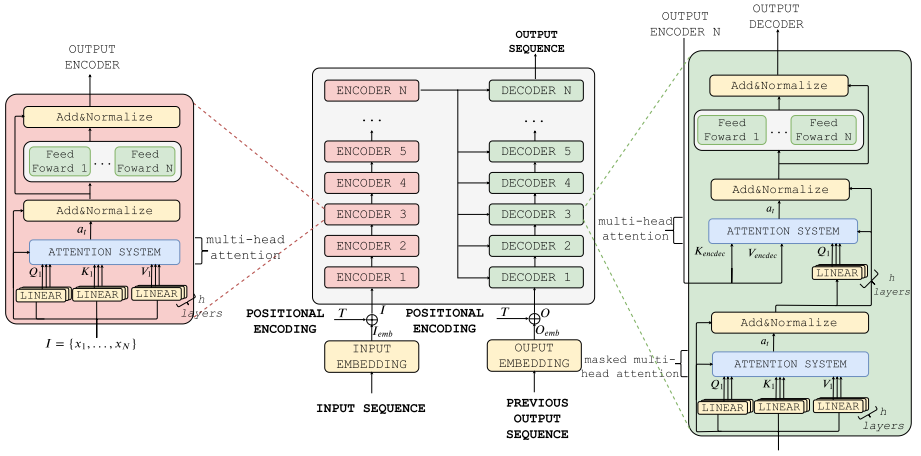

Fig. 19 Illustration of Neural Transformer (Vaswani et al. 2017) for machine translation. The architecture consists of several stacked attentional encoders and decoders. The encoder process is massively parallel and eliminates recurrences, and the decoder generates the translated words sequentially. Each encoder uses multiple heads of the self-attention mechanism followed by fusion and normalization layers. Similarly, to generate the output, each decoder has multiple heads of self-attention and masked-self attention to mask words not yet generated

system that calculates a probability distribution for all possible output symbols from a context vector. RNNSearch surpassed in all test cases the classic RNNencdec machine translation framework using the BLEU metric. While the RNNencdec-30 has a 13.93 BLEU score and RNNencdec-50 has a 17.82 BLEU score when the sequence size goes from 30 to 50, RNNSearch goes from 21.50 to 26.75 when the sequence size increases. It is visible that increasing the sequence size impacts RNNSearch much less than the classic encoder–decoder for machine translation (i.e., RNNencdec).

The Neural Transformer (Fig. 19), on the other hand presents an end-to-end attention alternative to machine translation. It is the base model for state-of-the-art results in NLP. The architecture consists of an arbitrary amount of stacked encoders and decoders. Each encoder has linear layers, an attention system, feed-forward neural networks, and normalization layers. The attention system has several parallel heads. Each head has N attentional subsystems that perform the same task but have different contextual inputs. The encoder receives a word embedding matrix $I = \{x_1, ..., x_N\}$, $I \in \mathbb{R}^{N \times d_{emb}}$ as input. As the architecture does not use recurrences, the input tokens' position information is not explicit but necessary. To represent the spatial position information, the Transformer adds a positional encoding to each embedding vector. Positional encoding is fixed and uses sinusoidal functions.

The input I goes through linear layers and generates, for each word, a query vector (q_i), a key vector (k_i), and a value vector (v_i). The attentional system receives all Q, K, and V arrays as input and uses several parallel attention heads. The motivation for using a multi-head structure is to explore multiple subspaces since each head gets a different projection of the data. Each head learns a different aspect of attention to the input, calculating different attentional distributions. Having multiple heads on the Transformer is similar to having multiple feature extraction filters on CNNs. The head outputs an attentional mask that relates all queries to a certain key. In a simplified way, the operation performed by a head is a matrix multiplication between a matrix of queries and keys.

Finally, the data is added to the residual output from the previous layer and normalized, representing the encoder output. This data is input to the next encoder. The last encoder's data are transformed into the attention matrices K_{encdec} and V_{encdec}. They are input to all decoder layers. This data help the decoder to focus on the appropriate locations in the input sequence. The decoder has two layers of attention, Feed-Foward layers, and normalization layers. The attentional layers are the masked multi-head attention and the decoder multi-head attention.

The masked multi-head attention is very similar to the encoder multi-head attention. The main difference is that the attention matrices Q, K, and V are created only with the previous data words, masking future positions with $-\infty$ values before the softmax step. The decoder multi-head attention is equal to the encoder multi-head attention, except it creates the Q matrix from the previous layer data and uses K_{encdec} and V_{encdec} matrices of the encoder output. The K_{encdec} and V_{encdec} matrices are the memory structure of the network. They store context information of the input sequence. Given the previous words in the output decoder, the relevant information is selected in memory to predict the next word. Finally, a linear layer followed by a softmax function projects the decoder vector of the last decoder into a probability vector. In this vector, each position defines the probability of the output word being a given vocabulary word. At each time step t, the position with the highest probability value is chosen, and the word associated with it is the output.

On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, the model outperforms the BLEU score, establishing a new state-of-the-art score of 28.4 to German translation and less than 1/4 training cost to French translation. Currently,

this model contributes significantly to NLP by the new developments. The BERT (Devlin et al. 2019), a transformer-based model, allowed significant advances in transfer learning. It obtains new state-of-the-art results on eleven natural language processing tasks, including improvements of the 7.7% point absolute in the GLUE (Wang et al. 2018a) score, 4.6% absolute improvement in the MultiNLI accuracy, 1.5 and 5.1 points improvements in the SQuAD v1.1 (Rajpurkar et al. 2016) and SQuAD v2.0 (Savchuk et al. 2018) question answering tests, respectively.

In QA, alignment usually occurs between a query and the content, looking for key terms to answer the question. The classic QA approaches do not support very long sequences and fail to model the meaning of context-dependent words correctly. Different words can have different meanings, which increases the difficulty of extracting each sentence's essential semantic logical flow in different paragraphs of context. These models are unable to address uncertain situations that require additional information to answer a particular question. In contrast, attention networks allow rich dialogues through addressing mechanisms for explicit memories or alignment structures in the query-context and context-query directions. To minimize such problems, external memory networks, BERT, and other self-attentive neural networks are the most significant developments. The End-to-End Memory Networks (Fig. 18b) is an key development. It uses attentional interfaces to manipulate external memories. Attention to looks for the memory elements most related to query q using an alignment function that dispenses the RNNs' complex structure. It consists of memory and a stack of identical attentional systems. Each layer i takes as input set $\{x_1, ..., x_N\}$ to store in the memory. The input set is converted in memory vectors $\{m_1, ..., m_N\}$ and $\{h_1, ..., h_N\}$, in the simplest case using the embedding matrix $A^i \in \mathbb{R}^{d \times V}$ to generate each $m_i \in \mathbb{R}^d$, and the matrix $C^i \in \mathbb{R}^{d \times V}$ to generate each $h_i \in \mathbb{R}^d$. In the first, layer the query q is also embedded, via embedding matrix B^1 to obtain an internal state u^1. From the second layer, the internal state u^{i+1} is the sum of the i layer output and the internal state u^i. Finally, the last layer generates $\hat{a}$.

Attention also contributes to summarize or classify texts/documents. It mainly helps build more effective embeddings that generally consider contextual, semantic, and hierarchical information between words, phrases, and paragraphs. Specifically, in summarization tasks, attention minimizes critical problems involving: (1) modeling of keywords; (2) summary of abstract sentences; (3) capture of the sentence's hierarchical structure; (4) repetitions of inconsistent phrases; and (5) generation of short sentences preserving their meaning. Hierarchical models of attention in encoder–decoder frameworks are quite typical in the field. Nallapati et al. (2016) has developed an important approach based on HAN (Seo et al. 2016) and Ptr-Nets Vinyals et al. (2015a) in which a pointer generator (i.e., selective attention mechanism) deals with out-of-vocabulary words. Intuitively, the mechanism acts as an identifier for keywords rare within the text but extremely important for summarization. In the decoder, the model has a switch capable of deciding between generating a word in a normal fashion or pointing to a word position from the source sequence. Then word in position is copied directly to the summary. At the same time, the hierarchical structure simultaneously captures important sentences and important words within each sentence, creating context-based embeddings at the word and sentence level.

Several other growing areas are exploring attention in NLP. Most of them explore frameworks such as RNNSearch, and numerous transformer-based models quickly emerge in several applications.

Table 2 Summary state-of-art approaches in computer vision sub-areas

Application	References
Computer vision	
Action recognition	Sharma et al. (2015), Girdhar et al. (2019), Girdhar and Ramanan (2017), Song et al. (2017b), Zhang et al. (2018d), Chen et al. (2018e), Liu et al. (2017d), Bazzani et al. (2017), Du et al. (2017), Li et al. (2018h), Ma et al. (2018a), Wang et al. (2018h), Li et al. (2018a), Zhang et al. (2017c), Zheng et al. (2019b), Gammulle et al. (2017), Wang et al. (2016c), Fan et al. (2018b), Si et al. (2019), Sudhakaran et al. (2019) and Liu et al. (2017b)
Counting crowds	Liu et al. (2018a), Liu et al. (2018b), Hossain et al. (2019), Zhang et al. (2019o) and Liu et al. (2019a)
Depth estimation	Xu et al. (2018a) and Liu et al. (2019b)
Facial detection	Tian et al. (2018b) , Zhang et al. (2019n) and Xiao et al. (2016)
Facial expression recognition	Xie et al. (2019c), Minaee and Abdolrashidi (2019) and Li et al. (2018g)
Image classification	Fu et al. (2017), Serra et al. (2018), Han et al. (2018), Sermanet et al. (2015), Hackel et al. (2018), Kang et al. (2018), Mnih et al. (2014), Chen et al. (2018e), Ke et al. (2018), Rodríguez et al. (2018), Wang et al. (2017a), Jaderberg et al. (2015), Seo et al. (2016), Vinyals et al. (2016), Hu et al. (2020a), Wang et al. (2017c), Bello et al. (2019), Wu et al. (2018b), Zhao et al. (2017a), Zhao et al. (2017a), Ye et al. (2018), Pesce et al. (2017), Mei et al. (2019), Zhao et al. (2018a), Yin et al. (2019), Xiao et al. (2015), Fukui et al. (2019b), Woo et al. (2018), Hu et al. (2019b), Guan et al. (2018), Fang et al. (2019a), Zhang et al. (2018b), Wang et al. (2018e), Kim et al. (2018b), Ren et al. (2019), Wang et al. (2018d) and Doughty et al. (2019), Peng et al. (2017)
Image restoration	Zhang et al. (2019j), Suganuma et al. (2019) and Qian et al. (2018)
Information retriever	Ji et al. (2018), Jin et al. (2020) and Yang et al. (2019d)
Image generation	Gregor et al. (2015) Zhang et al. (2019a), Kastaniotis et al. (2018), Yu et al. (2018b), Bornschein et al. (2017), Reed et al. (2018), Parmar et al. (2018), Chen et al. (2018d), Li et al. (2019d) and Tang et al. (2019a)
Image-to-Image Translation	Mejjati et al. (2018) and Tang et al. (2019b)
Medical image analysis	Kastaniotis et al. (2018), Schlemper et al. (2019) and Jiang et al. (2019)
Object recognition	Chu et al. (2017a), Zheng et al. (2017), Zheng et al. (2018), Liu et al. (2017e), He et al. (2018a), Chen et al. (2018c), Si et al. (2018), Song et al. (2018), and Zhou and Shao (2018)
Object detection	Zhang et al. (2018h), Kong et al. (2018b), Li et al. (2018b), Chen et al. (2018b), Chen and Li (2019) and Wang et al. (2019c)
Others	Lee et al. (2019a), Meng et al. (2018), Park et al. (2019), Ilse et al. (2018), Lee and Osindero (2016), Chu et al. (2017b), Zhang et al. (2018k), Lu et al. (2018), Yuan et al. (2019), Yang et al. (2017b) and Yang et al. (2019c)
Person detection	Zhang et al. (2019c) and Zhang et al. (2018e)

Table 2 (continued)

Application	References
Person recognition	Li et al. (2018f), Liu et al. (2017a), Liao et al. (2018), Chen et al. (2018a), Zhao et al. (2017b), Zhou et al. (2017), Xu et al. (2018b), Li et al. (2018e), Zheng et al. (2019a), Ouyang et al. (2019), Wang et al. (2018b), Xu et al. (2017b) and Zhang et al. (2019g)
Segmentation	Fu et al. (2019), Li et al. (2018d), Ren and Zemel (2017), Zhang et al. (2019f), Chen et al. (2016b), Jetley et al. (2018), Liu et al. (2019c), Yuan and Wang (2018), Li et al. (2018c) and Hu et al. (2018), Zeng et al. (2019b), Shuai et al. (2017), Hu et al. (2019c), Oktay et al. (2018) Chen et al. (2019b), Li et al. (2019g), Kong and Fowlkes (2019), Li and Loy (2018) and Zhao et al. (2018b)
Saliency detection	Kuen et al. (2016), Liu et al. (2018c), Liu et al. (2018c), Cornia et al. (2018) and Hu et al. (2020b)
Text recognition	He et al. (2018b), Cheng et al. (2017a), Luo et al. (2019), Xie et al. (2019b), Li et al. (2019b), and Cheng et al. (2017b)
Tracking targets	Fu et al. (2019), Li et al. (2018d), Ren and Zemel (2017), Zhang et al. (2019f), Chen et al. (2016b), Jetley et al. (2018), Liu et al. (2019c), Yuan and Wang (2018), Li et al. (2018c), Hu et al. (2018), Zeng et al. (2019b), Shuai et al. (2017), Hu et al. (2019c) and Oktay et al. (2018), Chen et al. (2019b)
Transfer learning	Zagoruyko and Komodakis (2017), Zhang et al. (2019i) and Li et al. (2019i)
Text detection	He et al. (2018b), Wojna et al. (2017), He et al. (2017) and Bhunia et al. (2019)
Video classification	Bielski and Trzcinski (2018) and Long et al. (2018)

5.2 Computer vision (CV)

Visual attention has become popular in many CV tasks. Action recognition, counting crowds, image classification, image generation, object detection, person recognition, segmentation, saliency detection, text recognition, and tracking targets are the most explored sub-areas, as shown in Table 2. Applications in other sub-areas still have few representative works, such as clustering, compression, deblurring, depth estimation, image restoration, among others, as shown in Table 2.

Visual attention in image classification tasks was first addressed by Mnih et al. (2014). In this domain, there are sequential approaches inspired by human saccadic movements (Mnih et al. 2014) and feedforward-augmented structures CNNs (Sect. 4). The general goal is usually to amplify fine-grained recognition, improve classification in the presence of occlusions, sudden variations in points of view, lighting, and rotation. Some approaches aim to learn to *look* at the most relevant parts of the input image, while others try to discern between discriminating regions through feature recalibration and ensemble predictors via attention. To fine-grained recognition, important advances have been

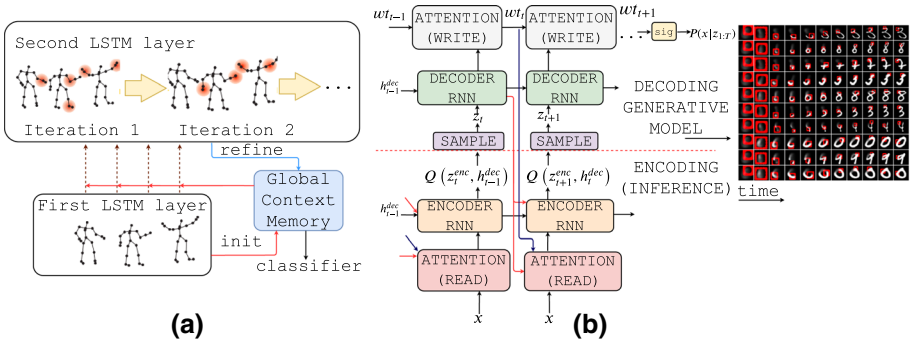

Fig. 20 Illustration of global context-aware attention (Liu et al. 2017c) for action recognition, and DRAW (Gregor et al. 2015) for image generation. **a** The Context-Aware Attention Network. The first LSTM layer encodes the skeleton sequence and generates an initial global context memory. The second layer performs attention over the inputs with global context memory assistance and generates a refined representation. The refine representation is then used back to the global context. Multiple attention iterations are carried out to refine the global context progressively. Finally, memory information is used for classification. **b** The DRAW architecture. At each time step t, the input is read by attention and passed to the encoder RNN. The encoder's output is used to compute the approximate posterior over the latent variables. On the right, an illustration shows the iterative process of generating some images. Each row shows successive stages in the generation of a single digit. The network guided by attention draws and refine regions successively. The red rectangle delimits the area attended to by the network at each time-step

achieved through recurrent convolutional networks in the classification of bird subspecies (Fu et al. 2017) and architectures trained via RL to classify vehicle subtypes (Zhao et al. 2016).

Visual attention also provides significant benefits for action recognition tasks by capturing spatio-temporal relationships. The biggest challenge's classical approaches are capturing discriminative features of movement in the sequences of images or videos. The attention allows the network to focus the processing only on the relevant joints or on the movement features easily. Generally, the main approaches use the following strategies: (1) saliency maps: spatiotemporal attention models learn *where to look* in video directly human fixation data. These models express the probability of saliency for each pixel. Deep 3D CNNs extract features only high saliency regions to represent spatial and short time relations at clip level, and LSTMs expand the temporal domain from few frames to seconds (Bazzani et al. 2017; 2) self-attention: modeling context-dependencies. The person being classified is the Query (Q), and the clip around the person is the memory, represented by keys (K) and values (V) vectors. The network process the query and memory to generate an updated query vector. Intuitively self-attention adds context to other people and objects in the clip to assist in subsequent classification (Girdhar et al. 2019; 3) recurrent attention mechanisms: captures relevant positions of joints or movement features and, through a recurring structure, refines the attentional focus at each time step (Liu et al. 2017c; Du et al. 2017); and (4) temporal attention: captures relevant spatial-temporal locations (Li et al. 2018h, h; Song et al. 2017b; Xin et al. 2016; Zang et al. 2018a; Pei et al. 2017).

Liu et al. (2017c) model is a recurrent attention approach to capturing the person's relevant positions. This model presented a pioneering approach using two layers of LSTMs and the context memory cell that recurrently interact with each other, as shown in Fig. 20a. First, a layer of LSTMs generates an encoding of a skeleton sequence, initializing the

context memory cell. The memory representation is input to the second layer of LSTMs and helps the network selectively focus on each frame's informational articulations. Finally, attentional representation feeds back the context memory cell to refine the focus's orientation by paying attention more reliably. Experiments on NTU RGB+D (Shahroudy et al. 2016), UT-Kinect (Xia et al. 2012), and SBU-Kinect Interaction (Yun et al. 2012) datasets show accuracy higher than 82% in the test sets. At the same time, the classical approaches reached 77%. Similarly, Du et al. (2017) proposed RPAN - a recurrent attention approach between sequentially modeling by LSTMs and convolutional features extractors. First, CNNs extract features from the current frame. Then, the attentional mechanism guided by the LSTM's previous hidden state estimates a series of features related to human articulations related to the semantics of movements of interest. Then, these highly discriminative features feed LSTM time sequences.

In image generation, there were also notable benefits. DRAW (Gregor et al. 2015) introduced visual attention with an innovative approach—image patches are generated sequentially and gradually refined, in which to generate the entire image in a single pass (Fig. 20b). The model was trained and tested using three datasets of distinct visual complexities: the cluttered MNIST (LeCun et al. 1998), the Street View House Numbers (SVHN) (Netzer et al. 2011), and the CIFAR-10 (Krizhevsky et al. 2009). On MNIST, the network hits an error of 3.36% while CNNs with two layers hit 14.35%. Without attentional mechanisms, the model performs as well as generative models DARN (Gregor et al. 2014), and DBMS (Salakhutdinov and Hinton 2009). However, MNIST is a simple dataset in terms of visual structures. In the SVHN dataset, the model is capable of generating highly realistic street digit images. In the CIFAR-10, the model's main challenge is to deal with a few diversified and low-resolution natural images. In this case, the model also showed good performance and correctly captured objects' shape, color, and composition from the original images.

Subsequently, attentional mechanisms emerged in GANs to minimize the challenges in modeling images with structural constraints. Naturally, GANs efficiently synthesize elements differentiated by texture (i.e., oceans, sky, natural landscapes) but suffer to generate geometric patterns (i.e., faces, animals, people, fine details). The central problem is the convolutions that fail to model dependencies between distant regions. Besides, the statistical and computational efficiency of the model suffers from the stacking of many layers. The attentional mechanisms, especially self-attention, offered a computationally inexpensive alternative to easily model long-range dependencies. Self-attention as a complement to convolution contributes significantly to the advancement of the area with approaches capable of generating fine details (Zhang et al. 2019a), high-resolution images, and intricate geometric patterns (Chen et al. 2020). The Self-Attention Generative Adversarial Networks (SAGAN) (Zhang et al. 2019a) is the main development in the area. It demonstrates the feasibility of using self-attention to support convolutional operations in unsupervised training. SAGAN significantly improves the best-published Inception score from 36.8 to 52.52. The FID results obtained (18.65 in the interior and 83.7 in the intra) also pointed out that self-attention helps the network to get closer to the distribution of data.

