Attention mechanism in neural networks: where it comes and where it goes

Derya Soydaner

Received: 22 July 2021 / Accepted: 27 April 2022 / Published online: 25 May 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
A long time ago in the machine learning literature, the idea of incorporating a mechanism inspired by the human visual system into neural networks was introduced. This idea is named the attention mechanism, and it has gone through a long development period. Today, many works have been devoted to this idea in a variety of tasks. Remarkable performance has recently been demonstrated. The goal of this paper is to provide an overview from the early work on searching for ways to implement attention idea with neural networks until the recent trends. This review emphasizes the important milestones during this progress regarding different tasks. By this way, this study aims to provide a road map for researchers to explore the current development and get inspired for novel approaches beyond the attention.

Keywords Attention mechanism · Neural networks · Deep learning · Survey

1 Introduction

Human eye sees the world in an interesting way. We suppose as if we see the entire scene at once, but this is an illusion created by the subconscious part of our brain [1]. According to the Scanpath theory [2, 3], when the human eye looks at an image, it can see only a small patch in high resolution. This small patch is called the fovea. It can see the rest of the image in low resolution which is called the periphery. To recognize the entire scene, the eye performs feature extraction based on the fovea. The eye is moved to different parts of the image until the information obtained from the fovea is sufficient for recognition [4]. These eye movements are called saccades. The eye makes successive fixations until the recognition task is complete. This sequential process happens so quickly that we feel as if it happens all at once.

Biologically, this is called visual attention system. Visual attention has been extensively studied in psychology and neuroscience; for reviews see [6–10]. Besides, there is a large amount of literature on modeling eye movements [11–14]. These studies have been a source of inspiration for many artificial intelligence tasks. It has been discovered that the attention idea is useful from image recognition to machine translation. Therefore, different types of attention mechanisms inspired from the human visual system have been developed for years. Since the success of deep neural networks has been at the forefront for these artificial intelligence tasks, these mechanisms have been integrated into neural networks for a long time.

This survey is about the journey of attention mechanisms used with neural networks. Researchers have been investigating ways to strengthen neural network architectures with attention mechanisms for many years. The primary aim of these studies is to reduce computational burden and to improve the model performance as well. Previous work reviewed the attention mechanisms from different perspectives [15], or examined them in context of natural language processing (NLP) [16, 17]. However, in this study, we examine the development of attention mechanisms over the years, and recent trends. We begin with the first attempts to integrate the visual attention idea to neural networks, and continue until the most modern neural networks armed with attention mechanisms. One of
them is the Transformer, which is used for many studies including the GPT-3 language model [18], goes beyond convolutions and recurrence by replacing them with only attention layers [19]. Finally, we discuss how much more can we move forward, and what’s next?

2 From the late 1980s to early 2010s: the attention awakens

The first attempts at adapting attention mechanisms to neural networks go back to the late 1980s. One of the early studies is the improved version of the Neocognitron [20] with selective attention [21]. This study is then modified to recognize and segment connected characters in cursive handwriting [22]. Another study describes VISIT, a novel model that concentrates on its relationship to a number of visual areas of the brain [5]. Also, a novel architecture named Signal Channelling Attentional Network (SCAN) is presented for attentional scanning [23].

Early work on improving the attention idea for neural networks includes a variety of tasks such as target detection [24]. In another study, a visual attention system extracts regions of interest by combining the bottom-up and top-down information from the image [25]. A recognition model based on selective attention which analyses only a small part of the image at each step, and combines results in time is described [4]. Besides, a model based on the concept of selective tuning is proposed [26]. As the years go by, several studies that use the attention idea in different ways have been presented for visual perception and recognition [27–30].