In expression recognition, attention optimizes the entire segmentation process. It scans inputs as a whole sequence, choosing the most relevant region to describe a segmented symbol or implicit space operator (Zhang et al. 2017b). In information retriever, it helps obtain appropriate semantic resources using individual class semantic resources. Progressively, it orients visual aids to generate an attention map to ponder the importance of different local regions. (Ji et al. 2018). In medical image analysis, attention helps implicitly learn to suppress irrelevant areas in an input image while highlighting useful resources for

a specific task. This allows us to eliminate the need to use explicit external tissue/organ localization modules using CNNs. Besides, it allows generating both images and maps of attention in unsupervised learning useful for data annotation. For this, there are the ATA-GANS (Kastaniotis et al. 2018) and attention gate (Schlemper et al. 2019) modules that work with unsupervised and supervised learning, respectively.

For person recognition, attention has become essential in in-person re-identification (re-id) (Han et al. 2018; Zheng et al. 2019a; Li et al. 2018f). Re-id aims to search for people seen from a surveillance camera implanted in different locations. In classical approaches, the bounding boxes of detected people were not optimized for re-identification. Thus, suffering from misalignment problems, background disorder, occlusion, and absent body parts. Misalignment is one of the biggest challenges, as people are often captured in various poses, and the system needs to compare different images. In this sense, neural attention models started to lead the developments mainly with multiple attentional alignment mechanisms between different bounding boxes.

There are still less popular applications, but for which attention plays an essential role. Self-attention models iterations between the input set for clustering tasks (Lee et al. 2019a). Attention refines and merges multi-scale feature maps in-depth estimation and edge detection (Xu et al. 2017a, 2018a). In video classification, attention helps capture global and local resources generating a comprehensive representation (Xie et al. 2019a). It also measures each time interval's relevance in a sequence (Pei et al. 2017), promoting a more intuitive interpretation of the impact of content on the video's popularity, providing the regions that contribute the most to the prediction (Bielski and Trzcinski 2018). In face detection, attention dynamically selects the main reference points of the face (Xiao et al. 2016). It improves deblurring in each convolutional layer in deblurring, preserving fine details (Park et al. 2019). Finally, in emotion recognition, it captures complex relationships between audio and video data by obtaining regions where both signals relate to emotion (Zhang et al. 2019m).

5.3 Multimodal tasks (CV/NLP)

Attention has been used extensively in multimodal learning, mainly for mapping complex relationships between different sensory modalities. In this domain, the importance of attention is quite intuitive, given that communication and human sensory processing are completely multimodal. The first approaches emerged from 2015, inspired by an attentive encoder–decoder framework entitled "Show, attend and tell: Neural image caption generation with visual attention" by Xu et al. (2015). In this framework, depicted in Fig. 21a at each time t, attention generates a vector with a dynamic context of visual features based on the words previously generated—a principle very similar to that presented in RNN-Search (Bahdanau et al. 2015). This framework collaborated with quite significant quantitative and qualitative results. Tests performed on Flickr9k, Flickr30k,[3] and MS COCO (Lin et al. 2014) even computed improvements of more than 20% using the BLEU-1 metric against the Google NIC (Vinyals et al. 2015c) and Log Bilinear models (Kiros et al. 2014). Qualitatively, results demonstrated that the learned alignments correspond well with human intuition. Hence, opening space for a vast area of research for attention and multimodality. Later, more elaborate methods using visual and textual sources were developed

[3] Publicly available on https://www.kaggle.com/hsankesara/flickr-image-dataset.

Table 3 Summary of state-of-art approaches in multimodal tasks (CV/NLP)

Application	References
Multimodal tasks	
Emotion recognition	Tan et al. (2019), Zadeh et al. (2018a) and Zadeh et al. (2018b)
Expression comprehension	Yu et al. (2018c)
Image captioning	Anderson et al. (2018), Zhu et al. (2018b), Xu et al. (2015), Ma et al. (2018b), Lu et al. (2017b), You et al. (2016), Chen et al. (2017b), He et al. (2019a), Huang et al. (2018b) Fang et al. (2015), Yang et al. (2016d), Pedersoli et al. (2017), Yao et al. (2018), Li et al. (2019k), Liang et al. (2017), Zhang et al. (2019h), Rohrbach et al. (2016), Kaiser et al. (2017), Lee et al. (2018b), Pan et al. (2020) and Anderson et al. (2018)
Image-to-text generation	Poulos and Valle (2021)
Image classification	Wang et al. (2018g)
Text-to-image generation	Xu et al. (2018d)
Visual question answering	Kim et al. (2020b), Jiang et al. (2020), Liang et al. (2018a), Hudson and Manning (2018), Gülçehre et al. (2019) Osman and Samek (2019) Lu et al. (2016) Nam et al. (2017) Deng et al. (2018) Kim et al. (2018a), Nguyen and Okatani (2018), Andreas et al. (2015), Kim et al. (2016), Shih et al. (2016), Xiong et al. (2016), Zhu et al. (2016), Lu et al. (2017a), Yu et al. (2018d), Agrawal et al. (2018), Zhang et al. (2018j) and Gan et al. (2019), Xu and Saenko (2016), Zhang et al. (2019k), Zhang et al. (2019p), Liang et al. (2019), Kim et al. (2020a), Yang et al. (2016b), Yu et al. (2017a), Anderson et al. (2018), Yu et al. (2017b)
Video captioning	Cho et al. (2015), Wu et al. (2018a), Zhou et al. (2018b), Zhu and Yang (2020) Pu et al. (2018), Yao et al. (2015), Yu et al. (2016), Hori et al. (2017a), Krishna et al. (2017), Ma et al. (2018a), Bin et al. (2018), Zhou et al. (2018b), Baraldi et al. (2017), Olivastri et al. (2019), Ji et al. (2019), Song et al. (2017a), Li et al. (2019j), Gao et al. (2017) and Wang et al. (2018c)

mainly in image captioning, video captioning, and visual question answering, as shown in Table 3.

For image captioning, Yang et al. (2016d) extended the seminal framework by Xu et al. (2015) with review attention, a sequence of modules that capture global information in various stages of reviewing hidden states and generate more compact, abstract, and global context vectors. Zhu et al. (2018b) presented a triple attention model which enhances object information at the text generation stage. Two attention mechanisms capture semantic visual information in input, and a mechanism in the prediction stage integrates word and image information better. Lu et al. (2017b) presented an adaptive attention encoder–decoder framework that decides when to trust visual signals and when to trust only the language model. Specifically, their mechanism has two complementary elements. The *visual sentinel* vector decides when to look at the image, and the *sentinel gate* decides how much new information the decoder wants from the image. Recently, Pan et al. (2020)

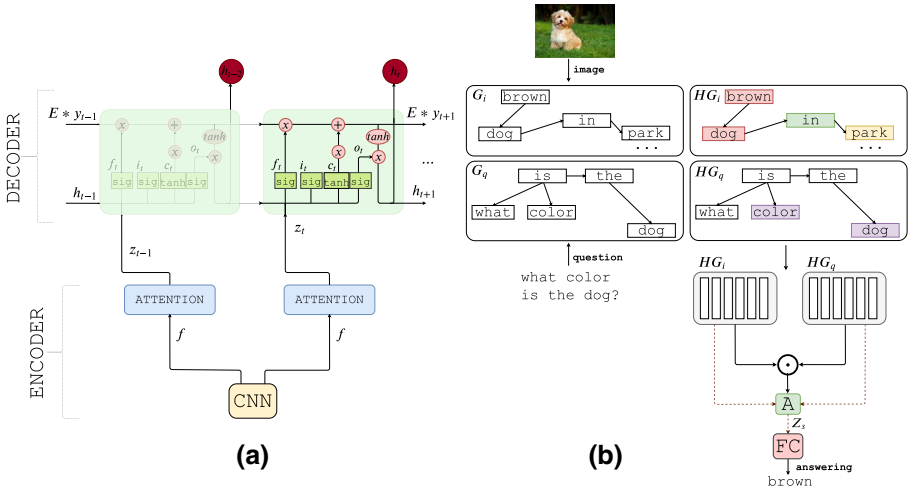

Fig. 21 Illustration of classic attentive encoder–decoder framework by Xu et al. (2015) and Hypergraph Attention Network (Kim et al. 2020a) for question answering tasks. **a** The *show, attend and tell* framework. At each time step t, attention takes as input visual feature maps and previous hidden state of the decoder and produces a dynamic context vector with essential visual features to predict the next word. **b** The Hypergraph Attention Network. For a given pair of images and questions, two symbolic graphs G_i, and G_q are constructed. After, two hypergraphs HG_i, and HG_q with random-walk hyperedge are constructed and combined via co-attention map A. Finally, the final representation Z_s is used to predict an answer for the given question

created attentional mechanisms based on bilinear pooling capable of capturing high order interactions between multi-modal features, unlike the classic mechanisms that capture only first-order feature interactions.

Similarly, in visual question-answering tasks, methods seek to align salient textual features with visual features via feedforward or recurrent soft attention methods (Anderson et al. 2018; Osman and Samek 2019; Lu et al. 2016; Yang et al. 2016b). Liang et al. (2018a) used attention to capture hierarchical relationships between sequences of image-text pairs not directly related. The objective is to answer questions and justify what results in the system were based on answers. More recent approaches aim to generate complex inter-modal representations. In this line, Kim et al. (2020a) proposed Hypergraph Attention Networks (HANs), a solution to minimize the disparity between different levels of abstraction from different sensory sources. So far, HANs is the first approach to define a common semantic space with symbolic graphs of each modality and extract an inter-modal representation based on co-attention maps in the constructed semantic space, as shown in Fig. 21b. HANs contributes significantly to AI; it presents an attentional interface as a key element for multimodality and as a link between symbolic and connectionist representations. In addition, to evaluate the model, they employed quite challenging datasets for the area. The Graph Question Answering (GQA) (Hudson and Manning 2019) challenges the model to capture several facts to answer a given question. In the VQA v2 (Antol et al. 2015), each question is associated with multiple possible answers. Compared to BAN (Kim et al. 2018a) and MFB (Yu et al. 2018e) models, the co-attention maps learned by HANs were quite effective for performing the task, presenting accuracy of approximately 70% in test sets, while the state-of-the-art approaches presented only approximately 54%.

For video captioning most approaches generally align textual features and spatio-temporal representations of visual features via simple soft attention mechanisms (Cho et al. 2015; Yu et al. 2016; Yao et al. 2015; Hori et al. 2017a). For example, Pu et al. (2018) design soft attention to adaptively emphasize different CNN layers while also imposing attention within local spatiotemporal regions of the feature maps at particular layers. These mechanisms define the importance of regions and layers to produce a word based on word-history information. Recently, self-attention mechanisms have also been used to capture more complex and explicit relationships between different modalities. Zhu and Yang (2020) introduced ActBERT, a transformer-based approach trained via self-supervised learning to encode complex relations between global actions and local, regional objects and linguistic descriptions. Zhou et al. (2018b) proposed a multimodal transformer via supervised learning, which employs a masking network to restrict its attention to the proposed event over the encoding feature.

Other applications also benefit from the attention. The main approaches use memory fusion structures inspired by the human brain's communication understanding mechanisms in the emotion recognition domain. Biologically, different regions process and understand different modalities connected via neural links to integrate multimodal information over time. Similarly, in existing approaches, attentional components model view-specific dynamics within each modality via recurrent neural networks. A second component simultaneously finds multiple cross-view dynamics in each recurrence timestep by storing them in hybrid memories. Memory updates occur based on all the sequential data seen. Finally, to generate the output, the predictor integrates the two levels of information: view-specific and multiple cross-view memory information (Zadeh et al. 2018a, b).

There are still few multimodal methods for classification. Wang et al. (2018g) presented a pioneering framework for classifying and describing image regions simultaneously from textual and visual sources. Their framework detects, classifies, and generates explanatory reports regarding abnormalities observed in chest X-ray images through multi-level attentional modules end-to-end in LSTMs and CNNs. In LSTMs, attention combines all hidden states, generating a dynamic context vector. Then, a spatial mechanism guided by a textual mechanism highlights the regions of the image with more meaningful information. Intuitively, the salient features of the image are extracted based on high-relevance textual regions.

5.4 Recommender systems (RS)

Attention has also been used in recommender systems for the behavioral modeling of users. Capturing user interests is a challenging problem for neural networks, as some iterations are transient, some clicks are unintentional, and interests can change quickly in the same session. Classical approaches (i.e., Markov Chains and RNNs) have limited performance predicting the user's next actions. They also present different performances in sparse and dense datasets and have long-term memory problems. In this sense, attention has been used mainly to assign weights to a user's interacted items capturing long and short-term interests more effectively than traditional ones. Self-attention and memory approaches have been explored to improve the area's development. STAMP (Liu et al. 2018d) model, based on attention and memory, manages users' general interests in long-term memories and current interests in short-term memories resulting in behavioral representations that are more coherent. This model presented predictive accuracy and average of reciprocal

Table 4 Summary of main state-of-art approaches in reinforcement learning tasks

Application	References
Reinforcement learning	
Computer vision	Ba et al. (2015), Li et al. (2019a), Rao et al. (2017), Stollenga et al. (2014), Cao et al. (2017), Zhao et al. (2016), Hu et al. (2019a), Edel and Lausch (2016) and Yeung et al. (2016)
Graph reasoning	Lee et al. (2018a)
Natural language processing	Santoro et al. (2018)
Navigation	Mishra et al. (2018), Parisotto and Salakhutdinov (2018), Zambaldi et al. (2019), Santoro et al. (2018) and Baker et al. (2020)

ranks superior to the classic Item-KNN (Sarwar et al. 2001), GRU4Rec (Hidasi et al. 2015), GRU4Rec+ (Tan et al. 2016), and NARM (Li et al. 2017) models using the Yoochoose (Ben-Shimon et al. 2015) and Diginetica[4] datasets. Very significant results were obtained mainly using the Yoochoose sessions. The results showed that when session size increases by 4 to 30, the performance of the STAMP model drops from 70% to 60%. The classic NARM model drops from 70% to approximately 25%, demonstrating the efficiency of attention in capturing key user behaviors in longer sessions.

Finally, the Collaborative Filtering (ACF) (Chen et al. 2017a) framework explored soft attention in capturing long-term semantics for finding the most relevant items in user's history. Results from the hit ratio (HR) and normalized discounted cumulative gain (NDCG) metrics using the Pinterest (Geng et al. 2015) and Vine (Chen 2016) datasets demonstrated the robustness of the ACF when compared to state-of-the-art field methods (e.i., Item-KNN (Sarwar et al. 2001), SVD++ (Koren 2008), and DeepHybrid (Van Den Oord et al. 2013)). The results demonstrate that the ACF performance is higher than other models, even in very sparse datasets, such as Vine. The ACF without attentional mechanism degrades the performance. In contrast, when attention is applied at the items and components' level, the performance of the recommendation based on multimedia data is significantly improved.

5.5 Reinforcement learning (RL)

Attention has been gradually introduced in reinforcement learning to deal with unstructured environments in which rewards and actions depend on past states and are challenging to guarantee the Markov property. Specifically, the goals are to increase the agent's generalizability and minimize long-term memory problems. Currently, the main attentional reinforcement learning approaches are computer vision, graph reasoning, natural language processing, and virtual navigation, as shown in Table 4.

Some approaches use attention in the policy network to increase the ability to generalize in partially observable environments. Mishra et al. (2018) used attention to easily capture long-term temporal dependencies in convolutions in an agent's visual navigation task in

[4] Publicly available on http://cikm2016.cs.iupui.edu/cikm-cup.

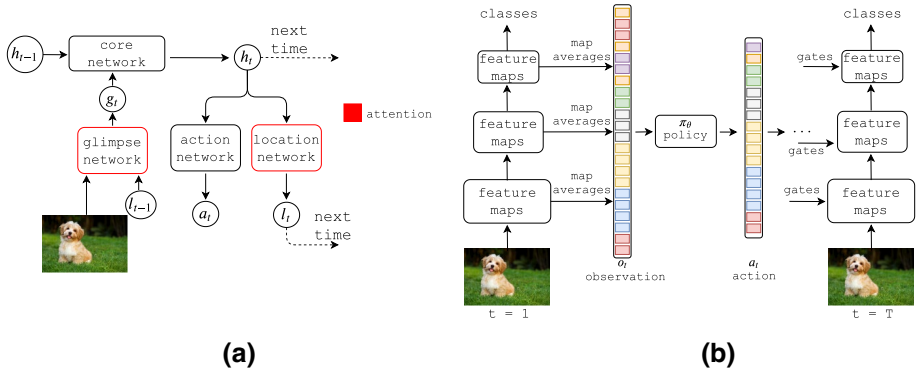

Fig. 22 Illustration of RAM (Mnih et al. 2014) and dasNet (Stollenga et al. 2014) for image classification tasks. **a** The RAM framework. At each time step t, the glimpse networks extracts a retina-like representation based on input image and an focus location l_{t-1}. The core network takes the glimpse representation as input and combining it with the internal representation at the previous time step and produces the new internal state of the model h_t. The location network and action network use the h_t to produce the next location l_t to attend and the classification. **b** The dasNet network. Each image is classified after T passes through CNNs. After each forward propagation, the averages of feature maps are combined into an observation vector o_t that is used by a deterministic policy to choose an action a_t that changes the focus of all the feature maps for the next pass of the same image. Finally, after pass T times, the output is used to classify the image

random mazes. At each time step t, the model receives as input the current observation o_t and previous sequences of observations, rewards, and actions so that attention allows the policy to maintain a long memory of past episodes. Other approaches implement attention directly to the representation of the state. State representation is a classic and critical problem in RL. As the state space dimension is one of the major bottlenecks for speed, efficiency, and generalization of training techniques. In this sense, the importance of attention on this topic is quite intuitive.

However, there are still few approaches exploring the representation of states. The neural map (Parisotto and Salakhutdinov 2018) maintains an internal memory in the agent controlled via attention mechanisms. While the agent navigates the environment, an attentional mechanism alters the internal memory, dynamically constructing a history summary. At the same time, another generates a representation o_t, based on the contextual information of the memory and the state's current observation. Then, the policy network receives o_t as input and generates the distribution of shares. The model managed to solve an average of 96% of the labyrinths in the test set of the Goal-Search (Oh et al. 2016) environment against LSTMs (57.4%) and memory networks (83.4%). Some more recent approaches affect the representation of the current o_t observation of the state via self-attention in iterative reasoning between entities in the scene (Zambaldi et al. 2019; Baker et al. 2020), or between the current observation o_t and memory units (Santoro et al. 2018) to guide model-free policies.

The most discussed topic is the use of the policy network to guide the attentional focus of the agent's glimpses sensors on the environment so that the representation of the state refers to only a small portion of the entire operating environment. This approach emerged initially by (Mnih et al. 2014) using policy gradient methods (i.e., REINFORCE algorithm) in the hybrid training of recurrent networks in image classification tasks. Their model consists of a glimpse sensor that captures only a portion of the input image, a core network that maintains a summary of the history of patches seen by the agent, an action network

Table 5 Summary state-of-art main approaches in robotics

Application	References
Robotics	
Control	Duan et al. (2017) Abolghasemi et al. (2019)
Human–robot interaction	Zang et al. (2018b)
Navigation	Sadeghian et al. (2019) Vemula et al. (2018) Chen et al. (2019a) Fang et al. (2019b)
Visual odometry	Xue et al. (2020) Johnston and Carneiro (2020) Kuo et al. (2020) Damirchi et al. (2020) Gao et al. (2020a) Li et al. (2021)

that estimates the class of the image seen, and a location network trained via RL which estimates the focus of the glimpse on the next time step, as shown in Fig. 22a. This structure considers the network as the agent, the image as the environment, and the reward is the number of correct network ratings in an episode. Although the approach is quite promising, it has been successfully tested only on the MNIST[5] dataset on 28×28 and 64×64 images. The results in the test sets show misclassification 1.84% against 7.56% and 2.31% of feed-forward and convolutional neural networks.

Stollenga et al. (2014) proposed a similar approach, however directly focused on CNNs, as shown in Fig. 22b. The structure allows each layer to influence all the others through attentional bottom-up and top-down connections that modulate convolutional filters' activity. After supervised training, the attentional connections' weights implement a control policy via RL and SNES (Schaul et al. 2011). The policy learns to suppress or enhance features at various levels by improving the classification of difficult cases not captured by the initial supervised training. The method improves over the state-of-the-art reference implementation using CIFAR-100[6] dataset. Subsequently, variants similar to these approaches appeared in multiple image classification (Ba et al. 2015; Edel and Lausch 2016; Zhao et al. 2016), action recognition (Yeung et al. 2016), and face hallucination (Cao et al. 2017).

5.6 Robotics

In robotics, there are still few applications with neural attentional models. A small portion of current work is focused on control, visual odometry, navigation, and human–robot interaction, as shown in Table 5.

Navigation and visual odometry (VO) are the most explored domains, although still with very few published works. For classic DL approaches, navigating tasks in real or complex environments are still very challenging. These approaches have limited performance in dynamic and unstructured environments and over long horizon tasks. In real environments, the robot must deal with dynamic and unexpected changes in humans and other obstacles around it. Also, decision-making depends on the information received in the past and the ability to infer the environment's future state. Some seminal approaches in the literature

[5] Publicly available on http://yann.lecun.com/exdb/mnist/.

[6] Publicly available on the url https://www.cs.toronto.edu/~kriz/cifar.html.

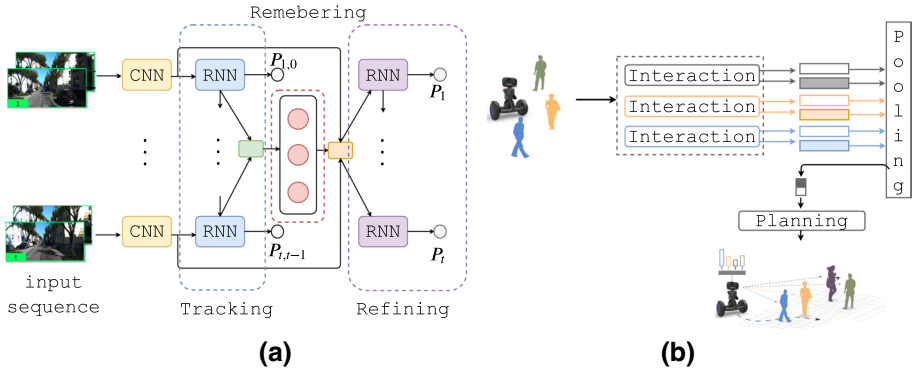

Fig. 23 Illustration of deep visual odometry with Adaptive Memory (Xue et al. 2020) and Crowd-Robot Interaction (Chen et al. 2019a) for navigation. **a** The deep visual odometry with adaptive memory framework. This framework introduces two important components called *remembering* and *refining*, as opposed to classic frameworks that treat VO as just a tracking task. *Remembering* preserves long-time information by adopting an adaptive context selection, and *refining* improves previous outputs using spatial-temporal feature reorganization mechanism. **b** Crowd–Robot Interaction network. The network consists of three modules: interaction, pooling, and planning. Interaction extracts robot–human interactions in crowds. Pooling aggregates all interaction information, and planning estimates the state's value based on robots and humans for navigation in crowds

have demonstrated the potential of attention to minimizing these problems without compromising the techniques' computational cost. Sadeghian et al. (2019) proposed Sophie: an interpretable framework based on GANs for robotic agents in environments with human crowds. Their framework via attention extracts two levels of information. A physical extractor learns spatial and physical constraints generating a C context vector that focuses on viable paths for each agent. In contrast, a social extractor learns the interactions between agents and their influence on each agent's future path. Finally, LSTMs based on GAN generate realistic samples capturing the nature of future paths. The attentional mechanisms allow the framework to predict physically and socially feasible paths for agents, achieving cutting-edge performances on several different trajectories. The Stanford Drone (Robicquet et al. 2016) dataset results showed that Sophie has an ADE error of 21.08 pixels against LSTM with 30.48, CAR-Net (Sadeghian et al. 2018) with 30.92, and S-GAN (Gupta et al. 2018) with 22.24. In the complex setups of the dataset, Sophie presented even more discrepant results, only 15.61 error, while the other models presented more than 24.

Vemula et al. (2018) proposed a trajectory prediction model that captures each person's relative importance when navigating in the crowd, regardless of their proximity, via spatio-temporal graphs. Chen et al. (2019a) proposed the crowd-aware robot navigation with attention-based deep reinforcement learning. Specifically, a self-attention mechanism models interactions between human–robot and human–human pairs, improving the robot's inference capacity of future environment states. It also captures how human–human interactions can indirectly affect decision-making, as shown in Fig. 23 in a). Fang et al. (2019b) proposed the novel memory-based policy (i.e., scene memory transformer—SMT) for embodied agents in long-horizon tasks. The SMT policy consists of two modules. A scene memory stores all past observations in an embedded form, whereas an attention-based policy network uses the updated scene memory to compute a distribution over actions. The SMT model is based on an encoder–decoder Transformer and showed strong performance as the agent moves in a large environment where the number of observations grows rapidly.

In VO, the classic learning-based methods consider the VO task a problem of pure tracking through the recovery of camera poses from fragments of the image, leading to the accumulation of errors. Such approaches often disregard crucial global information to alleviate accumulated errors. However, it is challenging to preserve this information in end-to-end systems effectively. Attention represents an alternative that is still little explored in this area to alleviate such disadvantages. Xue et al. (2020) proposed an adaptive memory approach to avoid the network's catastrophic forgetfulness. Their framework consists mainly of a memory, a remembering, and a refining module, as shown in Fig. 23b). First, it remembers to select the main hidden states based on camera movement while preserving selected hidden states in the memory slot to build a global map. The memory stores the global information of the entire sequence, allowing refinements on previous results. Finally, the refining module estimates each view's absolute pose, allowing previously refined outputs to pass through recurrent units, thus improving the next estimate. This approach reduced by more than 50% the translational and rotational RMSE errors of sequences 03, 04, 05, 06, 07, and 10 of the KITTI (Geiger et al. 2013) dataset, against 13 supervised and unsupervised models that are state of the art in the field.

Another common problem in VO classical approaches is selecting the features to derive ego-motion between consecutive frames. In scenes, there are dynamic objects and non-textured surfaces that generate inconsistencies in the estimation of movement. Recently, self-attention mechanisms have been successfully employed in dynamic reweighting of features and the semantic selection of image regions to extract more refined egomotion (Kuo et al. 2020; Damirchi et al. 2020; Gao et al. 2020a). Additionally, self-attentive neural networks have been used to replace traditional recurrent networks that consume training time and are inaccurate in the temporal integration of long sequences (Li et al. 2021).

In human–robot interaction, Zang et al. (2018b) proposed a framework that interprets navigation instructions in natural language and finds a mapping of commands in an executable navigation plan. The attentional mechanisms correlate navigation instructions very efficiently with the commands to be executed by the robot in only one trainable end-to-end model, unlike the classic approaches that use decoupled training and external interference during the system's operation. In control, existing applications mainly use manipulator robots in visual-motor tasks. Duan et al. (2017) used attention to improve the model's generalization capacity in imitation learning approaches with a complex manipulator arm. The objective is to build a one-shot learning system capable of successfully performing instances of tasks not seen in the training. Thus, it employs soft attention mechanisms to process a long sequence of (states, actions) demonstration pairs. Finally, Abolghasemi et al. (2019) proposed a deep visual engine policy through task-focused visual attention to make the policy more robust and to prevent the robot from releasing manipulated objects even under physical attacks.

5.7 Interpretability

A long-standing criticism of neural network models is their lack of interpretability (Li et al. 2016b). Academia and industry have a great interest in the development of interpretable models mainly for the following aspects: (1) critical decisions: when critical decisions need to be made (e.i., medical analysis, stock market, autonomous cars), it is essential to provide explanations to increase the confidence of the specialist human results; (2) failure analysis: an interpretable model can retrospectively inspect where bad decisions were made and understand how to improve the system; (3) verification: there is no evidence of the models'

robustness and convergence even with small errors in the test set. It is difficult to explain the influence of spurious correlations on performance and why the models are sometimes excellent in some test cases and flawed in others; and (4) model improvements: interpretability can guide improvements in the model's structure if the results are not acceptable.

Attention as an interpretability tool is still an open discussion. For some researchers, it allows inspecting the models' internal dynamics. In this case, the hypothesis is that the attentional weights' magnitude is correlated with the data's relevance for predicting the output. Li et al. (2016b) proposed a general methodology to analyze the effect of erasing particular representations of neural networks' input. They found that attentional focuses are essential to understanding networks' internal functioning when analyzing erasure effects. In Serrano and Smith (2019) the results showed that higher attentional weights generally contribute with a more significant impact to the model's decision. However, multiple weights generally do not fully identify the most relevant representations for the final decision. In this investigation, the researchers concluded that attention is an ideal tool to identify which elements are responsible for the output but do not yet fully explain the model's decisions.

Some studies have also shown that attention encodes linguistic notions relevant to understanding NLP models (Vig and Belinkov 2019; Tenney et al. 2019; Clark et al. 2019). However, Jain and Wallace (2019) showed that although attention improves NLP results, its ability to provide transparency or significant explanations for the model's predictions is questionable. Specifically, the researchers investigated the relationship between attentional weights and model results by answering the following questions: (i) to what extent do attention weights correlate with metrics of the importance of features, specifically those resulting from gradients? Moreover, (ii) do different attentional maps produce different predictions? The results showed that the correlation between intuitive metrics about the features' importance (e.i., gradient approaches, erasure of features) and attentional weights is low in recurrent encoders. Besides, the selection of features other than the attentional distribution did not significantly impact the output as attentional weights exchanged at random also induced minimal output changes. The researchers also concluded that such results depend significantly on the type of architecture, given that feedforward encoders obtained more coherent relationships between attentional weights and output than other models.

Vashishth et al. (2019) systematically investigated explanations for the researchers' distinct views through experiments on NLP tasks with single sequence models, pair sequence models, and self-attentive neural networks. The experiments showed that attentional weights in single sequences tasks work like gates and do not reflect the reasoning behind the model's prediction, justifying the observations made by Jain and Wallace (2019). However, for pair sequence tasks, attentional weights were essential to explaining the model's reasoning. Manual tests have also shown that attentional weights are highly similar to the manual assessment of human observers' attention. Recently, Wiegreffe and Pinter (2019) also investigated these issues in depth through an extensive protocol of experiments. The authors observed that attention as an explanation depends on the definition of explainability considered. If the focus is on plausible explainability, the authors concluded that attention could help interpret model insights. However, if the focus is a faithful and accurate interpretation of the model's link between inputs and outputs, results are not always positive. These authors confirmed that good alternative distributions could be found in LSTMs and classification tasks, as hypothesized by Jain and Wallace (2019). However, in some experiments, adversarial training's alternative distributions had poor performances concerning attention's traditional mechanisms. These results indicate that the attention mechanisms trained mainly in RNNs learn something significant about the relationship between

tokens and prediction, which cannot be easily hacked. In the end, they showed that attention efficiency as an explanation depends on the data set and the model's properties.

6 Trends and opportunities

Attention has been one of the most influential ideas in the Deep Learning community in recent years, with several profound advances, mainly in computer vision and natural language processing. However, there is much space to grow, and many contributions are still to appear. In this section, we highlight some gaps and opportunities in this scenario.

6.1 End-to-end attention models

Over the past eight years, most of the papers published in the literature have involved attentional mechanisms. Models that are state of the art in DL use attention. Specifically, we note that end-to-end attention networks, such as Transformers (Vaswani et al. 2017) and Graph Attention Networks (Veličković et al. 2018), have been expanding significantly and have been used successfully in tasks across multiple domains (Sect. 2). In particular, Transformer has introduced a new form of computing in which the neural network's core is fully attentional. Transformer-based language models like BERT (Devlin et al. 2019), GPT2 (Radford et al. 2019), and GPT3 (Brown et al. 2020) are the most advanced language models in NLP. Image GPT (Chen et al. 2020) has recently revolutionized the results of unsupervised learning in imaging. It is already a trend to propose Transfomer based models with sparse attentional mechanisms to reduce the Transformer's complexity from quadratic to linear and use attentional mechanisms to deal with multimodality in GATs. However, Transformer is still an autoregressive architecture in the decoder and does not use other cognitive mechanisms such as memory. As research in attention and DL is still at early stages, there is still plenty of space in the literature for new attentional mechanisms, and we believe that end-to-end attention architectures might be very influential in Deep Learning's future models.

6.2 Learning multimodality

Attention has played a crucial role in the growth of learning from multimodal data. Multimodality is extremely important for learning complex tasks. Human beings use different sensory signals all the time to interpret situations and decide which action to take. For example, while recognizing emotions, humans use visual data, gestures, and voice tones to analyze feelings. Attention allowed models to learn the synergistic relationship between the different sensory data, even if they are not synchronized, allowing the development of increasingly complex applications mainly in emotion recognition, (Zhang et al. 2019m), feelings (Tan et al. 2019), and language-based image generation (Ramesh et al. 2021). We note that multimodal applications are continually growing in recent years. However, most research efforts are still focused on relating a pair of sensory data, mostly visual and textual data. Architectures that can scale easily to handle more than one pair of sensors are not yet widely explored. Multimodal learning exploring voice data, RGBD images, images from monocular cameras, data from various sensors, such as accelerometers, gyroscopes, GPS, RADAR, biomedical sensors, are still scarce in the literature.

6.3 Cognitive elements

Attention proposed a new way of thinking about the architecture of neural networks. For many years, the scientific community neglected using other cognitive elements in neural network architectures, such as memory and logic flow control. Attention has made possible including in neural networks other elements that are widely important in human cognition. Memory Networks (Weston et al. 2014), and Neural Turing Machine (Graves et al. 2014) are essential approaches in which attention makes updates and recoveries in external memory. However, research on this topic is at an early stage. The Neural Turing Machine has not yet been explored in several application domains, being used only in simple datasets for algorithmic tasks, with a slow and unstable convergence. We believe that there is plenty of room to explore the advantages of NTM in a wide range of problems and develop more stable and efficient models. Still, Memory Networks (Weston et al. 2014) presents some developments (Sect. 2), but few studies explore the use of attention to managing complex and hierarchical structures of memory. Attention to managing different memory types simultaneously (i.e., working memory, declarative, non-declarative, semantic, and long and short term) is still absent in the literature. To the best of our knowledge, the most significant advances have been made in Dynamic Memory Networks (Kumar et al. 2016) with the use of episodic memory. Another open challenge is how to use attention to plug external knowledge into memory and make training faster. Finally, undoubtedly one of the biggest challenges still lies in including other human cognition elements such as imagination, reasoning, creativity, and consciousness working in harmony with attentional structures.

6.4 Computer vision

Recurrent Attention Models (RAM) (Mnih et al. 2014) introduced a new form of image computing using glimpses and hard attention. The architecture is simple, scalable, and flexible. Spatial Transformer (STN) (Jaderberg et al. 2015) presented a simple module for learning image transformations that can be easily plugged into different architectures. We note that RAM has a high potential for many tasks in which convolutional neural networks have difficulties, such as large, high-resolution images. However, currently, RAM has been explored with simple datasets. We believe that it is interesting to validate RAM in complex classification and regression tasks. Another proposal is to add new modules to the architecture, such as memory, multimodal glimpses, and scaling. It is interesting to explore STN in conjunction with RAM in classification tasks or use STN to predict transformations between sets of images. RAM aligned with STN can help address robostusnees to spatial transformation, learn the system dynamics in Visual Odometry tasks, enhance multiple-instance learning, addressing multiple view-points.

6.5 Capsule neural network

Capsule networks (CapsNets), a new class of deep neural network architectures proposed recently by Sabour et al. (2017), have shown excellent performance in many fields, particularly in image recognition and natural language processing. However, few studies in the literature implement attention in capsule networks. AR CapsNet (Choi et al. 2019a) implements a dynamic routing algorithm where routing between capsules is made through an attention module. The attention routing is a fast forward-pass while keeping spatial

information. DA-CapsNet (Huang and Zhou 2020) proposes a dual attention mechanism, the first layer is added after the convolution layer, and the second layer is added after the primary caps. SACN (Hoogi et al. 2019) is the first model that incorporates the self-attention mechanism as an integral layer. Recently, Tsai et al. (2020) introduced a new attentional routing mechanism in which a daughter capsule is routed to a parent capsule-based between the father's state and the daughter's vote. We particularly believe that attention is essential to improve the relational and hierarchical nature that CapsNets propose. The development of works aiming at the dynamic attentional routing of the capsules and incorporating attentional capsules of self-attention, soft and hard attention can bring significant results to current models.

6.6 Neural-symbolic learning and reasoning

According to LeCun (LeCun et al. 2015) one of the great challenges of artificial intelligence is to combine the robustness of connectionist systems (i.e., neural networks) with symbolic representation to perform complex reasoning tasks. While symbolic representation is highly recursive and declarative, neural networks encode knowledge implicitly by adjusting weights. For many decades exploring the fusion between connectionist and symbolic systems has been overlooked by the scientific community. Only over the past decade, research with hybrid approaches using the two families of AI methodologies has grown again. Approaches such as statistical relational learning (SRL) and neural-symbolic learning were proposed. Recently, attention mechanisms have been integrated into some neural-symbolic models, the development of which is still at an early stage. Memory Networks (Weston et al. 2014) (Sect. 2) and Neural Turing Machine (Graves et al. 2014) (Sect. 2) were the first initiatives to include reasoning in deep connectionist models.

In the context of neural logic programming, attention has been exploited to reason about knowledge graphs or memory structures to combine the learning of parameters and structures of logical rules. Neural Logic Programming (Yang et al. 2017a) uses attention on a neural controller that learns to select a subset of operations and memory content to execute first-order rules. Logic Attention Networks (Wang et al. 2019b) facilitates inductive KG embedding and uses attention to aggregate information coming from graph neighbors with rules and attention weights. A pGAT (Harsha Vardhan et al. 2020) uses attention to knowledge base completion, which involves the prediction of missing relations between entities in a knowledge graph. While producing remarkable advances, recent approaches to reasoning with deep networks do not adequately address the task of symbolic reasoning. Current efforts are only about using attention to ensure efficient memory management. We believe that attention can be better explored to understand which pieces of knowledge are relevant to formulate a hypothesis to provide a correct answer, which are rarely present in current neural systems of reasoning.

6.7 Incremental learning

Incremental learning is one of the challenges for the DL community in the coming years. Machine learning classifiers are trained to recognize a fixed set of classes. However, it is desirable to have the flexibility to learn additional classes with limited data without retraining in the complete training set. Attention can significantly contribute to advances in the area and has been little explored. Ren et al. (2019) were the first to introduce seminal

work in the area. They use Attention Attractor Networks to regularize the learning of new classes. In each episode, a set of new weights is trained to recognize new classes until they converge. Attention Attractor Networks helps recognize new classes while remembering the classes beforehand without revising the original training set.

6.8 Credit assignment problem (CAP)

In reinforcement learning (RL), an action that leads to a higher final cumulative reward should have more value. Therefore, more "credit" should be assigned to it than an action that leads to a lower final reward. However, measuring the individual contribution of actions to future rewards is not simple and has been studied by the RL community for years. There are at least three variations of the CAP problem that have been explored. The temporal CAP refers to identifying which actions were useful or useless in obtaining the final feedback. The structural CAP seeks to find the set of sensory situations in which a given sequence of actions will produce the same result. Transfer CAP refers to learning how to generalize a sequence of actions in tasks. Few works in the literature explore attention to the CAP problem. We believe that attention will be fundamental to advance credit assignment research. Recently, Ferret et al. (2020) started the first research in the area by proposing a seminal work with attention to learn how to assign credit through a separate supervised problem and transfer credit assignment capabilities to new environments.

6.9 Attention and interpretability

There are investigations to verify attention as an interpretability tool. Some recent studies suggest that attention can be considered reliable for this purpose. However, other researchers criticize the use of attention weights as an analytical tool. Jain and Wallace (Jain and Wallace 2019) proved that attention is not consistent with other explainability metrics and that it is easy to create distributions similar to those of the trained model but to produce a different result. Their conclusion is that changing attention weights does not significantly affect the model's prediction, contrary to research by Rudin (2018) and Riedl (2019) (Sect. 5.7). On the other hand, some studies have found how attention in neural models captures various notions of syntax and co-reference (Vig and Belinkov 2019) (Clark et al. 2019) (Tenney et al. 2019). Amid such confusion, Vashishth et al. (2019) investigated attention more systematically. They attempted to justify the two types of observation (that is, when attention is interpretable and not), employing various experiments on various NLP tasks. The conclusion was that attention weights are interpretable and are correlated with metrics of the importance of features. However, this is only valid for cases where weights are essential for predicting models and cannot simply be reduced to a gating unit. Despite the existing studies, there are numerous research opportunities to develop systematic methodologies to analyze attention as an interpretability tool. The current conclusions are based on experiments with few architectures in a specific set of applications in NLP.

6.10 Unsupervised learning

In the last decade, unsupervised learning has also been recognized as one of the most critical challenges of machine learning since, in fact, human learning is mainly

unsupervised (LeCun et al. 2015). Some works have recently successfully explored attention within purely unsupervised models. In GANs, attention has been used to improve the global perception of a model (i.e., the model learns which part of the image gives more attention to the others). SAGAN (Zhang et al. 2019a) was one of the pioneering efforts to incorporate self-attention in Convolutional Gans to improve the quality of the images generated. Image Transformer is an end-to-end attention network created to generate high-resolution images that significantly surpassed state-of-the-art in ImageNet in 2018. Att-Gan (He et al. 2019b) uses attention to easily take advantage of multimodality to improve the generation of images. Combining a region of the image with a corresponding part of the word-context vector helps to generate new features with more details in each stage.