By the 2000s, the studies on making attention mechanisms more useful for neural networks continued. In the early years, a model that integrates an attentional orienting where pathway and an object recognition what pathway is presented [31]. A computational model of human eye movements is proposed for an object class detection task [32]. A serial model is presented for visual pattern recognition gathering Markov models and neural networks with selective attention on the handwritten digit recognition and face recognition problems [33]. In that study, a neural network analyses image parts and generates posterior probabilities as observations to the Markov model. Also, attention idea is used for object recognition [34], and the analysis of a scene [35]. An interesting study proposes to learn sequential attention in real-world visual object recognition using a Q-learner [36]. Besides, a computational model of visual selective attention is described to automatically detect the most relevant parts of a color picture displayed on a television screen [37]. The attention idea is also used for identifying and tracking objects in multi-resolution digital video of partially cluttered environments [38].

In 2010, the first implemented system inspired by the fovea of human retina was presented for image classification [39]. This system jointly trains a restricted Boltzmann machine (RBM) and an attentional component called the fixation controller. Similarly, a novel attentional model is implemented for simultaneous object tracking and recognition that is driven by gaze data [40]. By taking advantage of reinforcement learning, a novel recurrent neural network (RNN) is described for image classification [41]. Deep Attention Selective Network (DasNet), a deep neural network with feedback connections that are learned through reinforcement learning to direct selective attention to certain features extracted from images, is presented [42].

Table 1 Summary of Notation

| Symbol | Definition |
|--------|------------|
| a      | Annotation |
| c      | Context vector |
| z      | Weight |
| e      | Energy |
| f      | Feedforward neural network |
| h      | Hidden state |
| φ      | Hard (stochastic) / soft (deterministic) attention |
| s      | Location variable |
| p      | Source position |
| K, Q, V| Keys, queries and values matrices, respectively |
| W_q, W_k, W_v| Weight matrices for queries, keys and values, respectively |
Additionally, a deep learning-based framework using attention has been proposed for generative modeling [43].

3 2015: the rise of attention

It can be said that 2015 is the golden year of attention mechanisms. Because the number of attention studies has grown like an avalanche after three main studies presented in that year. The first one proposed a novel approach for neural machine translation (NMT) [44]. As it is known, most of the NMT models belong to a family of encoder-decoders [45, 46], with an encoder and a decoder for each language. However, compressing all the necessary information of a source sentence into a fixed-length vector is an important disadvantage of this encoder-decoder approach. This usually makes it difficult for the neural network to capture all the semantic details of a very long sentence [1].

The idea that [44] introduced is an extension to the conventional NMT models. This extension is composed of an encoder and decoder as shown in Fig 1. The first part, encoder, is a bidirectional RNN (BiRNN) [47] that takes word vectors as input. The forward and backward states of BiRNN are computed. Then, an annotation $a_j$ for each word $x_j$ is obtained by concatenating these forward and backward hidden states. Thus, the encoder maps the input sentence to a sequence of annotations $\langle a_1, \ldots, a_T \rangle$. By using a BiRNN rather than conventional RNN, the annotation of each word can summarize both the preceding words and the following words. Besides, the annotation $a_j$ can focus on the words around $x_j$ because of the inherent nature of RNNs that representing recent inputs better.

In decoder, a weight $a_{ij}$ of each annotation $a_j$ is obtained by using its associated energy $e_{ij}$ that is computed by a feedforward neural network $f$ as in Eq. (1). This neural network $f$ is defined as an alignment model that can be jointly trained with the proposed architecture. In order to reduce computational burden, a multilayer perceptron (MLP) with a single hidden layer is proposed as $f$. This alignment model tells us about the relation between the inputs around position $j$ and the output at position $i$. By this way, the decoder applies an attention mechanism. As it is seen in Eq. (2), the $x_j$ is the output of softmax function:

$$e_{ij} = f(h_{i-1}, a_j)$$

$$x_j = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

Here, the probability $a_{ij}$ determines the importance of annotation $a_j$ with respect to the previous hidden state $h_{i-1}$. Finally, the context vector $c_i$ is computed as a weighted sum of these annotations as follows [44]:

$$c_i = \sum_{j=1}^{T} a_{ij} a_j$$

Based on the decoder state, the context and the last generated word, the target word $y_t$ is predicted. In order to generate a word in a translation, the model searches for the most relevant information in the source sentence to concentrate. When it finds the appropriate source positions, it makes the prediction. By this way, the input sentence is encoded into a sequence of vectors and a subset of these vectors is selected adaptively by the decoder that is relevant to predicting the target [44]. Thus, it is no longer necessary to compress all the information of a source sentence into a fixed-length vector.