Attention has still been little explored to make generative models simpler, scalable, and more stable. Perhaps the only approach in the literature to explore such aspects more deeply is DRAW (Gregor et al. 2015), which presents a sequential and straightforward way to generate images, being possible to refine image patches while more information is captured sequentially. However, the architecture was tested only in simple datasets, leaving open spaces for new developments. There is not much exploration of attention using autoencoders. Using VAEs, Bornschein et al. (2017) increased the generative models with external memory and used an attentional system to address and retrieve the corresponding memory content.

In natural language processing, attention is explored in unsupervised models mainly to extract aspects of sentiment analysis. It is also used within autoencoders to generate semantic representations of phrases (Zhang et al. 2017a; Tian and Fang 2019). However, most studies still use supervised learning attention, and few approaches still focus on computer vision and NLP. Therefore, we believe that there is still a great path for research and exploration of attention in the unsupervised context, particularly we note that the construction of purely bottom-up attentional systems is not explored in the literature and especially in the context of unsupervised learning, these systems can great value, accompanied by inhibition and return mechanisms.

6.11 New tasks and robotics

Although attention has been used in several domains, there are still potential applications that can benefit from it. The prediction of time series, medical applications, and robotics applications are little-explored areas of the literature. Predicting time series becomes challenging as the size of the series increases. Attentional neural networks can contribute significantly to improving results. Specifically, we believe that exploring RAM (Mnih et al. 2014) with multiple glimpses looking at different parts of the series or different frequency ranges can introduce a new way of computing time series. In medical applications, there are still few works that explore biomedical signals in attentional architectures. There are opportunities to apply attention to all applications, ranging from segmentation and image classification, support for disease diagnosis to support treatments such as Parkinson's, Alzheimer's, and other chronic diseases.

For robotics, there are countless opportunities. For years the robotics community has been striving for robots to perform tasks in a safe manner and with behaviors closer to humans. However, DL techniques need to cope well with multimodality, active learning, incremental learning, identify unknowns, uncertainty estimation, object and scene semantics, reasoning, awareness, and planning for this task. Architectures like RAM (Mnih et al.

2014), DRAW (Gregor et al. 2015) and Transformer (Vaswani et al. 2017) can contribute a lot by being applied to visual odometry, SLAM and mapping tasks.

7 Conclusions

In this survey, we presented a systematic review of the literature on attention in Deep Learning to overview the area from its main approaches, historical landmarks, uses of attention, applications, and research opportunities. In total, we critically analyzed more than 600 relevant papers published from 2014 to the present. To the best of our knowledge, this is the broadest survey in the literature, given that most of the existing reviews cover only particular domains with a slightly smaller number of reviewed works. Throughout the paper, we have identified and discussed the relationship between attention mechanisms in established deep neural network models, emphasizing CNNs, RNNs, and generative models. We discussed how attention led to performance gains, improvements in computational efficiency, and a better understanding of networks' knowledge. We present an exhaustive list of application domains discussing the main benefits of attention, highlighting each domain's most representative instances. We also showed recent discussions about attention on the explanation and interpretability of models, a branch of research that is widely discussed today. Finally, we present what we consider trends and opportunities for new developments around attentive models. We hope that this survey will help the audience understand the different existing research directions and provide significant scientific community background in generating future research.

It is worth mentioning that our survey results from an extensive and exhaustive process of searching, filtering, and critical analysis of papers published between 01/01/2014 until 15/02/2021 in the central publication repositories for machine learning and related areas. In total, we collected more than 20,000 papers. After successive automatic and manual filtering, we selected approximately 650 papers for critical analysis and more than 6000 for quantitative analyses. We identify the main application domains, places of publication, and main architectures. For automatic filtering, we use keywords from the area and set up different combinations of filters to eliminate noise from psychology and classic computational visual attention techniques (i.e., saliency maps). In manual filtering, we separate the papers by year and define the work's originality and the number of citations as the main selection criteria. In the appendix, we provide our complete methodology and links to our search codes to improve future revisions on any topic in the area.

We are currently complementing this survey with a theoretical analysis of the main neural attention models. This complementary survey will help to address an urgent need for an attentional framework supported by taxonomies based on theoretical aspects of attention, which predate the era of Deep Learning. The few existing taxonomies in the area do not yet use theoretical concepts. Hence, they are challenging to extend to various architectures and application domains. Taxonomies inspired by classical concepts are essential to understand how attention has acted in deep neural networks and whether the roles played corroborate with theoretical foundations studied for more than 40 years in psychology and neuroscience. This study is already in the final stages of development by our team and will hopefully help researchers develop new attentional structures with functions still little explored in the literature. We hope to make it available to the scientific community as soon as possible.

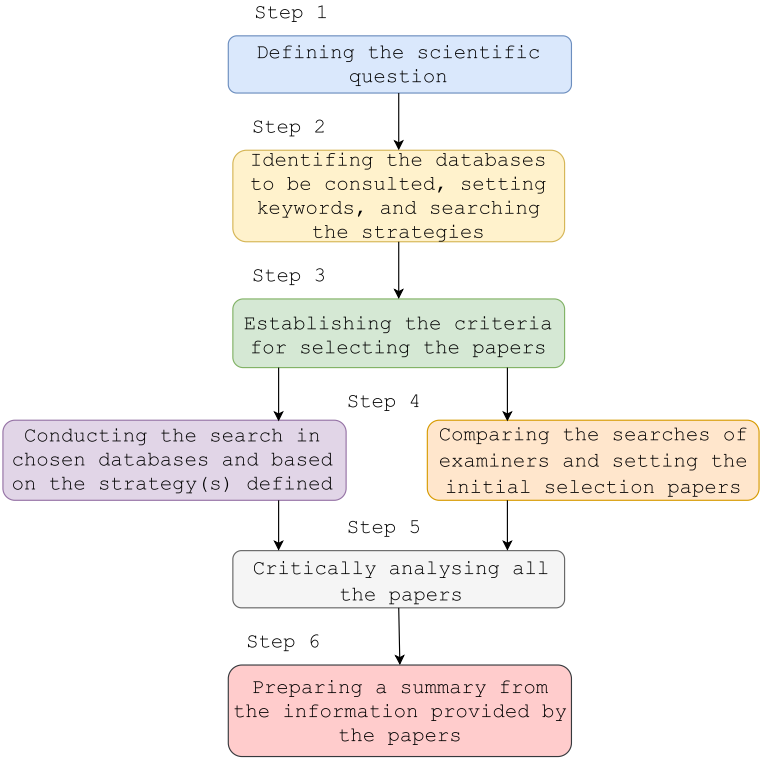

Fig. 24 Steps of the systematic review used in this survey

Appendix

This survey employs a systematic review (SR) approach aiming to collect, critically evaluate, and synthesize the results of multiple primary studies concerning attention in Deep Learning. The selection and evaluation of the works should be meticulous and easily reproducible. Also, SR should be objective, systematic, transparent, and replicable. Although recent, the use of attention in Deep Learning is extensive. Therefore, we systematically reviewed the literature, collecting works from a variety of sources. SR consists of the following steps: defining the scientific questions, identifying the databases, establishing the criteria for selecting papers, searching the databases, performing a critical analysis to choose the most relevant works, and preparing a critical summary of the most relevant papers, as shown Fig. 24.

This survey covers the following aspects: (1) The uses of attention in Deep Learning; (2) Attention mechanisms; (3) Uses of attention; (4) Attention applications; (5) Attention and interpretability; (6) Trends and challenges. These aspects provide the main topics regarding attention in Deep Learning, which can help understand the field's fundamentals. The second step identifies the main databases in the machine learning area, such as arXiv, DeepMind, Google AI, OpenAI, Facebook AI research, Microsoft research, Amazon research, Google Scholar, IEEE Xplore, DBLP, ACM, NIPS, ICML, ICLR, AAAI, CVPR, ICCV, CoRR, IJCNN, Neurocomputing, and Google general search (including blogs,

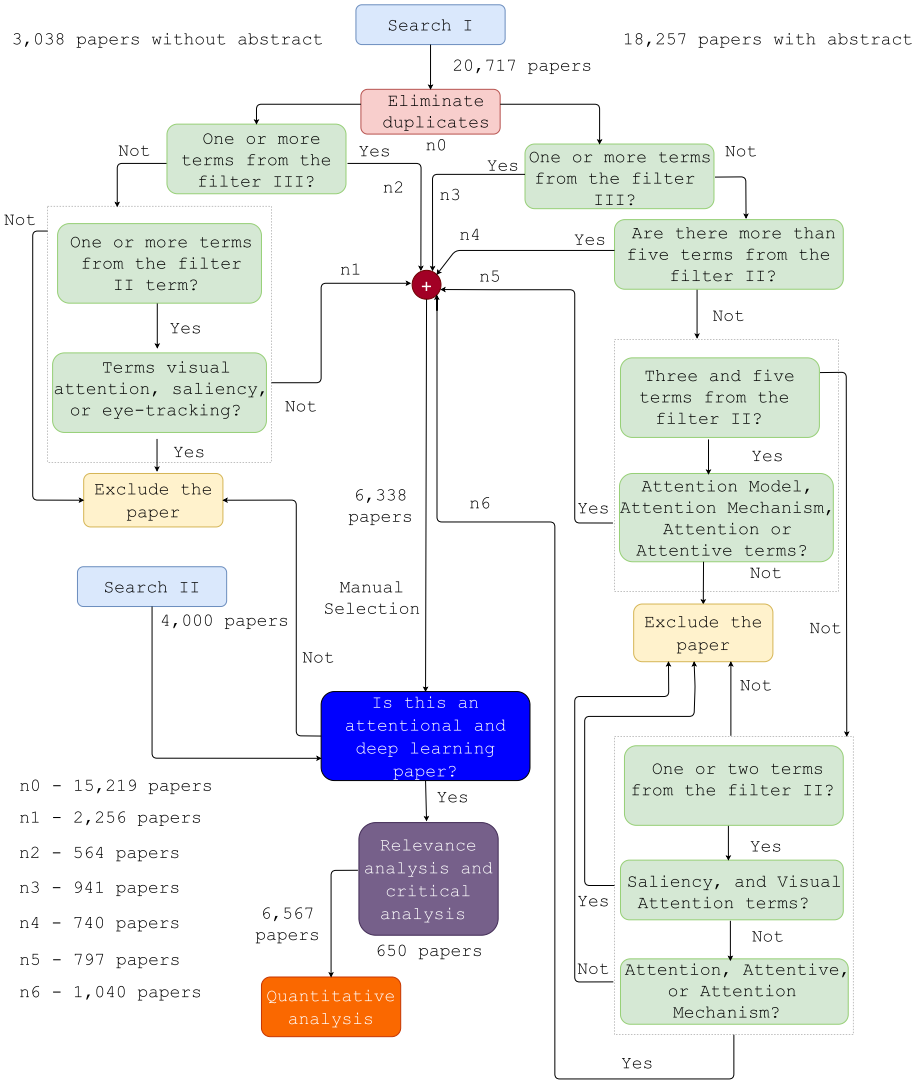

Fig. 25 The filter strategies for selecting the relevant works. Search I corresponds to articles collected between 01/01/2014 to 06/30/2019 (first stage), and Search II corresponds to papers collected between 07/01/2019 to 02/15/2021 (second stage)

distill, and Quora). Our searching period comprises 01/01/2014 to 06/30/2019 (first stage) and 07/01/2019 to 02/15/2021 (second stage), and the search was performed via a Phyton script.[7] The papers' title, abstract, year, DOI, and source publication were downloaded and stored in a JSON file. The most appropriate set of keywords to perform the searches was

[7] https://github.com/larocs/attention_dl.

defined by partially searching the field with expert knowledge from our research group. The final set of keywords wore: attention, attentional, and attentive.

However, these keywords are also relevant in psychology and visual attention. Hence, we performed a second selection to eliminate these papers and remmove duplicate papers unrelated to the DL field. After removing duplicates, 18,257 different papers remained. In the next selection step, we performed a sequential combination of three types of filters: (1) Filter I: Selecting the works with general terms of attention (i.e., attention, attentive, attentional, saliency, top-down, bottom-up, memory, focus, and mechanism); (2) Filter II: Selecting the works with terms related to DL (i.e. deep learning, neural network, ann, dnn deep neural, encoder, decoder, recurrent neural network, recurrent network, rnn, long short term memory, long short-term memory, lstm, gated recurrent unit, gru, autoencoder, ae, variational autoencoder, vae, denoising ae, dae, sparse ae, sae, markov chain, mc, hopfield network, boltzmann machine, em, restricted boltzmann machine, rbm, deep belief network, dbn, deep convolutional network, dcn, deconvolution network, dn, deep convolutional inverse graphics network, dcign, generative adversarial network, gan, liquid state machine, lsm, extreme learnng, machine, elm, echo state network, esn, deep residual network, drn, konohen network, kn, turing machine, ntm, convolutional network, cnn, and capsule network); (3) Filter III: Selecting the works with specific words of attention in Deep Learning (i.e., attention network, soft attention, hard attention, self-attention, self attention deep attention, hierarchical attention, transformer, local attention, global attention, coattention, co-attention, flow attention, attention-over-attention, way attention, intra-attention, self-attentive, and self attentive).

The decision tree with the second selection is shown in Fig. 25. The third filtering selects works with at least one specific term of attention in deep learning. In the next filtering, we remove papers without abstract, the collection of filters verify if there is at least one specific term of Deep Learning and remove the works with the following keywords: visual attention, saliency, and eye-tracking. For the papers with abstract, the selection is more complex, requiring three cascade conditions: (1) First condition: Selecting the works that have more than five filter terms from filter II; (2) Second condition: selecting the works that have between three and five terms from filter II and where there is at least one of the following: attention model, attention mechanism, attention, or attentive; (3) Third condition: Selecting the works with one or two terms from filter II; without the terms: salience, visual attention, attentive, and attentional mechanism. A total of 6338 works remained for manual selection. We manually excluded the papers without a title, abstract, or introduction related to the DL field. After manual selection, 3567 works were stored in Zotero. Given the number of papers, we grouped them by year and chose those above a threshold (average citations in the group). Only works above average were read and classified as relevant or not for critical analysis. To find the number of citations, we automated the process with a Python script. 650 papers were considered relevant for this survey's critical analysis, and 6567 were used to perform quantitative analyzes.

References

Abdulnabi AH, Shuai B, Winkler S, Wang G (2017) Episodic camn: contextual attention-based memory networks with iterative feedback for scene labeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5561–5570

Abolghasemi P, Mazaheri A, Shah M, Boloni L (2019) Pay attention!-robustifying a deep visuomotor policy through task-focused visual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4254–4262

Abu-El-Haija S, Perozzi B, Al-Rfou R, Alemi AA (2018) Watch your step: learning node embeddings via graph attention. In: Advances in neural information processing systems, pp 9180–9190

Agrawal A, Batra D, Parikh D, Kembhavi A (2018) Don't just assume; look and answer: overcoming priors for visual question answering. In: Proceedings of the IEEE CVPR, pp 4971–4980

Ahmadi S (2017) Attention-based encoder-decoder networks for spelling and grammatical error correction. PhD thesis, Paris Descartes University

Ahmadi AHK, Hassani K, Moradi P, Lee L, Morris Q (2020) Memory-based graph networks. In: 8th International conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net. https://openreview.net/forum?id=r1laNeBYPB

Allamanis M, Peng H, Sutton C (2016) A convolutional attention network for extreme summarization of source code. In: International conference on machine learning, pp 2091–2100

Alt C, Hübner M, Hennig L (2018) Improving relation extraction by pre-trained language representations. In: Automated knowledge base construction (AKBC)

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086

Andreas J, Rohrbach M, Darrell T, Klein D (2015) Deep compositional question answering with neural module networks. arXiv:abs/1511.02799

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433

Arjovsky M (2020) Out of distribution generalization in machine learning. PhD thesis, New York University

Ba J, Mnih V, Kavukcuoglu K (2015) Multiple object recognition with visual attention. In: ICLR (Poster). arXiv:1412.7755

Baevski A, Auli M (2019) Adaptive input representations for neural language modeling. In: International conference on learning representations. https://openreview.net/forum?id=ByxZX20qFQ

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7=-9, 2015, conference track proceedings

Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y (2016) End-to-end attention-based large vocabulary speech recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4945–4949

Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2020) Emergent tool use from multi-agent autocurricula. In: International conference on learning representations. https://openreview.net/forum?id=SkxpxJBKwS

Baluja S, Pomerleau DA (1997) Expectation-based selective attention for visual monitoring and control of a robot vehicle. Robot Auton Syst 22(3–4):329–344

Baraldi L, Grana C, Cucchiara R (2017) Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1657–1666

Bartunov S, Vetrov DP (2017) Fast adaptation in generative models with generative matching networks. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Workshop track proceedings, OpenReview.net. https://openreview.net/forum?id=r1IvyjVYl

Bastings J, Titov I, Aziz W, Marcheggiani D, Sima'an K (2017) Graph convolutional encoders for syntax-aware neural machine translation. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, Copenhagen, Denmark, pp 1957–1967. https://doi.org/10.18653/v1/D17-1209

Bauer L, Wang Y, Bansal M (2018) Commonsense for generative multi-hop question answering tasks. In: Proceedings of the empirical methods in natural language processing

Baziotis C, Pelekis N, Doulkeridis C (2017) Datastories at semeval-2017 task 4: deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 747–754

Bazzani L, Larochelle H, Torresani L (2017) Recurrent mixture density network for spatiotemporal visual attention. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference track proceedings, OpenReview.net. https://openreview.net/forum?id=SJRpRfKxx

Bello I, Zoph B, Vaswani A, Shlens J, Le QV (2019) Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3286–3295

Ben-Shimon D, Tsikinovsky A, Friedmann M, Shapira B, Rokach L, Hoerle J (2015) Recsys challenge 2015 and the yoochoose dataset. In: Proceedings of the 9th ACM conference on recommender systems, pp 357–358

Bhunia AK, Konwer A, Bhunia AK, Bhowmick A, Roy PP, Pal U (2019) Script identification in natural scene image and video frames using an attention based convolutional-lstm network. Pattern Recognit 85:172–184. https://doi.org/10.1016/j.patcog.2018.07.034

Bielski A, Trzcinski T (2018) Pay attention to virality: understanding popularity of social media videos with the attention mechanism. In: Proceedings of the IEEE CVPR workshops, pp 2335–2337

Bin Y, Yang Y, Shen F, Xie N, Shen HT, Li X (2018) Describing video with attention-based bidirectional lstm. IEEE Trans Cybern 49(7):2631–2641

Bornschein J, Mnih A, Zoran D, Rezende DJ (2017) Variational memory addressing in generative models. In: Proceedings of the 31st international conference on neural information processing systems, pp 3923–3932

Breazeal C, Scassellati B (1999) A context-dependent attention system for a social robot. rn 255:3

Broadbent DE (2013) Perception and communication. Elsevier, New York

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual

Burgess CP, Matthey L, Watters N, Kabra R, Higgins I, Botvinick M, Lerchner A (2019) Monet: unsupervised scene decomposition and representation. arXiv:abs/1901.11390

Cao Q, Lin L, Shi Y, Liang X, Li G (2017) Attention-aware face hallucination via deep reinforcement learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 690–698

Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. 2016 IEEE international conference on acoustics. speech and signal processing (ICASSP). IEEE, pp 4960–4964

Chaudhari S, Polatkan G, Ramanath R, Mithal V (2019) An attentive survey of attention models. arXiv:190402874

Chen J (2016) Multi-modal learning: study on a large-scale micro-video data collection. In: Proceedings of the 24th ACM international conference on Multimedia, pp 1454–1458

Chen H, Li Y (2019) Three-stream attention-aware network for rgb-d salient object detection. IEEE Trans Image Process 28(6):2825–2835

Chen H, Sun M, Tu C, Lin Y, Liu Z (2016a) Neural sentiment classification with user and product attention. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1650–1659

Chen LC, Yang Y, Wang J, Xu W, Yuille AL (2016b) Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3640–3649

Chen J, Zhang H, He X, Nie L, Liu W, Chua TS (2017a) Attentive collaborative filtering: multimedia recommendation with item-and component-level attention. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, pp 335–344

Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS (2017b) Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Chen P, Sun Z, Bing L, Yang W (2017c) Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 452–461

Chen D, Li H, Xiao T, Yi S, Wang X (2018a) Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1169–1178

Chen S, Tan X, Wang B, Hu X (2018b) Reverse attention for salient object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 234–250

Chen X, Li LJ, Fei-Fei L, Gupta A (2018c) Iterative visual reasoning beyond convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7239–7248

Chen X, Xu C, Yang X, Tao D (2018d) Attention-gan for object transfiguration in wild images. In: Proceedings of the European conference on computer vision (ECCV), pp 164–180

Chen Y, Kalantidis Y, Li J, Yan S, Feng J (2018e) A^2-nets: Double attention networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31. Curran Associates, Inc., pp 352–361. http://papers.nips.cc/paper/7318-a2-nets-double-attention-networks.pdf

Chen C, Liu Y, Kreiss S, Alahi A (2019a) Crowd-robot interaction: crowd-aware robot navigation with attention-based deep reinforcement learning. In: 2019 IEEE ICRA. IEEE, pp 6015–6022

Chen X, Zhang R, Yan P (2019b) Feature fusion encoder decoder network for automatic liver lesion segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE, pp 430–433

Chen M, Radford A, Child R, Wu J, Jun H, Dhariwal P, Luan D, Sutskever I (2020) Generative pretraining from pixels. In: Proceedings of the 37th international conference on machine learning, vol 1

Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading. In: Su J, Carreras X, Duh K (eds) Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, pp 551–561. https://doi.org/10.18653/v1/d16-1053

Cheng Z, Bai F, Xu Y, Zheng G, Pu S, Zhou S (2017a) Focusing attention: towards accurate text recognition in natural images. In: 2017 IEEE international conference on computer vision (ICCV), pp 5086–5094, https://doi.org/10.1109/ICCV.2017.543

Cheng Z, Bai F, Xu Y, Zheng G, Pu S, Zhou S (2017b) Focusing attention: towards accurate text recognition in natural images. In: Proceedings of the IEEE ICCV, pp 5076–5084

Chiang T, Huang C, Su S, Chen Y (2020) Learning multi-level information for dialogue response selection by highway recurrent transformer. Comput Speech Lang 63:101073. https://doi.org/10.1016/j.csl.2020.101073

Cho K, van Merrienboer B, Bahdanau D, Bengio Y (2014a) On the properties of neural machine translation: encoder-decoder approaches. In: Wu D, Carpuat M, Carreras X, Vecchi EM (eds) Proceedings of SSST@EMNLP 2014, Eighth workshop on syntax, semantics and structure in statistical translation, Doha, Qatar, 25 October 2014, Association for Computational Linguistics, pp 103–111. https://doi.org/10.3115/v1/W14-4012, https://www.aclweb.org/anthology/W14-4012/

Cho K, van Merrienboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014b) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, pp 1724–1734. https://doi.org/10.3115/v1/d14-1179

Cho K, Courville A, Bengio Y (2015) Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans Multimed 17(11):1875–1886. https://doi.org/10.1109/TMM.2015.2477044

Choi J, Chang HJ, Yun S, Fischer T, Demiris Y, Choi JY (2017) Attentional correlation filter network for adaptive visual tracking. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 4828–4837. https://doi.org/10.1109/CVPR.2017.513

Choi J, Seo H, Im S, Kang M (2019a) Attention routing between capsules. In: Proceedings of the IEEE/CVF ICCV workshops, pp 0–0

Choi M, Park C, Yang S, Kim Y, Choo J, Hong SR (2019b) Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–12

Chopra S, Auli M, Rush AM (2016) Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, San Diego, California, pp 93–98. https://doi.org/10.18653/v1/N16-1012, http://aclweb.org/anthology/N16-1012

Chorowski J, Bahdanau D, Cho K, Bengio Y (2014) End-to-end continuous speech recognition using attention-based recurrent NN: first results. arXiv:abs/1412.1602

Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Proceedings of the 28th international conference on neural information processing systems, Vol 1, pp 577–585

Chowdhury FRR, Wang Q, Moreno IL, Wan L, (2018) Attention-based models for text-dependent speaker verification. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5359–5363

Chu Q, Ouyang W, Li H, Wang X, Liu B, Yu N (2017a) Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In: Proceedings of the IEEE international conference on computer vision, pp 4836–4845

Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X (2017b) Multi-context attention for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1831–1840

Chun MM, Golomb JD, Turk-Browne NB (2011) A taxonomy of external and internal attention. Ann Rev Psychol 62:73–101

Clark JJ, Ferrier NJ (1988) Modal control of an attentive vision system. In: IEEE ICCV. IEEE, pp 514–523

Clark JJ, Ferrier NJ (1992) Attentive visual servoing. In: Active vision, Citeseer

Clark K, Khandelwal U, Levy O, Manning CD (2019) What does bert look at? An analysis of Bert's attention. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, pp 276–286

Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N (2018) A discourse-aware attention model for abstractive summarization of long documents. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers), Association for Computational Linguistics, pp 615–621. https://doi.org/10.18653/v1/n18-2097

Colombini EL, da Silva Simoes A, Ribeiro C (2014) An attentional model for intelligent robotics agents. PhD thesis, Instituto Tecnológico de Aeronáutica, São José dos Campos, Brazil

Cornia M, Baraldi L, Serra G, Cucchiara R (2018) Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Trans Image Process 27(10):5142–5154

Cui Y, Liu T, Chen Z, Wang S, Hu G (2016) Consensus attention-based neural networks for Chinese reading comprehension. In: Calzolari N, Matsumoto Y, Prasad R (eds) COLING 2016, 26th international conference on computational linguistics, Proceedings of the conference: technical papers, December 11–16, 2016, Osaka, Japan, ACL, pp 1777–1786. https://www.aclweb.org/anthology/C16-1167/

Cui Y, Chen Z, Wei S, Wang S, Liu T, Hu G (2017) Attention-over-attention neural networks for reading comprehension. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 593–602

Dai N, Liang J, Qiu X, Huang X (2019a) Style transformer: unpaired text style transfer without disentangled latent representation. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 5997–6007. https://doi.org/10.18653/v1/p19-1601

Dai Z, Yang Z, Yang Y, Carbonell JG, Le QV, Salakhutdinov R (2019b) Transformer-xl: attentive language models beyond a fixed-length context. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 2978–2988. https://doi.org/10.18653/v1/p19-1285

Damirchi H, Khorrambakht R, Taghirad HD (2020) Exploring self-attention for visual odometry. arXiv:abs/2011.08634

Daniluk M, Rocktäschel T, Welbl J, Riedel S (2017) Frustratingly short attention spans in neural language modeling. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=ByIAPUcee

Das R, Neelakantan A, Belanger D, McCallum A (2017) Chains of reasoning over entities, relations, and text using recurrent neural networks. In: Lapata M, Blunsom P, Koller A (eds) Proceedings of the 15th conference of the european chapter of the association for computational linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, volume 1: long papers. Association for Computational Linguistics, pp 132–141. https://doi.org/10.18653/v1/e17-1013

Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser L (2019) Universal transformers. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net. https://openreview.net/forum?id=HyzdRiR9Y7

Deng Y, Kim Y, Chiu J, Guo D, Rush A (2018) Latent alignment and variational attention. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31. Curran Associates, Inc., pp 9712–9724. http://papers.nips.cc/paper/8179-latent-alignment-and-variational-attention.pdf

Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 4171–4186. https://doi.org/10.18653/v1/n19-1423

Dhingra B, Liu H, Yang Z, Cohen WW, Salakhutdinov R (2017) Gated-attention readers for text comprehension. In: Barzilay R, Kan M (eds) Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, volume 1: long papers, association for computational linguistics. pp 1832–1846. https://doi.org/10.18653/v1/P17-1168

Dong L, Lapata M (2016) Language to logical form with neural attention. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 33–43

Dong L, Wang F, Xu B (2019) Self-attention aligner: a latency-control end-to-end model for asr using self-attention network and chunk-hopping. ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 5656–5660

Doughty H, Mayol-Cuevas W, Damen D (2019) The pros and cons: rank-aware temporal attention for skill determination in long videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7862–7871

Dozat T, Manning CD (2017) Deep biaffine attention for neural dependency parsing. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings. OpenReview.net. https://openreview.net/forum?id=Hk95PK9le

Du W, Wang Y, Qiao Y (2017) Rpan: an end-to-end recurrent pose-attention network for action recognition in videos. In: Proceedings of the IEEE international conference on computer vision, pp 3725–3734

Du J, Gui L, He Y, Xu R, Wang X (2019) Convolution-based neural attention with applications to sentiment classification. IEEE Access 7:27983–27992

Duan Y, Andrychowicz M, Stadie B, Ho OJ, Schneider J, Sutskever I, Abbeel P, Zaremba W (2017) One-shot imitation learning. In: Advances in neural information processing systems, pp 1087–1098

Edel M, Lausch J (2016) Capacity visual attention networks. In: GCAI, pp 72–80

Eriguchi A, Hashimoto K, Tsuruoka Y (2016) Tree-to-sequence attentional neural machine translation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 823–833

Escolano C, Costa-jussà MR, Fonollosa JAR (2018) (self-attentive) autoencoder-based universal language representation for machine translation. arXiv:abs/1810.06351

Eyzaguirre C, Soto A (2020) Differentiable adaptive computation time for visual reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12817–12825

Fakoor R, Mohamed A, Mitchell M, Kang SB, Kohli P (2016) Memory-augmented attention modelling for videos. arXiv:abs/1611.02261

Fan A, Lewis M, Dauphin YN (2018a) Hierarchical neural story generation. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, volume 1: long papers. Association for Computational Linguistics, pp 889–898. https://doi.org/10.18653/v1/P18-1082, https://www.aclweb.org/anthology/P18-1082/

Fan Z, Zhao X, Lin T, Su H (2018b) Attention-based multiview re-observation fusion network for skeletal action recognition. IEEE Trans Multimed 21(2):363–374

Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC, et al. (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482

Fang B, Li Y, Zhang H, Chan JCW (2019a) Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism. Remote Sens 11(2):159

Fang K, Toshev A, Fei-Fei L, Savarese S (2019b) Scene memory transformer for embodied agents in long-horizon tasks. In: Proceedings of the IEEE CVPR, pp 538–547

Feng S, Wang Y, Liu L, Wang D, Yu G (2019) Attention based hierarchical lstm network for context-aware microblog sentiment classification. World Wide Web 22(1):59–81

Ferret J, Marinier R, Geist M, Pietquin O (2020) Self-attentional credit assignment for transfer in reinforcement learning. In: Bessiere C (ed) Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020, ijcai.org, pp 2655–2661. https://doi.org/10.24963/ijcai.2020/368

Figurnov M, Collins MD, Zhu Y, Zhang L, Huang J, Vetrov D, Salakhutdinov R (2017) Spatially adaptive computation time for residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1039–1048

Firat O, Cho K, Bengio Y (2016) Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Knight K, Nenkova A, Rambow O (eds) NAACL HLT 2016, The 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, San Diego California, USA, June 12–17, 2016. The Association for Computational Linguistics, pp 866–875. https://doi.org/10.18653/v1/n16-1101

Frintrop S (2006) VOCUS: a visual attention system for object detection and goal-directed search, vol 3899. Springer, Berlin

Frintrop S, Jensfelt P (2008) Attentional landmarks and active gaze control for visual slam. IEEE Trans Robot 24(5):1054–1065

Frintrop S, Rome E, Christensen H (2010a) Computational visual attention systems and their cognitive foundations: a survey. ACM Trans Appl Percept. https://doi.org/10.1145/1658349.1658355

Frintrop S, Rome E, Christensen HI (2010b) Computational visual attention systems and their cognitive foundations: a survey. ACM Trans Appl Percept (TAP) 7(1):1–39

Fu J, Zheng H, Mei T (2017) Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Honolulu, pp 4476–4484. https://doi.org/10.1109/CVPR.2017.476, http://ieeexplore.ieee.org/document/8099959/

Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, June 16–20, 2019, Computer Vision Foundation/IEEE, pp 3146–3154. https://doi.org/10.1109/CVPR.2019.00326

Fukui H, Hirakawa T, Yamashita T, Fujiyoshi H (2019a) Attention branch network: learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10705–10714

Fukui H, Hirakawa T, Yamashita T, Fujiyoshi H (2019b) Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10705–10714

Galassi A, Lippi M, Torroni P (2020) Attention in natural language processing. In: IEEE transactions on neural networks and learning systems

Gammulle H, Denman S, Sridharan S, Fookes C (2017) Two stream lstm: A deep fusion framework for human action recognition. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 177–186

Gan Z, Cheng Y, Kholy AE, Li L, Liu J, Gao J (2019) Multi-step reasoning via recurrent dual attention for visual dialog. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, July 28– August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 6463–6474. https://doi.org/10.18653/v1/p19-1648

Ganea O, Hofmann T (2017) Deep joint entity disambiguation with local neural attention. In: Palmer M, Hwa R, Riedel S (eds) Proceedings of the 2017 conference on empirical methods in natural language processing, EMNLP 2017, Copenhagen, September 9–11, 2017. Association for Computational Linguistics, pp 2619–2629. https://doi.org/10.18653/v1/d17-1277

Gao L, Guo Z, Zhang H, Xu X, Shen HT (2017) Video captioning with attention-based lstm and semantic consistency. IEEE Trans Multimed 19(9):2045–2055

Gao F, Yu J, Shen H, Wang Y, Yang H (2020a) Attentional separation-and-aggregation network for self-supervised depth-pose learning in dynamic scenes. arXiv:abs/2011.09369

Gao J, Li P, Chen Z, Zhang J (2020b) A survey on deep learning for multimodal data fusion. Neural Comput 32(5):829–864

Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. Int J Robot Res (IJRR)

Geng X, Zhang H, Bian J, Chua TS (2015) Learning image and user features for recommendation in social networks. In: Proceedings of the IEEE international conference on computer vision, pp 4274–4282

Girdhar R, Ramanan D (2017) Attentional pooling for action recognition. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, pp 34–45

Girdhar R, Carreira J, Doersch C, Zisserman A (2019) Video action transformer network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 244–253

Goyal A, Lamb A, Hoffmann J, Sodhani S, Levine S, Bengio Y, Schölkopf B (2019) Recurrent independent mechanisms. arXiv:190910893

Graves A (2016) Adaptive computation time for recurrent neural networks. arXiv:abs/1603.08983

Graves A, Wayne G, Danihelka I (2014) Neural turing machines. arXiv:abs/1410.5401

Grefenstette E, Hermann KM, Suleyman M, Blunsom P (2015) Learning to transduce with unbounded memory. In: Advances in neural information processing systems, pp 1828–1836

Gregor K, Danihelka I, Mnih A, Blundell C, Wierstra D (2014) Deep autoregressive networks. In: International conference on machine learning, PMLR, pp 1242–1250

Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D (2015) DRAW: A recurrent neural network for image generation. In: Bach FR, Blei DM (eds) Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, 6–11 July 2015, JMLR.org, JMLR workshop and conference proceedings, vol 37, pp 1462–1471. http://proceedings.mlr.press/v37/gregor15.html

Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y (2018) Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. arXiv:abs/1801.09927

Gülçehre Ç, Denil M, Malinowski M, Razavi A, Pascanu R, Hermann KM, Battaglia PW, Bapst V, Raposo D, Santoro A, de Freitas N (2019) Hyperbolic attention networks. In: 7th International conference on learning representations, ICLR 2019, New Orleans, May 6–9, 2019, OpenReview.net. https://openreview.net/forum?id=rJxHsjRqFQ

Guo Q, Qiu X, Liu P, Shao Y, Xue X, Zhang Z (2019a) Star-transformer. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 1315–1325. https://doi.org/10.18653/v1/n19-1133

Guo X, Zhang H, Yang H, Xu L, Ye Z (2019b) A single attention-based combination of cnn and rnn for relation classification. IEEE Access 7:12467–12475

Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A (2018) Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2255–2264

Hackel T, Usvyatsov M, Galliani S, Wegner JD, Schindler K (2018) Inference, learning and attention mechanisms that exploit and preserve sparsity in convolutional networks. arXiv:abs/1801.10585

Hamker FH (2005) The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Comput Vis Image Underst 100(1–2):64–106

Hamker FH (2006) Modeling feature-based attention as an active top-down inference process. BioSystems 86(1–3):91–99

Han K, Guo J, Zhang C, Zhu M (2018) Attribute-aware attention model for fine-grained representation learning. In: Boll S, Lee KM, Luo J, Zhu W, Byun H, Chen CW, Lienhart R, Mei T (eds) 2018 ACM Multimedia conference on multimedia conference, MM 2018, Seoul, Republic of Korea, October 22–26, 2018. ACM, pp 2040–2048. https://doi.org/10.1145/3240508.3240550

Han Z, Wang X, Vong C, Liu Y, Zwicker M, Chen CLP (2019) 3dviewgraph: learning global features for 3d shapes from A graph of unordered views with attention. In: Kraus S (ed) Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao, China, August 10–16, 2019, ijcai.org, pp 758–765. https://doi.org/10.24963/ijcai.2019/107

Hao Y, Zhang Y, Liu K, He S, Liu Z, Wu H, Zhao J (2017) An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 221–231

Hao J, Wang X, Yang B, Wang L, Zhang J, Tu Z (2019) Modeling recurrence for transformer. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 1198–1207. https://doi.org/10.18653/v1/n19-1122

Harsha Vardhan LV, Jia G, Kok S (2020) Probabilistic logic graph attention networks for reasoning. Companion Proc Web Conf 2020:669–673

He P, Huang W, He T, Zhu Q, Qiao Y, Li X (2017) Single shot text detector with regional attention. In: Proceedings of the IEEE international conference on computer vision, pp 3047–3055

He A, Luo C, Tian X, Zeng W (2018a) A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4834–4843

He T, Tian Z, Huang W, Shen C, Qiao Y, Sun C (2018b) An end-to-end textspotter with explicit alignment and attention. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, June 18–22, 2018. IEEE Computer Society, pp 5020–5029. https://doi.org/10.1109/CVPR.2018.00527, http://openaccess.thecvf.com/content_cvpr_2018/html/He_An_End-to-End_TextSpotter_CVPR_2018_paper.html

He X, Yang Y, Shi B, Bai X (2019a) Vd-san: visual-densely semantic attention network for image caption generation. Neurocomputing 328:48–55. https://doi.org/10.1016/j.neucom.2018.02.106

He Z, Zuo W, Kan M, Shan S, Chen X (2019b) Attgan: facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
Hendrycks D, Liu X, Wallace E, Dziedzic A, Krishnan R, Song D (2020) Pretrained transformers improve out-of-distribution robustness. arXiv:200406100
Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Advances in neural information processing systems, pp 1693–1701
Hidasi B, Karatzoglou A, Baltrunas L, Tikk D (2015) Session-based recommendations with recurrent neural networks. arXiv:151106939
Hieber F, Domhan T, Denkowski M, Vilar D (2020) Sockeye 2: A toolkit for neural machine translation. In: Forcada ML, Martins A, Moniz H, Turchi M, Bisazza A, Moorkens J, Arenas AG, Nurminen M, Marg L, Fumega S, Martins B, Batista F, Coheur L, Escartín CP, Trancoso I (eds) Proceedings of the 22nd annual conference of the european association for machine translation, EAMT 2020, Lisboa, Portugal, November 3–5, 2020. European Association for Machine Translation, pp 457–458. https://www.aclweb.org/anthology/2020.eamt-1.50/
Hole KJ, Ahmad S (2021) A thousand brains: toward biologically constrained ai. SN Appl Sci 3(8):1–14
Hoogi A, Wilcox B, Gupta Y, Rubin DL (2019) Self-attention capsule networks for image classification. arXiv:abs/1904.12483
Hori C, Hori T, Lee T, Zhang Z, Harsham B, Hershey JR, Marks TK, Sumi K (2017a) Attention-based multimodal fusion for video description. In: IEEE international conference on computer vision, ICCV 2017, Venice, Italy, October 22–29, 2017. IEEE Computer Society, pp 4203–4212. https://doi.org/10.1109/ICCV.2017.450, http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.450
Hori T, Watanabe S, Zhang Y, Chan W (2017b) Advances in joint ctc-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. In: Lacerda F (ed) Interspeech 2017, 18th annual conference of the international speech communication association. Stockholm, Sweden, August 20–24, 2017, ISCA, pp 949–953. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1296.html
Hossain M, Hosseinzadeh M, Chanda O, Wang Y (2019) Crowd counting using scale-aware attention networks. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1280–1288
Hu D (2019) An introductory survey on attention mechanisms in nlp problems. In: Proceedings of SAI intelligent systems conference. Springer, pp 432–448
Hu X, Zhu L, Fu CW, Qin J, Heng PA (2018) Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7454–7462
Hu D, Zhou S, Shen Q, Zheng S, Zhao Z, Fan Y (2019a) Digital image steganalysis based on visual attention and deep reinforcement learning. IEEE Access 7:25924–25935
Hu H, Xiao A, Zhang S, Li Y, Shi X, Jiang T, Zhang L, Zhang L, Zeng J (2019b) Deephint: understanding hiv-1 integration via deep learning with attention. Bioinformatics 35(10):1660–1667
Hu X, Yang K, Fei L, Wang K (2019c) Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 1440–1444
Hu J, Shen L, Albanie S, Sun G, Wu E (2020a) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
Hu X, Fu CW, Zhu L, Wang T, Heng PA (2020b) Sac-net: Spatial attenuation context for salient object detection. IEEE Trans Circuits Syst Video Technol
Huang W, Zhou F (2020) Da-capsnet: dual attention mechanism capsule network. Sci Rep 10(1):1–13
Huang CZA, Vaswani A, Uszkoreit J, Simon I, Hawthorne C, Shazeer N, Dai AM, Hoffman MD, Dinculescu M, Eck D (2018a) Music transformer: generating music with long-term structure. In: International conference on learning representations
Huang F, Zhang X, Zhao Z, Li Z (2018b) Bi-directional spatial-semantic attention networks for image-text matching. IEEE Trans Image Process 28(4):2008–2020
Huang H, Zhu C, Shen Y, Chen W (2018c) Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In: 6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=BJIgi_eCZ
Huang J, Zhou W, Zhang Q, Li H, Li W (2018d) Video-based sign language recognition without temporal segmentation. In: Thirty-second AAAI conference on artificial intelligence