The second study is the first visual attention model in image captioning [48]. Different from the previous study [44], it uses a deep convolutional neural network (CNN) as an encoder. This architecture is an extension of the neural network [49] that encodes an image into a compact representation, followed by an RNN that generates a corresponding sentence. Here, the annotation vectors $a_i \in R^D$ are extracted from a lower convolutional layer, each of which is a $D$-dimensional representation corresponding to a part of the image. Thus, the decoder selectively focuses on certain parts of an image by weighting a subset of all the feature vectors [48]. This extended architecture uses attention for salient features to dynamically come to the forefront instead of compressing the entire image into a static representation.
The context vector $c_t$ represents the relevant part of the input image at time $t$. The weight $a_i$ of each annotation vector is computed similar to Eq. (2), whereas its associated energy is computed similar to Eq. (1) by using an MLP conditioned on the previous hidden state $h_{t-1}$. The remarkable point of this study is a new mechanism $\phi$ that computes $c_t$ from the annotation vectors $a_i$ corresponding to the features extracted at different image locations:

$$c_t = \phi(\{a_i\}, \{z_i\})$$  \hspace{1cm} (4)

The definition of the $\phi$ function causes two variants of attention mechanisms: The hard (stochastic) attention mechanism is trainable by maximizing an approximate variational lower bound, i.e., by REINFORCE [50]. On the other side, the soft (deterministic) attention mechanism is trainable by standard backpropagation methods. The hard attention defines a location variable $s_t$, and uses it to decide where to focus attention when generating the $t$-th word. When the hard attention is applied, the attention locations are considered as intermediate latent variables. It assigns a multinomial distributionparametrized by $z_i$, and $c_t$ becomes a random variable. Here, $s_{t,i}$ is defined as a one-hot variable which is set to 1 if the $i$-th location is used to extract visual features [48]:

$$p(s_{t,i} = 1 | s_{<t,i}, a) = z_{t,i}$$ \hspace{1cm} (5)

$$c_t = \sum_i s_{t,i} a_i$$  \hspace{1cm} (6)

Whereas learning hard attention requires sampling the attention location $s_t$ each time, the soft attention mechanism computes a weighted annotation vector similar to [44] and takes the expectation of the context vector $c_t$ directly:

$$E_{p(s_t|x_t)}[c_t] = \sum_{i=1}^L z_{t,i} a_i$$ \hspace{1cm} (7)

Furthermore, in training the deterministic version of the model, an alternative method namely doubly stochastic attention, is proposed with an additional constraint added to the training objective to encourage the model to pay equal attention to all parts of the image.

The third study should be emphasized presents two classes of attention mechanisms for NMT: the global attention that always attends to all source words, and the local attention that only looks at a subset of source words at a time [51]. These mechanisms derive the context vector $c_t$ in different ways: Whereas the global attention considers all the hidden states of the encoder, the local one selectively focuses on a small window of context. In global attention, a variable-length alignment vector is derived similar to Eq. (2). Here, the current target hidden state $h_t$ is compared with each source hidden state $h_s$ by using a score function instead of the associated energy $e_{ij}$. Thus, the alignment vector whose size equals the number of time steps on the source side is derived. Given the alignment vector as weights, the context vector $c_t$ is computed as the weighted average over all the source hidden states. Here, score is referred as a content-based function, and three different alternatives are considered [51].