Huang KY, Wu CH, Su MH (2019) Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. Pattern Recognit 88:668–678

Hudson DA, Manning CD (2018) Compositional attention networks for machine reasoning. In: 6th International conference on learning representations, ICLR 2018, Vancouver. April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=S1Euwz-Rb

Hudson DA, Manning CD (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6700–6709

Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, PMLR, proceedings of machine learning research, vol 80, pp 2132–2141. http://proceedings.mlr.press/v80/ilse18a.html

Irie K, Zeyer A, Schlüter R, Ney H (2019) Language modeling with deep transformers. In: Kubin G, Kacic Z (eds) Interspeech 2019, 20th annual conference of the international speech communication association, Graz, Austria, 15–19 September 2019, ISCA, pp 3905–3909. https://doi.org/10.21437/Interspeech.2019-2225

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE PAMI 20(11):1254–1259

Jaderberg M, Simonyan K, Zisserman A, kavukcuoglu k (2015) Spatial transformer networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28. Curran Associates, Inc., pp 2017–2025. http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

Jain S, Wallace BC (2019) Attention is not explanation. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 3543–3556. https://doi.org/10.18653/v1/n19-1357

James W (1890) The principles of psychology. Dover Publications, New York

Jetley S, Lord NA, Lee N, Torr PHS (2018) Learn to pay attention. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=HyzbhfWRW

Ji G, Liu K, He S, Zhao J (2017) Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the AAAI conference on artificial intelligence

Ji Z, Fu Y, Guo J, Pang Y, Zhang ZM, et al. (2018) Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Advances in neural information processing systems, pp 5995–6004

Ji Z, Xiong K, Pang Y, Li X (2019) Video summarization with attention-based encoder-decoder networks. IEEE Trans Circuits Syst Video Technol 30(6):1709–1717

Jiang H, Shi T, Bai Z, Huang L (2019) Ahcnet: an application of attention mechanism and hybrid connection for liver tumor segmentation in ct volumes. IEEE Access 7:24898–24909

Jiang M, Chen S, Yang J, Zhao Q (2020) Fantastic answers and where to find them: Immersive question-directed visual attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2980–2989

Jin S, Yao H, Sun X, Zhou S, Zhang L, Hua X (2020) Deep saliency hashing for fine-grained retrieval. IEEE Trans Image Process 29:5336–5351. https://doi.org/10.1109/TIP.2020.2971105

Johansen-Berg H, Lloyd DM (2000) The physiology and psychology of selective attention to touch. Front Biosci 5:D894–D904

Johnston A, Carneiro G (2020) Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4756–4765

Joulin A, Mikolov T (2015) Inferring algorithmic patterns with stack-augmented recurrent nets. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, December 7–12, 2015, Montreal, pp 190–198

Kadlec R, Schmid M, Bajgar O, Kleindienst J (2016) Text understanding with the attention sum reader network. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 908–918

Kahneman D (1973) Attention and effort, vol 1063. Citeseer

Kaiser L, Gomez AN, Shazeer N, Vaswani A, Parmar N, Jones L, Uszkoreit J (2017) One model to learn them all. arXiv:abs/1706.05137

Kang G, Zheng L, Yan Y, Yang Y (2018) Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision-ECCV 2018—15th european conference, Munich, September 8–14, 2018, Proceedings, Part XI, Springer, Lecture Notes in Computer Science, vol 11215, pp 420–436. https://doi.org/10.1007/978-3-030-01252-6_25

Kastaniotis D, Ntinou I, Tsourounis D, Economou G, Fotopoulos S (2018) Attention-aware generative adversarial networks (ata-gans). In: 2018 IEEE 13th image, video, and multidimensional signal processing workshop (IVMSP). IEEE, pp 1–5

Ke NR, Goyal A, Bilaniuk O, Binas J, Mozer MC, Pal C, Bengio Y (2018) Sparse attentive backtracking: temporal credit assignment through reminding. In: Proceedings of the 32nd international conference on neural information processing systems, pp 7651–7662

Kim JH, Lee SW, Kwak D, Heo MO, Kim J, Ha JW, Zhang BT (2016) Multimodal residual learning for visual qa. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29. Curran Associates, Inc., pp 361–369. http://papers.nips.cc/paper/6446-multimodal-residual-learning-for-visual-qa.pdf

Kim S, Hori T, Watanabe S (2017a) Joint ctc-attention based end-to-end speech recognition using multitask learning. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4835–4839

Kim Y, Denton C, Hoang L, Rush AM (2017b) Structured attention networks. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=HkE0Nvqlg

Kim JH, Jun J, Zhang BT (2018a) Bilinear attention networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31. Curran Associates, Inc., pp 1564–1574. http://papers.nips.cc/paper/7429-bilinear-attention-networks.pdf

Kim W, Goyal B, Chawla K, Lee J, Kwon K (2018b) Attention-based ensemble for deep metric learning. In: Proceedings of the European conference on computer vision (ECCV), pp 736–751

Kim Y, Kim D, Kumar A, Sarikaya R (2018c) Efficient large-scale neural domain classification with personalized attention. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, July 15–20, 2018, volume 1: long papers. Association for Computational Linguistics, pp 2214–2224. https://doi.org/10.18653/v1/P18-1206, https://www.aclweb.org/anthology/P18-1206/

Kim S, Kang I, Kwak N (2019) Semantic sentence matching with densely-connected recurrent and co-attentive information. In: Proceedings of the AAAI conference on artificial intelligence, pp 6586–6593

Kim ES, Kang WY, On KW, Heo YJ, Zhang BT (2020a) Hypergraph attention networks for multimodal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14581–14590

Kim J, Ma M, Pham T, Kim K, Yoo CD (2020b) Modality shifting attention network for multi-modal video question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10106–10115

Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International conference on machine learning, PMLR, pp 595–603

Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer, pp 115–141

Kong S, Fowlkes C (2019) Pixel-wise attentional gating for scene parsing. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1024–1033

Kong Q, Xu Y, Wang W, Plumbley MD (2018a) Audio set classification with attention model: a probabilistic perspective. In: 2018 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 316–320

Kong T, Sun F, Tan C, Liu H, Huang W (2018b) Deep feature pyramid reconfiguration for object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 169–185

Koren Y (2008) Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 426–434

Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J (2017) Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision, pp 706–715

Krizhevsky A, Hinton G, et al. (2009) Learning multiple layers of features from tiny images

Kuen J, Wang Z, Wang G (2016) Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3668–3677

Kumar A, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Zhong V, Paulus R, Socher R (2016) Ask me anything: Dynamic memory networks for natural language processing. In: International conference on machine learning, PMLR, pp 1378–1387

Kumar A, Sangwan SR, Arora A, Nayyar A, Abdel-Basset M et al (2019) Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access 7:23319–23328

Kuncoro A, Ballesteros M, Kong L, Dyer C, Neubig G, Smith NA (2017) What do recurrent neural network grammars learn about syntax? In: Lapata M, Blunsom P, Koller A (eds) Proceedings of the 15th conference of the european chapter of the association for computational linguistics, EACL 2017, Valencia, April 3–7, 2017, volume 1: long papers. Association for Computational Linguistics, pp 1249–1258. https://doi.org/10.18653/v1/e17-1117,

Kuo XY, Liu C, Lin KC, Lee CY (2020) Dynamic attention-based visual odometry. In: Proceedings of the IEEE/CVF CVPR workshops, pp 36–37

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee CY, Osindero S (2016) Recursive recurrent nets with attention modeling for ocr in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2231–2239

Lee JB, Rossi R, Kong X (2018a) Graph classification using structural attention. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1666–1674

Lee KH, Chen X, Hua G, Hu H, He X (2018b) Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 201–216

Lee J, Lee Y, Kim J, Kosiorek AR, Choi S, Teh YW (2019a) Set transformer: A framework for attention-based permutation-invariant neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, PMLR, Proceedings of machine learning research, vol 97, pp 3744–3753. http://proceedings.mlr.press/v97/lee19d.html

Lee JB, Rossi RA, Kim S, Ahmed NK, Koh E (2019b) Attention models in graphs: a survey. ACM Trans Knowl Discov Data 13(6):62:1–62:25. https://doi.org/10.1145/3363574

Li X, Loy CC (2018) Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proceedings of the European conference on computer vision (ECCV), pp 90–105

Li C, Zhu J, Zhang B (2016a) Learning to generate with memory. In: International conference on machine learning, pp 1177–1186

Li J, Monroe W, Jurafsky D (2016b) Understanding neural networks through representation erasure. arXiv: abs/1612.08220

Li J, Ren P, Chen Z, Ren Z, Lian T, Ma J (2017) Neural attentive session-based recommendation. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1419–1428

Li D, Yao T, Duan LY, Mei T, Rui Y (2018a) Unified spatio-temporal attention networks for action recognition in videos. IEEE Trans Multimed 21(2):416–428

Li G, Gan Y, Wu H, Xiao N, Lin L (2018b) Cross-modal attentional context learning for rgb-d object detection. IEEE Trans Image Process 28(4):1591–1601

Li H, Xiong P, An J, Wang L (2018c) Pyramid attention network for semantic segmentation. In: British machine vision conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, BMVA Press, p 285. http://bmvc2018.org/contents/papers/1120.pdf

Li K, Wu Z, Peng K, Ernst J, Fu Y (2018d) Tell me where to look: Guided attention inference network. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, June 18–22, 2018. IEEE Computer Society, pp 9215–9223. https://doi.org/10.1109/CVPR.2018.00960

Li S, Bak S, Carr P, Wang X (2018e) Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 369–378

Li W, Zhu X, Gong S (2018f) Harmonious attention network for person re-identification. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Li Y, Zeng J, Shan S, Chen X (2018g) Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans Image Process 28(5):2439–2450

Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG (2018h) Videolstm convolves, attends and flows for action recognition. Comput Vis Image Underst 166:41–50

Li H, Chen J, Hu R, Yu M, Chen H, Xu Z (2019a) Action recognition using visual attention with reinforcement learning. In: Kompatsiaris I, Huet B, Mezaris V, Gurrin C, Cheng WH, Vrochidis S (eds) MultiMedia modeling. Lecture Notes in Computer Science. Springer, pp 365–376

Li H, Wang P, Shen C, Zhang G (2019b) Show, attend and read: A simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 8610–8617

Li J, Yang B, Dou ZY, Wang X, Lyu MR, Tu Z (2019c) Information aggregation for multi-head attention with routing-by-agreement. In: NAACL-HLT (1), pp 3566–3575. https://aclweb.org/anthology/papers/N/N19/N19-1359/

Li J, Yang J, Hertzmann A, Zhang J, Xu T (2019d) Layoutgan: Generating graphic layouts with wireframe discriminators. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net. https://openreview.net/forum?id=HJxB5sRcFQ

Li L, Gan Z, Cheng Y, Liu J (2019e) Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 10313–10322

Li N, Liu S, Liu Y, Zhao S, Liu M (2019f) Neural speech synthesis with transformer network. In: Proceedings of the AAAI conference on artificial intelligence, pp 6706–6713

Li R, Li M, Li J (2019g) Connection sensitive attention U-NET for accurate retinal vessel segmentation. arXiv:abs/1903.05558

Li X, Chebiyyam V, Kirchhoff K (2019h) Multi-stream network with temporal attention for environmental sound classification. In: Kubin G, Kacic Z (eds) Interspeech 2019, 20th Annual conference of the international speech communication Association, Graz, Austria, 15–19 September 2019, ISCA, pp 3604–3608. https://doi.org/10.21437/Interspeech.2019-3019

Li X, Xiong H, Wang H, Rao Y, Liu L, Huan J (2019i) Delta: deep learning transfer using feature map with attention for convolutional networks. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019, OpenReview.net. https://openreview.net/forum?id=rkgbwsAcYm

Li X, Zhou Z, Chen L, Gao L (2019j) Residual attention-based lstm for video captioning. World Wide Web 22(2):621–636

Li Y, Yao T, Pan Y, Chao H, Mei T (2019k) Pointing novel objects in image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12497–12506

Li X, Hou Y, Wang P, Gao Z, Xu M, Li W (2021) Transformer guided geometry model for flow-based unsupervised visual odometry. Neural Comput Appl 1–12

Liang X, Hu Z, Zhang H, Gan C, Xing EP (2017) Recurrent topic-transition gan for visual paragraph generation. In: Proceedings of the IEEE ICCV, pp 3362–3371

Liang J, Jiang L, Cao L, Li J, Haupmann A (2018a) Focal visual-text attention for visual question answering. In: CVPR

Liang Y, Ke S, Zhang J, Yi X, Zheng Y (2018b) Geoman: multi-level attention networks for geo-sensory time series prediction. In: IJCAI, pp 3428–3434

Liang J, Jiang L, Cao L, Kalantidis Y, Li LJ, Hauptmann AG (2019) Focal visual-text attention for memex question answering. IEEE PAMI 41(8):1893–1908

Liao X, He L, Yang Z, Zhang C (2018) Video-based person re-identification via 3d convolutional networks and non-local attention. In: Asian conference on computer vision. Springer, pp 620–634

Libovický J, Helcl J, Marecek D (2018) Input combination strategies for multi-source transformer decoder. In: Bojar O, Chatterjee R, Federmann C, Fishel M, Graham Y, Haddow B, Huck M, Jimeno-Yepes A, Koehn P, Monz C, Negri M, Névéol A, Neves ML, Post M, Specia L, Turchi M, Verspoor K (eds) Proceedings of the third conference on machine translation: research papers. WMT 2018, Belgium, Brussels, October 31–November 1, 2018. Association for Computational Linguistics, pp 253–260. https://doi.org/10.18653/v1/w18-6326,

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp 740–755

Lin Y, Shen S, Liu Z, Luan H, Sun M (2016) Neural relation extraction with selective attention over instances. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 2124–2133

Lin Z, Feng M, dos Santos CN, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference track proceedings. OpenReview.net, https://openreview.net/forum?id=BJC_jUqxe

Liu B, Lane I (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. In: Morgan N (ed) Interspeech 2016, 17th annual conference of the international speech communication association, San Francisco, September 8–12, 2016, ISCA, pp 685–689. https://doi.org/10.21437/Interspeech.2016-1352,

Liu J, Zhang Y (2017) Attention modeling for targeted sentiment. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers, pp 572–577

Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. Neurocomputing 337:325–338

Liu Y, Lapata M (2019) Hierarchical transformers for multi-document summarization. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 5070–5081. https://doi.org/10.18653/v1/p19-1500

Liu Y, Sun C, Lin L, Wang X (2016) Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv:abs/1605.09090

Liu H, Feng J, Qi M, Jiang J, Yan S (2017a) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506

Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC (2017b) Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Trans Image Process 27(4):1586–1599

Liu J, Wang G, Hu P, Duan LY, Kot AC (2017c) Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1647–1656

Liu J, Wang G, Hu P, Duan LY, Kot AC (2017d) Global context-aware attention lstm networks for 3d action recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017e) Hydraplus-net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp 350–359

Liu J, Gao C, Meng D, Hauptmann AG (2018a) Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5197–5206

Liu L, Wang H, Li G, Ouyang W, Lin L (2018b) Crowd counting using deep recurrent spatial-aware network. In: Lang J (ed) Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI 2018, July 13–19, 2018, Stockholm, Sweden, ijcai.org, pp 849–855. https://doi.org/10.24963/ijcai.2018/118

Liu N, Han J, Yang MH (2018c) Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3089–3098

Liu Q, Zeng Y, Mokhosi R, Zhang H (2018d) Stamp: short-term attention/memory priority model for session-based recommendation. In: Proceedings of the 24th ACM SIGKDD international conference on on knowledge discovery and data mining, pp 1831–1839

Liu N, Long Y, Zou C, Niu Q, Pan L, Wu H (2019a) Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3225–3234

Liu S, Johns E, Davison AJ (2019b) End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1871–1880

Liu S, Johns E, Davison AJ (2019c) End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1871–1880

Liu S, Zhang S, Zhang X, Wang H (2019d) R-trans: Rnn transformer network for Chinese machine reading comprehension. IEEE Access 7:27736–27745

Liu X, Wang Z, Shao J, Wang X, Li H (2019e) Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1950–1959

Long X, Gan C, De Melo G, Wu J, Liu X, Wen S (2018) Attention clusters: Purely attention based local feature integration for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7834–7843

Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29. Curran Associates, Inc., pp 289–297

Lu J, Kannan A, Yang J, Parikh D, Batra D (2017a) Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in Neural Information Processing Systems 30: Annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 314–324

Lu J, Xiong C, Parikh D, Socher R (2017b) Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 375–383

Lu N, Wu Y, Feng L, Song J (2018) Deep learning for fall detection: three-dimensional cnn combined with lstm on video kinematic data. IEEE J Biomed Health Inf 23(1):314–323

Lu X, Wang W, Danelljan M, Zhou T, Shen J, Gool LV (2020) Video object segmentation with episodic graph memory networks. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer vision-ECCV 2020—16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III, Springer, Lecture Notes in Computer Science, vol 12348, pp 661–679. https://doi.org/10.1007/978-3-030-58580-8_39

Luo C, Jin L, Sun Z (2019) Moran: a multi-object rectified attention network for scene text recognition. Pattern Recogn 90:109–118

Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp 1412–1421. https://doi.org/10.18653/v1/D15-1166, http://aclweb.org/anthology/D15-1166

Lüscher C, Beck E, Irie K, Kitza M, Michel W, Zeyer A, Schlüter R, Ney H (2019) Rwth asr systems for librispeech: hybrid vs attention. Proc Interspeech 2019:231–235

Ma D, Li S, Zhang X, Wang H (2017) Interactive attention networks for aspect-level sentiment classification. In: Sierra C (ed) Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, Melbourne, August 19–25, 2017. ijcai.org, pp 4068–4074. https://doi.org/10.24963/ijcai.2017/568

Ma C, Kadav A, Melvin I, Kira Z, AlRegib G, Graf HP (2018a) Attend and interact: higher-order object interactions for video understanding. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, June 18–22, 2018. IEEE Computer Society, pp 6790–6800. https://doi.org/10.1109/CVPR.2018.00710

Ma S, Fu J, Chen CW, Mei T (2018b) Da-gan: Instance-level image translation by deep attention generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5657–5666

Ma Y, Peng H, Cambria E (2018c) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In: Proceedings of the AAAI conference on artificial intelligence

Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) Dialoguernn: an attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence, pp 6818–6825

Matthews D, Behne T, Lieven E, Tomasello M (2012) Origins of the human pointing gesture: a training study. Dev Sci 15(6):817–829

Mei X, Pan E, Ma Y, Dai X, Huang J, Fan F, Du Q, Zheng H, Ma J (2019) Spectral-spatial attention networks for hyperspectral image classification. Remote Sens 11(8):963

Mejjati YA, Richardt C, Tompkin J, Cosker D, Kim KI (2018) Unsupervised attention-guided image-to-image translation. In: Advances in neural information processing systems, pp 3693–3703

Meng X, Deng X, Zhu S, Liu S, Wang C, Chen C, Zeng B (2018) Mganet: a robust model for quality enhancement of compressed video. arXiv:181109150

Mensch A, Blondel M (2018) Differentiable dynamic programming for structured prediction and attention. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, PMLR, proceedings of machine learning research, vol 80, pp 3459–3468. http://proceedings.mlr.press/v80/mensch18a.html

Miller AH, Fisch A, Dodge J, Karimi A, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. In: Su J, Carreras X, Duh K (eds) Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016. The Association for Computational Linguistics, pp 1400–1409. https://doi.org/10.18653/v1/d16-1147

Minaee S, Abdolrashidi A (2019) Deep-emotion: facial expression recognition using attentional convolutional network. arXiv:190201019

Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 2227–2231

Mishra N, Rohaninejad M, Chen X, Abbeel P (2018) A simple neural attentive meta-learner. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3,

2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=B1DmUzWAW
Mittal S, Lamb A, Goyal A, Voleti V, Shanahan M, Lajoie G, Mozer M, Bengio Y (2020) Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. In: International conference on machine learning, PMLR, pp 6972–6986
Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, December 8–13 2014, Montreal, pp 2204–2212
Moon S, Shah P, Kumar A, Subba R (2019) Memory graph networks for explainable memory-grounded question answering. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL), pp 728–736
Munkhdalai T, Yu H (2017) Neural tree indexers for text understanding. In: Proceedings of the conference. Association for Computational Linguistics. Meeting, NIH Public Access, vol 1, p 11
Nallapati R, Zhou B, dos Santos C, Gulçehre Ç, Xiang B (2016) Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Proceedings of The 20th SIGNLL conference on computational natural language learning, pp 280–290
Nam H, Ha JW, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299–307
Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimizing detection speed. In: 2006 IEEE CVPR), vol 2. IEEE, pp 2049–2056
Neelakantan A, Le QV, Sutskever I (2016) Neural programmer: inducing latent programs with gradient descent. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings. arXiv:1511.04834
Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY (2011) Reading digits in natural images with unsupervised feature learning
Neumann M, Vu NT (2017) Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In: Lacerda F (ed) Interspeech 2017, 18th annual conference of the international speech communication association, Stockholm, August 20–24, 2017, ISCA, pp 1263–1267. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0917.html
Nguyen DK, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Nguyen PX, Joty S (2018) Phrase-based attentions. arXiv:181003444
Norman DA (1968) Toward a theory of memory and attention. Psychol Rev 75(6):522
Norouzian A, Mazoure B, Connolly D, Willett D (2019) Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed. In: ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 7310–7314
Oh J, Chockalingam V, Lee H, et al. (2016) Control of memory, active perception, and action in minecraft. In: International conference on machine learning, PMLR, pp 2790–2799
Oh SW, Lee JY, Xu N, Kim SJ (2019) Video object segmentation using space-time memory networks. In: Proceedings of the IEEE international conference on computer vision, pp 9226–9235
Okabe K, Koshinaka T, Shinoda K (2018) Attentive statistics pooling for deep speaker embedding. In: Yegnanarayana B (ed) Interspeech 2018, 19th Annual conference of the international speech communication association, Hyderabad, 2–6 September 2018, ISCA, pp 2252–2256. https://doi.org/10.21437/Interspeech.2018-993
Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, et al. (2018) Attention u-net: learning where to look for the pancreas. arXiv:180403999
Olivastri S, Singh G, Cuzzolin F (2019) An end-to-end baseline for video captioning. arXiv:190402628
Osman A, Samek W (2019) DRAU: dual recurrent attention units for visual question answering. Comput Vis Image Underst 185:24–30. https://doi.org/10.1016/j.cviu.2019.05.001
Ouerhani N (2003) Visual attention: from bio-inspired modeling to real-time implementation. PhD thesis, Université de Neuchâtel
Ouyang D, Zhang Y, Shao J (2019) Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. Pattern Recognit Lett 117:153–160
Pan Y, Yao T, Li Y, Mei T (2020) X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10971–10980

Parikh AP, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference. In: Su J, Carreras X, Duh K (eds) Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016. The Association for Computational Linguistics, pp 2249–2255, https://doi.org/10.18653/v1/d16-1244

Parisotto E, Salakhutdinov R (2018) Neural map: structured memory for deep reinforcement learning. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=Bk9zbyZCZ

Park D, Kim J, Chun SY (2019) Down-scaling with learned kernels in multi-scale deep neural networks for non-uniform single image deblurring. arXiv:190310157

Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, Tran D (2018) Image transformer. In: International conference on machine learning. PMLR, pp 4055–4064

Paulus R, Xiong C, Socher R (2018) A deep reinforced model for abstractive summarization. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=HkAClQgA-

Pedersoli M, Lucas T, Schmid C, Verbeek J (2017) Areas of attention for image captioning. In: IEEE International conference on computer vision, ICCV 2017, Venice, Italy, October 22–29, 2017. IEEE Computer Society, pp 1251–1259. https://doi.org/10.1109/ICCV.2017.140

Pei W, Baltrusaitis T, Tax DM, Morency LP (2017) Temporal attention-gated model for robust sequence classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6730–6739

Peng Y, He X, Zhao J (2017) Object-part attention model for fine-grained image classification. IEEE Trans Image Process 27(3):1487–1500

Perera D, Zimmermann R (2018) LSTM networks for online cross-network recommendations. In: Lang J (ed) Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI 2018, July 13–19, 2018, Stockholm. ijcai.org, pp 3825–3833. https://doi.org/10.24963/ijcai.2018/532

Pesce E, Ypsilantis P, Withey S, Bakewell R, Goh V, Montana G (2017) Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv:171200996

Phaf RH, Van der Heijden A, Hudson PT (1990) Slam: a connectionist model for attention in visual selection tasks. Cognit Psychol 22(3):273–341

Poulos J, Valle R (2021) Character-based handwritten text transcription with attention networks. Neural Comput Appl 1–11

Prabhavalkar R, Sainath T, Wu Y, Nguyen P, Chen Z, Chiu CC, Kannan A (2018) Minimum word error rate training for attention-based sequence-to-sequence models. https://ai.google/research/pubs/pub46670

Pu Y, Min MR, Gan Z, Carin L (2018) Adaptive feature abstraction for translating video to text. In: Thirty-second AAAI Conference on artificial intelligence

Qian R, Tan RT, Yang W, Su J, Liu J (2018) Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2482–2491

Qin Y, Song D, Chen H, Cheng W, Jiang G, Cottrell GW (2017) A dual-stage attention-based recurrent neural network for time series prediction. In: Sierra C (ed) Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, Melbourne, August 19–25, 2017, ijcai.org, pp 2627–2633. https://doi.org/10.24963/ijcai.2017/366

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Raffel C, Eck D, Liu P, Weiss RJ, Luong T (2017) Online and linear-time attention by enforcing monotonic alignments. https://ai.google/research/pubs/pub46110

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. arXiv:160605250

Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34(6):96–108

Ramesh A, Pavlov M, Goh G, Gray S (2021) Dall·e: creating images from text

Ran Q, Li P, Hu W, Zhou J (2019) Option comparison network for multiple-choice reading comprehension. arXiv:190303033

Rao Y, Lu J, Zhou J (2017) Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE international conference on computer vision, pp 3931–3940

Reed SE, Chen Y, Paine T, van den Oord A, Eslami SMA, Rezende DJ, Vinyals O, de Freitas N (2018) Few-shot autoregressive density estimation: towards learning to learn distributions. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=r1wEFyWCW

Rekabdar B, Mousas C, Gupta B (2019) Generative adversarial network with policy gradient for text summarization. In: 2019 IEEE 13th international conference on semantic computing (ICSC). IEEE, pp 204–207

Ren M, Zemel RS (2017) End-to-end instance segmentation with recurrent attention. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Ren M, Liao R, Fetaya E, Zemel R (2019) Incremental few-shot learning with attention attractor networks. In: Advances in neural information processing systems, pp 5275–5285

Rensink RA (2000) The dynamic representation of scenes. Visual Cognit 7(1–3):17–42

Rezende DJ, Mohamed S, Danihelka I, Gregor K, Wierstra D (2016) One-shot generalization in deep generative models. arXiv:160305106

Riedl MO (2019) Human-centered artificial intelligence and machine learning. Human Behav Emerg Technol 1(1):33–36

Robicquet A, Sadeghian A, Alahi A, Savarese S (2016) Learning social etiquette: Human trajectory understanding in crowded scenes. In: European conference on computer vision. Springer, pp 549–565

Rocktäschel T, Grefenstette E, Hermann KM, Kociský T, Blunsom P (2016) Reasoning about entailment with neural attention. In: Bengio Y, LeCun Y (eds) 4th International conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings. http://arxiv.org/abs/1509.06664

Rodríguez P, Cucurull G, Gonzàlez J, Gonfaus JM, Roca X (2018) A painless attention mechanism for convolutional neural networks. ICLR 2018

Rohrbach A, Rohrbach M, Hu R, Darrell T, Schiele B (2016) Grounding of textual phrases in images by reconstruction. In: European conference on computer vision. Springer, pp 817–834

Rossi E, Chamberlain B, Frasca F, Eynard D, Monti F, Bronstein M (2020) Temporal graph networks for deep learning on dynamic graphs. arXiv:200610637

Rotenstein A, Andreopoulos A, Fazl E, Jacob D, Robinson M, Shubina K, Zhu Y, Tsotsos J (2007) Towards the dream of intelligent, visually-guided wheelchairs. In: Proceedings of the 2nd international conference on technology and aging

Rudin C (2018) Please stop explaining black box models for high stakes decisions. arXiv:181110154 1

Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Màrquez L, Callison-Burch C, Su J, Pighin D, Marton Y (eds) Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015. The Association for Computational Linguistics, pp 379–389. https://doi.org/10.18653/v1/d15-1044

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, pp 3856–3866

Sadeghian A, Legros F, Voisin M, Vesel R, Alahi A, Savarese S (2018) Car-net: clairvoyant attentive recurrent network. In: Proceedings of the European conference on computer vision (ECCV), pp 151–167

Sadeghian A, Kosaraju V, Sadeghian A, Hirose N, Rezatofighi H, Savarese S (2019) Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1349–1358

Salah AA, Alpaydin E, Akarun L (2002) A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. IEEE PAMI 24(3):420–425

Salakhutdinov R, Hinton G (2009) Deep Boltzmann machines. In: Artificial intelligence and statistics. PMLR, pp 448–455

Salazar J, Kirchhoff K, Huang Z (2019) Self-attention networks for connectionist temporal classification in speech recognition. ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 7115–7119

Santoro A, Faulkner R, Raposo D, Rae J, Chrzanowski M, Weber T, Wierstra D, Vinyals O, Pascanu R, Lillicrap T (2018) Relational recurrent neural networks. In: Proceedings of the 32nd international conference on neural information processing systems, pp 7310–7321

Santos Cd, Tan M, Xiang B, Zhou B (2016) Attentive pooling networks. arXiv:160203609

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, pp 285–295

Savchuk V, Shults V, Galeev F (2018) Question answering with squad v2. 0

Schaul T, Glasmachers T, Schmidhuber J (2011) High dimensions and heavy tails for natural evolution strategies. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation, pp 845–852

Scheier C, Egner S (1997) Visual attention in a mobile robot. In: ISIE'97 Proceeding of the IEEE international symposium on industrial electronics. vol 1. IEEE, pp SS48–SS52

Schick T, Schütze H (2019) Attentive mimicking: better word embeddings by attending to informative contexts. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 489–494. https://doi.org/10.18653/v1/n19-1048

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: learning to leverage salient regions in medical images. Medical Image Anal 53:197–207

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1073–1083

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 1715–1725. https://doi.org/10.18653/v1/P16-1162, https://www.aclweb.org/anthology/P16-1162

Seo PH, Lin Z, Cohen S, Shen X, Han B (2016) Hierarchical attention networks. arXiv:abs/160602393

Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H (2017) Bidirectional attention flow for machine comprehension. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=HJ0UKP9ge

Sermanet P, Frome A, Real E (2015) Attention for fine-grained categorization. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, May 7–9, 2015, workshop track proceedings. arXiv:abs/1412.7054

Serra J, Suris D, Miron M, Karatzoglou A (2018) Overcoming catastrophic forgetting with hard attention to the task. In: International conference on machine learning. PMLR, pp 4548–4557

Serrano S, Smith NA (2019) Is attention interpretable? In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 2931–2951. https://doi.org/10.18653/v1/p19-1282

Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019

Shao Y, Gouws S, Britz D, Goldie A, Strope B, Kurzweil R (2017) Generating high-quality and informative conversation responses with sequence-to-sequence models. In: Palmer M, Hwa R, Riedel S (eds) Proceedings of the 2017 conference on empirical methods in natural language processing, EMNLP 2017, Copenhagen, September 9–11, 2017. Association for Computational Linguistics, pp 2210–2219. https://doi.org/10.18653/v1/d17-1235

Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv:151104119

Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. https://ai.google/research/pubs/pub46989

She H, Wu B, Wang B, Chi R (2018) Distant supervision for relation extraction with hierarchical attention and entity descriptions. In: Proceedings of the IEEE IJCNN. IEEE, pp 1–8

Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C (2018a) Disan: Directional self-attention network for rnn/cnn-free language understanding. In: Proceedings of the AAAI conference on artificial intelligence

Shen T, Zhou T, Long G, Jiang J, Wang S, Zhang C (2018b) Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In: Lang J (ed) Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI 2018, July 13–19, 2018, Stockholm. ijcai.org, pp 4345–4352. https://doi.org/10.24963/ijcai.2018/604

Shen T, Zhou T, Long G, Jiang J, Zhang C (2018c) Bi-directional block self-attention for fast and memory-efficient sequence modeling. In: 6th International conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=H1cWzoxA-

Shih KJ, Singh S, Hoiem D (2016) Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4613–4621

Shin B, Lee T, Choi JD (2017) Lexicon integrated CNN models with attention for sentiment analysis. In: Balahur A, Mohammad SM, van der Goot E (eds) Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis, WASSA@EMNLP 2017,

Copenhagen, September 8, 2017. Association for Computational Linguistics, pp 149–158. https://doi.org/10.18653/v1/w17-5220
Shuai B, Zuo Z, Wang B, Wang G (2017) Scene segmentation with dag-recurrent neural networks. IEEE PAMI 40(6):1480–1493
Shuang K, Ren X, Yang Q, Li R, Loo J (2019) Aela-dlstms: attention-enabled and location-aware double lstms for aspect-level sentiment classification. Neurocomputing 334:25–34
Si J, Zhang H, Li CG, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE CVPR, pp 5363–5372
Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1227–1236
Song J, Gao L, Guo Z, Liu W, Zhang D, Shen HT (2017a) Hierarchical LSTM with adjusted temporal attention for video captioning. In: Sierra C (ed) Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, Melbourne. August 19–25, 2017, ijcai.org, pp 2737–2743. https://doi.org/10.24963/ijcai.2017/381
Song S, Lan C, Xing J, Zeng W, Liu J (2017b) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence
Song C, Huang Y, Ouyang W, Wang L (2018) Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE CVPR, pp 1179–1188
Song M, Park H, Ks Shin (2019a) Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in korean. Inf Process Manag 56(3):637–653
Song Y, Wang J, Jiang T, Liu Z, Rao Y (2019b) Attentional encoder network for targeted sentiment classification. arXiv:190209314
Sordoni A, Bachman P, Trischler A, Bengio Y (2016) Iterative alternating neural attention for machine reading. arXiv:160602245
Sperber M, Neubig G, Niehues J, Waibel A (2019) Attention-passing models for robust and data-efficient end-to-end speech translation. Trans Assoc Computat Linguist 7:313–325
Stollenga MF, Masci J, Gomez FJ, Schmidhuber J (2014) Deep networks with internal selective attention through feedback connections. In: NIPS
Strubell E, Verga P, Andor D, Weiss D, McCallum A (2018) Linguistically-informed self-attention for semantic role labeling. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, October 31–November 4, 2018. Association for Computational Linguistics, pp 5027–5038. https://www.aclweb.org/anthology/D18-1548/
Sudhakaran S, Escalera S, Lanz O (2019) Lsta: Long short-term attention for egocentric action recognition. In: Proceeedings of the IEEE/CVF CVPR, pp 9954–9963
Suganuma M, Liu X, Okatani T (2019) Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9039–9048
Sukhbaatar S, Weston J, Fergus R, et al. (2015) End-to-end memory networks. In: Advances in neural information processing systems, pp 2440–2448
Sukhbaatar S, Grave E, Bojanowski P, Joulin A (2019) Adaptive attention span in transformers. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 331–335. https://doi.org/10.18653/v1/p19-1032
Sun B, Zhu Y, Xiao Y, Xiao R, Wei Y (2018) Automatic question tagging with deep neural networks. IEEE Trans Learn Technol 12(1):29–43
Tachibana H, Uenoyama K, Aihara S (2018) Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. 2018 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 4784–4788
Tan M, Santos Cd, Xiang B, Zhou B (2015) Lstm-based deep learning models for non-factoid answer selection. arXiv:151104108
Tan YK, Xu X, Liu Y (2016) Improved recurrent neural networks for session-based recommendations. In: Proceedings of the 1st workshop on deep learning for recommender systems, pp 17–22
Tan Z, Wang M, Xie J, Chen Y, Shi X (2018) Deep semantic role labeling with self-attention. In: Proceedings of the AAAI conference on artificial intelligence

Tan ZX, Goel A, Nguyen TS, Ong DC (2019) A multimodal lstm for predicting listener empathic responses over time. In: 2019 14th IEEE international conference on automatic face and gesture recognition (FG 2019). IEEE, pp 1–4

Tang H, Xu D, Sebe N, Wang Y, Corso JJ, Yan Y (2019a) Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2417–2426

Tang H, Xu D, Sebe N, Yan Y (2019b) Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: 2019 International joint conference on neural networks (IJCNN). IEEE, pp 1–8

Tay Y, Luu AT, Hui SC, Su J (2018) Densely connected attention propagation for reading comprehension. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31. Curran Associates, Inc., pp 4906–4917

Tenney I, Das D, Pavlick E (2019) BERT rediscovers the classical NLP pipeline. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 4593–4601. https://doi.org/10.18653/v1/p19-1452

Tian T, Fang ZF (2019) Attention-based autoencoder topic model for short texts. Procedia Comput Sci 151:1134–1139

Tian J, Li C, Shi Z, Xu F (2018a) A diagnostic report generator from ct volumes on liver tumor with semi-supervised attention mechanism. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 702–710

Tian W, Wang Z, Shen H, Deng W, Meng Y, Chen B, Zhang X, Zhao Y, Huang X (2018b) Learning better features for face detection with feature fusion and segmentation supervision. arXiv:181108557

Tootell RB, Hadjikhani N, Hall EK, Marrett S, Vanduffel W, Vaughan JT, Dale AM (1998) The retinotopy of visual spatial attention. Neuron 21(6):1409–1422

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognit Psychol 12(1):97–136

Tsai YH, Srivastava N, Goh H, Salakhutdinov R (2020) Capsules with inverted dot-product attention routing. In: 8th International conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net. https://openreview.net/forum?id=HJe6uANtwH

Van Den Oord A, Dieleman S, Schrauwen B (2013) Deep content-based music recommendation. In: Neural information processing systems conference (NIPS 2013), neural information processing systems foundation (NIPS), vol 26

Vashishth S, Upadhyay S, Tomar GS, Faruqui M (2019) Attention interpretability across nlp tasks. arXiv:190911218

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, pp 5998–6008

Van der Velde F, de Kamps M et al (2004) Clam: closed-loop attention model for visual search. Neurocomputing 58:607–612

Veldhuizen MG, Bender G, Constable RT, Small DM (2007) Trying to detect taste in a tasteless solution: modulation of early gustatory cortex by attention to taste. Chem Sens 32(6):569–581

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In: International conference on learning representations. https://openreview.net/forum?id=rJXMpikCZ

Vemula A, Muelling K, Oh J (2018) Social attention: Modeling attention in human crowds. In: IEEE ICRA. IEEE, pp 1–7

Verga P, Strubell E, McCallum A (2018) Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2018, New Orleans, June 1–6, 2018, volume 1 (long papers). Association for Computational Linguistics, pp 872–884. https://doi.org/10.18653/v1/n18-1080

Vig J, Belinkov Y (2019) Analyzing the structure of attention in a transformer language model. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, pp 63–76

Vinyals O, Le Q (2015) A neural conversational model. arXiv:150605869

Vinyals O, Fortunato M, Jaitly N (2015a) Pointer networks. In: Advances in neural information processing systems, pp 2692–2700

Vinyals O, Kaiser Ł, Koo T, Petrov S, Sutskever I, Hinton G (2015b) Grammar as a foreign language. In: Advances in neural information processing systems, pp 2773–2781

Vinyals O, Toshev A, Bengio S, Erhan D (2015c) Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164

Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D (2016) Matching networks for one shot learning. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R (eds) Advances in neural information processing systems 29: annual conference on neural information processing systems 2016, December 5–10, 2016, Barcelona, pp 3630–3638

Walther D (2006) Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics. PhD thesis, California Institute of Technology

Walther D, Edgington DR, Koch C (2004) Detection and tracking of objects in underwater video. In: Proceedings of the IEEE CVPR, vol 1. IEEE, pp I–I

Wang S, Jiang J (2016) Learning natural language inference with LSTM. In: Knight K, Nenkova A, Rambow O (eds) NAACL HLT 2016, The 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, San Diego California, June 12–17, 2016. The Association for Computational Linguistics, pp 1442–1451. https://doi.org/10.18653/v1/n16-1170

Wang F, Tax DM (2016) Survey on the attention based rnn model and its applications in computer vision. arXiv:160106823

Wang B, Liu K, Zhao J (2016a) Inner attention based recurrent neural networks for answer selection. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1288–1297

Wang L, Cao Z, De Melo G, Liu Z (2016b) Relation classification via multi-level attention cnns. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1298–1307

Wang X, Gao L, Song J, Shen H (2016c) Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition. IEEE Signal Process Lett 24(4):510–514

Wang Y, Huang M, Zhu X, Zhao L (2016d) Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615

Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017a) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164

Wang W, Pan SJ, Dahlmeier D, Xiao X (2017b) Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Proceedings of the AAAI conference on artificial intelligence

Wang Z, Chen T, Li G, Xu R, Lin L (2017c) Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE ICCV, pp 464–472

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018a) Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv:180407461

Wang C, Zhang Q, Huang C, Liu W, Wang X (2018b) Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 365–381

Wang J, Jiang W, Ma L, Liu W, Xu Y (2018c) Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7190–7198

Wang Q, Liu S, Chanussot J, Li X (2018d) Scene classification with recurrent attention of vhr remote sensing images. IEEE Trans Geosci Remote Sens 57(2):1155–1167

Wang W, Xu Y, Shen J, Zhu SC (2018e) Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4271–4280

Wang W, Yan M, Wu C (2018f) Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 1705–1714

Wang X, Peng Y, Lu L, Lu Z, Summers RM (2018g) Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9049–9058

Wang Y, Jiang L, Yang MH, Li LJ, Long M, Fei-Fei L (2018h) Eidetic 3d lstm: a model for video prediction and beyond. In: International conference on learning representations

Wang L, Huang Y, Hou Y, Zhang S, Shan J (2019a) Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10296–10305

Wang P, Han J, Li C, Pan R (2019b) Logic attention based neighborhood aggregation for inductive knowledge graph embedding. Proc AAAI Conf Artif Intell 33:7152–7159
Wang X, Cai Z, Gao D, Vasconcelos N (2019c) Towards universal object detection by domain attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7289–7298
Wang X, Ji H, Shi C, Wang B, Ye Y, Cui P, Yu PS (2019d) Heterogeneous graph attention network. In: The world wide web conference, pp 2022–2032
Wang X, Li R, Mallidi SH, Hori T, Watanabe S, Hermansky H (2019e) Stream attention-based multi-array end-to-end speech recognition. ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 7105–7109
Wang Y, Fan X, Chen IF, Liu Y, Chen T, Hoffmeister B (2019f) End-to-end anchored speech recognition. In: ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 7090–7094
Watanabe S, Hori T, Kim S, Hershey JR, Hayashi T (2017) Hybrid ctc/attention architecture for end-to-end speech recognition. IEEE J Sel Top Signal Process 11(8):1240–1253
Weston J, Chopra S, Bordes A (2014) Memory networks. arXiv:14103916
Wiegreffe S, Pinter Y (2019) Attention is not not explanation. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, November 3–7, 2019. Association for Computational Linguistics, pp 11–20. https://doi.org/10.18653/v1/D19-1002
Wojna Z, Gorban AN, Lee DS, Murphy K, Yu Q, Li Y, Ibarz J (2017) Attention-based extraction of structured information from street view imagery. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), , vol 1. IEEE, pp 844–850
Woldorff MG, Gallen CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, Bloom FE (1993) Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc Natl Acad Sci 90(18):8722–8726
Wolfe JM, Cave KR, Franzel SL (1989) Guided search: an alternative to the feature integration model for visual search. J Exp Psychol Hum Percep Perform 15(3):419
Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. (2016) Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:160908144
Wu C, Wei Y, Chu X, Weichen S, Su F, Wang L (2018a) Hierarchical attention-based multimodal fusion for video captioning. Neurocomputing 315:362–370
Wu L, Wang Y, Li X, Gao J (2018b) Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Trans Cybern 49(5):1791–1802
Wu W, Chen Y, Xu J, Zhang Y (2018c) Attention-based convolutional neural networks for chinese relation extraction. In: Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer, pp 147–158
Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst
Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, pp 20–27
Xiao T, Xu Y, Yang K, Zhang J, Peng Y, Zhang Z (2015) The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE CVPR, pp 842–850
Xiao S, Feng J, Xing J, Lai H, Yan S, Kassim A (2016) Robust facial landmark detection via recurrent attentive-refinement networks. In: European conference on computer vision. Springer, pp 57–72
Xiao F, Li J, Zhao H, Wang R, Chen K (2019) Lattice-based transformer encoder for neural machine translation. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, July 28–August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 3090–3097. https://doi.org/10.18653/v1/p19-1298
Xie D, Deng C, Wang H, Li C, Tao D (2019a) Semantic adversarial network with multi-scale pyramid attention for video classification. Proc AAAI Confer Artif Intell 33:9030–9037
Xie H, Fang S, Zha ZJ, Yang Y, Li Y, Zhang Y (2019b) Convolutional attention networks for scene text recognition. ACM Trans Multimed Comput Commun Appl (TOMM) 15(1s):1–17

Xie S, Hu H, Wu Y (2019c) Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognit 92:177–191

Xin M, Zhang H, Sun M, Yuan D (2016) Recurrent temporal sparse autoencoder for attention-based action recognition. In: 2016 International joint conference on neural networks (IJCNN). IEEE, pp 456–463

Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma WY (2017) Topic aware neural response generation. In: Proceedings of the AAAI conference on artificial intelligence

Xing C, Wu Y, Wu W, Huang Y, Zhou M (2018) Hierarchical recurrent attention network for response generation. In: Proceedings of the AAAI conference on artificial intelligence

Xiong C, Merity S, Socher R (2016) Dynamic memory networks for visual and textual question answering. In: Balcan M, Weinberger KQ (eds) Proceedings of the 33nd international conference on machine learning, ICML 2016, New York City, June 19–24, 2016, JMLR.org, JMLR workshop and conference proceedings, vol 48, pp 2397–2406. http://proceedings.mlr.press/v48/xiong16.html

Xiong C, Zhong V, Socher R (2017) Dynamic coattention networks for question answering. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=rJeKjwvclx

Xu H, Saenko K (2016) Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European conference on computer vision. Springer, pp 451–466

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–205. http://proceedings.mlr.press/v37/xuc15.html

Xu D, Ouyang W, Alameda-Pineda X, Ricci E, Wang X, Sebe N (2017a) Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In: Advances in neural information processing systems, pp 3961–3970

Xu S, Cheng Y, Gu K, Yang Y, Chang S, Zhou P (2017b) Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 4733–4742

Xu D, Wang W, Tang H, Liu H, Sebe N, Ricci E (2018a) Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3917–3925

Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018b) Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE CVPR, pp 2119–2128

Xu K, Wu L, Wang Z, Feng Y, Witbrock M, Sheinin V (2018c) Graph2seq: graph to sequence learning with attention-based neural networks. arXiv:180400823

Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018d) Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324

Xue W, Li T (2018) Aspect based sentiment analysis with gated convolutional networks. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, volume 1: long papers, association for computational linguistics, pp 2514–2523. https://doi.org/10.18653/v1/P18-1234, https://www.aclweb.org/anthology/P18-1234/

Xue F, Wang X, Wang J, Zha H (2020) Deep visual odometry with adaptive memory. IEEE Trans Pattern Anal Mach Intell

Yang B, Mitchell TM (2017) Leveraging knowledge bases in lstms for improving machine reading. In: Barzilay R, Kan M (eds) Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, July 30–August 4, volume 1: long papers. Association for Computational Linguistics, pp 1436–1446. https://doi.org/10.18653/v1/P17-1132

Yang L, Ai Q, Guo J, Croft WB (2016a) anmm: Rranking short answer texts with attention-based neural matching model. In: Mukhopadhyay S, Zhai C, Bertino E, Crestani F, Mostafa J, Tang J, Si L, Zhou X, Chang Y, Li Y, Sondhi P (eds) Proceedings of the 25th ACM international conference on information and knowledge management, CIKM 2016, Indianapolis, October 24–28, 2016. ACM, pp 287–296. https://doi.org/10.1145/2983323.2983818

Yang Z, He X, Gao J, Deng L, Smola A (2016b) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016c) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489

Yang Z, Yuan Y, Wu Y, Cohen WW, Salakhutdinov R (2016d) Review networks for caption generation. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R (eds) Advances in neural information

processing systems 29: annual conference on neural information processing systems 2016, December 5–10, 2016, Barcelona, pp 2361–2369
Yang F, Yang Z, Cohen WW (2017a) Differentiable learning of logical rules for knowledge base reasoning. In: Advances in neural information processing systems, pp 2319–2328
Yang J, Ren P, Zhang D, Chen D, Wen F, Li H, Hua G (2017b) Neural aggregation network for video face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4362–4371
Yang B, Li J, Wong DF, Chao LS, Wang X, Tu Z (2019a) Context-aware self-attention networks. In: Proceedings of the AAAI conference on artificial intelligence, pp 387–394
Yang B, Wang L, Wong DF, Chao LS, Tu Z (2019b) Convolutional self-attention networks. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp 4040–4045
Yang F, Jin L, Lai S, Gao X, Li Z (2019c) Fully convolutional sequence recognition network for water meter number reading. IEEE Access 7:11679–11687
Yang Z, Raymond OI, Sun W, Long J (2019) Deep attention-guided hashing. IEEE Access 7:11209–11221
Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision, pp 4507–4515
Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV), pp 684–699
Yasuda Y, Wang X, Takaki S, Yamagishi J (2019) Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In: ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP). IEEE, pp 6905–6909
Ye HJ, Hu H, Zhan DC, Sha F (2018) Learning embedding adaptation for few-shot learning. arXiv:1812.03664
Yeung S, Russakovsky O, Mori G, Fei-Fei L (2016) End-to-end learning of action detection from frame glimpses in videos. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, Las Vegas, pp 2678–2687. https://doi.org/10.1109/CVPR.2016.293, http://ieeexplore.ieee.org/document/7780662/
Yin W, Schütze H, Xiang B, Zhou B (2016) Abcnn: attention-based convolutional neural network for modeling sentence pairs. Trans Assoc Comput Linguist 4:259–272
Yin Q, Wang J, Luo X, Zhai J, Jha SK, Shi YQ (2019) Quaternion convolutional neural network for color image classification and forensics. IEEE Access 7:20293–20301
You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Yu H, Wang J, Huang Z, Yang Y, Xu W (2016) Video paragraph captioning using hierarchical recurrent neural networks. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp 4584–4593. https://doi.org/10.1109/CVPR.2016.496
Yu D, Fu J, Mei T, Rui Y (2017a) Multi-level attention networks for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4709–4717
Yu Z, Yu J, Fan J, Tao D (2017b) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 1821–1830
Yu AW, Dohan D, Luong M, Zhao R, Chen K, Norouzi M, Le QV (2018a) Qanet: Combining local convolution with global self-attention for reading comprehension. In: 6th international conference on learning representations, ICLR 2018, Vancouver, April 30–May 3, 2018, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=B14TlG-RW
Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018b) Generative image inpainting with contextual attention. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, June 18–22, 2018. IEEE Computer Society, pp 5505–5514, https://doi.org/10.1109/CVPR.2018.00577
Yu L, Lin Z, Shen X, Yang J, Lu X, Bansal M, Berg TL (2018c) Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1307–1315
Yu Z, Yu J, Xiang C, Fan J, Tao D (2018d) Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
Yu Z, Yu J, Xiang C, Fan J, Tao D (2018e) Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
Yuan Y, Wang J (2018) Ocnet: Object context network for scene parsing. arXiv:180900916

Yuan Y, Xiong Z, Wang Q (2019) Vssa-net: vertical spatial sequence attention network for traffic sign detection. IEEE Trans Image Process 28(7):3423–3434

Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, pp 28–35

Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP (2018a) Memory fusion network for multi-view sequential learning. In: Thirty-second AAAI conference on artificial intelligence

Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency LP (2018b) Multi-attention recurrent network for human communication comprehension. In: Thirty-second AAAI conference on artificial intelligence

Zagoruyko S, Komodakis N (2017) Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings, OpenReview.net. https://openreview.net/forum?id=Sks9_ajex

Zambaldi V, Raposo D, Santoro A, Bapst V, Li Y, Babuschkin I, Tuyls K, Reichert D, Lillicrap T, Lockhart E, Shanahan M, Langston V, Pascanu R, Botvinick M, Vinyals O, Battaglia P (2019) Deep reinforcement learning with relational inductive biases. In: International conference on learning representations. https://openreview.net/forum?id=HkxaFoC9KQ

Zang J, Wang L, Liu Z, Zhang Q, Hua G, Zheng N (2018a) Attention-based temporal weighted convolutional neural network for action recognition. In: IFIP international conference on artificial intelligence applications and innovations. Springer, pp 97–108

Zang X, Pokle A, Vázquez M, Chen K, Niebles JC, Soto A, Savarese S (2018b) Translating navigation instructions in natural language to a high-level plan for behavioral robot navigation. In: EMNLP

Zelano C, Bensafi M, Porter J, Mainland J, Johnson B, Bremner E, Telles C, Khan R, Sobel N (2005) Attentional modulation in human primary olfactory cortex. Nat Neurosci 8(1):114–120

Zeng J, Ma X, Zhou K (2019a) Enhancing attention-based lstm with position context for aspect-level sentiment classification. IEEE Access 7:20462–20471

Zeng Z, Xie W, Zhang Y, Lu Y (2019b) Ric-unet: an improved neural network based on unet for nuclei segmentation in histology images. Ieee Access 7:21420–21428

Zenkel T, Wuebker J, DeNero J (2019) Adding interpretable attention to neural translation models improves word alignment. arXiv:190111359

Zeyer A, Irie K, Schlüter R, Ney H (2018) Improved training of end-to-end attention models for speech recognition. Proc Interspeech 2018:7–11

Zhang H, Li J, Ji Y, Yue H (2016) Understanding subtitles by character-level sequence-to-sequence learning. IEEE Trans Ind Inf 13(2):616–624

Zhang B, Xiong D, Su J (2017a) Battrae: didimensional attention-based recursive autoencoders for learning bilingual phrase embeddings. In: Proceedings of the AAAI conference on artificial intelligence

Zhang J, Du J, Dai L (2017b) A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol 1. IEEE, pp 902–907

Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2017c) View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE international conference on computer vision, pp 2117–2126

Zhang Y, Zhong V, Chen D, Angeli G, Manning CD (2017d) Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 35–45

Zhang B, Xiong D, Su J, Zhang M (2018a) Learning better discourse representation for implicit discourse relation recognition via attention networks. Neurocomputing 275:1241–1249. https://doi.org/10.1016/j.neucom.2017.09.074

Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S (2018b) Top-down neural attention by excitation backprop. Int J Comput Vis 126(10):1084–1102

Zhang L, Zhu G, Mei L, Shen P, Shah SAA, Bennamoun M (2018c) Attention in convolutional lstm for gesture recognition. In: Advances in neural information processing systems, pp 1953–1962

Zhang P, Xue J, Lan C, Zeng W, Gao Z, Zheng N (2018d) Adding attentiveness to the neurons in recurrent neural networks. In: Proceedings of the European conference on computer vision (ECCV), pp 135–151

Zhang S, Yang J, Schiele B (2018e) Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6995–7003

Zhang T, Liu B, Niu D, Lai K, Xu Y (2018f) Multiresolution graph attention networks for relevance matching. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp 933–942

Zhang X, Su J, Qin Y, Liu Y, Ji R, Wang H (2018g) Asynchronous bidirectional decoding for neural machine translation. In: Proceedings of the AAAI conference on artificial intelligence
Zhang X, Wang T, Qi J, Lu H, Wang G (2018h) Progressive attention guided recurrent network for salient object detection. In: IEEE CVPR)
Zhang Y, Du J, Wang Z, Zhang J, Tu Y (2018i) Attention based fully convolutional network for speech emotion recognition. In: 2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 1771–1775
Zhang Y, Hare J, Prügel-Bennett A (2018j) Learning to count objects in natural images for visual question answering. In: International conference on learning representations. https://openreview.net/forum?id=B12Js_yRb
Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018k) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
Zhang H, Goodfellow I, Metaxas D, Odena A (2019a) Self-attention generative adversarial networks. In: International conference on machine learning. PMLR, pp 7354–7363
Zhang JX, Ling ZH, Liu LJ, Jiang Y, Dai LR (2019b) Sequence-to-sequence acoustic modeling for voice conversion. IEEE/ACM Trans Audio Speech Lang Process 27(3):631–644
Zhang L, Liu Z, Zhang S, Yang X, Qiao H, Huang K, Hussain A (2019c) Cross-modality interactive attention network for multispectral pedestrian detection. Inf Fusion 50:20–29
Zhang M, Wang X, Fang F, Li H, Yamagishi J (2019d) Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. Proc Interspeech 2019:1298–1302
Zhang N, Deng S, Sun Z, Wang G, Chen X, Zhang W, Chen H (2019e) Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, June 2–7, 2019, volume 1 (long and short papers). Association for Computational Linguistics, pp 3016–3025. https://doi.org/10.18653/v1/n19-1306
Zhang P, Liu W, Wang H, Lei Y, Lu H (2019f) Deep gated attention networks for large-scale street-level scene segmentation. Pattern Recognit 88:702–714. https://doi.org/10.1016/j.patcog.2018.12.021
Zhang R, Li J, Sun H, Ge Y, Luo P, Wang X, Lin L (2019g) Scan: Self-and-collaborative attention network for video person re-identification. IEEE Trans Image Process 28(10):4870–4882
Zhang X, Wang X, Tang X, Zhou H, Li C (2019h) Description generation for remote sensing images using attribute attention mechanism. Remote Sens 11(6):612
Zhang XY, Shi H, Li C, Zheng K, Zhu X, Duan L (2019i) Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In: Proceedings of the AAAI conference on artificial intelligence, pp 9227–9234
Zhang Y, Li K, Li K, Zhong B, Fu Y (2019j) Residual non-local attention networks for image restoration. In: 7th International conference on learning representations, ICLR 2019, New Orleans, May 6–9, 2019, OpenReview.net. https://openreview.net/forum?id=HkeGhoA5FX
Zhang Y, Niebles JC, Soto A (2019k) Interpretable visual question answering by visual grounding from attention supervision mining. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 349–357
Zhang Y, Wang X, Jiang X, Shi C, Ye Y (2019l) Hyperbolic graph attention network. arXiv:191203046
Zhang Y, Wang ZR, Du J (2019m) Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In: 2019 International joint conference on neural networks (IJCNN). IEEE, pp 1–8
Zhang Y, Xu X, Liu X (2019n) Robust and high performance face detector. arXiv:190102350
Zhang Y, Zhou C, Chang F, Kot AC (2019o) Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing 329:144–152
Zhang Z, Liao L, Huang M, Zhu X, Chua TS (2019p) Neural multimodal belief tracker with adaptive attention for dialogue systems. In: The world wide web conference, pp 2401–2412
Zhao D, Chen Y, Lv L (2016) Deep reinforcement learning with visual attention for vehicle classification. IEEE Trans Cognit Dev Syst 9(4):356–367
Zhao B, Wu X, Feng J, Peng Q, Yan S (2017a) Diversified visual attention networks for fine-grained object classification. IEEE Trans Multimed 19(6):1245–1256
Zhao L, Li X, Zhuang Y, Wang J (2017b) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
Zhao B, Li X, Lu X, Wang Z (2018a) A cnn-rnn architecture for multi-label weather recognition. Neurocomputing 322:47–57

Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, Jia J (2018b) Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European conference on computer vision (ECCV), pp 267–283
Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision, pp 5209–5217
Zheng Z, Zheng L, Yang Y (2018) Pedestrian alignment network for large-scale person re-identification. IEEE Trans Circuits Syst Video Technol 29(10):3037–3045
Zheng M, Karanam S, Wu Z, Radke RJ (2019a) Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5735–5744
Zheng W, Li L, Zhang Z, Huang Y, Wang L (2019b) Relational network for skeleton-based action recognition. In: 2019 IEEE international conference on multimedia and expo (ICME). IEEE, pp 826–831
Zhong V, Xiong C, Keskar NS, Socher R (2019) Coarse-grain fine-grain coattention network for multi-evidence question answering. In: 7th International conference on learning representations, ICLR 2019, New Orleans, May 6–9, 2019, OpenReview.net. https://openreview.net/forum?id=Syl7OsRqY7
Zhou Y, Shao L (2018) Aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE CVPR, pp 6489–6498
Zhou J, Cao Y, Wang X, Li P, Xu W (2016a) Deep recurrent models with fast-forward connections for neural machine translation. Trans Assoc Comput Linguist 4:371–383
Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016b) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pp 207–212
Zhou Z, Huang Y, Wang W, Wang L, Tan T (2017) See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4747–4756
Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X (2018a) Commonsense knowledge aware conversation generation with graph attention. In: IJCAI, pp 4623–4629
Zhou L, Zhou Y, Corso JJ, Socher R, Xiong C (2018b) End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8739–8748
Zhou S, Dong L, Xu S, Xu B (2018c) Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin Chinese. In: Yegnanarayana B (ed) Interspeech 2018, 19th annual conference of the international speech communication association, Hyderabad, India, 2–6 September 2018, ISCA, pp 791–795. https://doi.org/10.21437/Interspeech.2018-1107
Zhou X, Li L, Dong D, Liu Y, Chen Y, Zhao WX, Yu D, Wu H (2018d) Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1118–1127
Zhu L, Yang Y (2020) Actbert: learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8746–8755
Zhu Y, Groth O, Bernstein MS, Fei-Fei L (2016) Visual7w: Grounded question answering in images. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, June 27–30, 2016. IEEE Computer Society, pp 4995–5004. https://doi.org/10.1109/CVPR.2016.540
Zhu C, Zeng M, Huang X (2018a) Sdnet: Contextualized attention-based deep network for conversational question answering. arXiv:181203593
Zhu X, Li L, Liu J, Li Z, Peng H, Niu X (2018b) Image captioning with triple-attention and stack parallel lstm. Neurocomputing 319:55–65. https://doi.org/10.1016/j.neucom.2018.08.069
Zhu Y, Ko T, Snyder D, Mak B, Povey D (2018c) Self-attentive speaker embeddings for text-independent speaker verification. In: Interspeech, pp 3573–3577