On the other side, the local attention is differentiable. Firstly, an aligned position $p_t$ is generated for each target word at a time $t$. Then, a window centered around the source position $p_t$ is used to compute the context vector as a weighted average of the source hidden states within the window. The local attention selectively focuses on a small window of context, and obtains the alignment vector from the current target state $h_t$ and the source states $h_s$ in the window [51].

The introduction of these novel mechanisms in 2015 triggered the rise of attention for neural networks. Based on the proposed attention mechanisms, significant research

Fig. 2 Examples of the attention mechanism in visual. (Top) Attending to the correct object in neural image caption generation [48]. (Bottom) Visualization of original image and question pairs, and co-attention maps namely word-level, phrase-level and question-level, respectively [52]

© Springer
has been conducted in a variety of tasks. In order to imagine the attention idea in neural networks better, two visual examples are shown in Fig. 2. A neural image caption generation task is seen in the top row that implements an attention mechanism [48]. Then, the second example shows how the attention mechanisms can be used for visual question answering [52]. Both examples demonstrate how attention mechanisms focus on parts of input images.

4 2015-2016: attack of the attention

During two years from 2015, the attention mechanisms were used for different tasks, and novel neural network architectures were presented applying these mechanisms. After the memory networks [53] that require a supervision signal instructing them how to use their memory cells, the introduction of the neural Turing machine [54] allows end-to-end training without this supervision signal, via the use of a content-based soft attention mechanism [1]. Then, end-to-end memory network [55] that is a form of memory network based on a recurrent attention mechanism is proposed.

In these years, an attention mechanism called self-attention, sometimes called intra-attention, was successfully implemented within a neural network architecture namely Long Short-Term Memory-Networks (LSTMN) [56]. It modifies the standard LSTM structure by replacing the memory cell with a memory network [53]. This is because memory networks have a set of key vectors and a set of value vectors, whereas LSTMs maintain a hidden vector and a memory vector [56]. In contrast to attention idea in [44], memory and attention are added within a sequence encoder in LSTMN. In order to compute a representation of a sequence, self-attention is described as relating different positions of it [19]. One of the first approaches of self-attention is applied for natural language inference [57].

Many attention-based models have been proposed for neural image captioning [58], abstractive sentence summarization [59], speech recognition [60, 61], automatic video captioning [62], neural machine translation [63], and recognizing textual entailment [64]. Different attention-based models perform visual question answering [65–67]. An attention-based CNN is presented for modeling sentence pairs [68]. A recurrent soft attention based model learns to focus selectively on parts of the video frames and classifies videos [69].

On the other side, several neural network architectures have been presented in a variety of tasks. For instance, Stacked Attention Network (SAN) is described for image question answering [70]. Deep Attention Recurrent Q-Network (DARQN) integrates soft and hard attention mechanisms into the structure of Deep Q-Network (DQN) [71]. Wake-Sleep Recurrent Attention Model (WS-RAM) speeds up the training time for image classification and caption generation tasks [72]. alignDRAW model, an extension of the Deep Recurrent Attention Writer (DRAW) [73], is a generative model of images from captions using a soft attention mechanism [74]. Generative Adversarial What-Where Network (GAWWN) synthesizes images given instructions describing what content to draw in which location [75].

5 The transformer: return of the attention

After the proposed attention mechanisms in 2015, researchers published studies that mostly modifying or implementing them to different tasks. However, in 2017, a novel neural network architecture, namely the Transformer, based entirely on self-attention was presented [19]. The Transformer achieved great results on two machine translation tasks in addition to English constituency parsing. The most impressive point about this architecture is that it contains neither recurrence nor convolution. The Transformer performs well by replacing the conventional recurrent layers in encoder-decoder architecture used for NMT with self-attention.

The Transformer is composed of encoder-decoder stacks each of which has six identical layers within itself. In Fig. 3, one encoder-decoder stack is shown to illustrate the model [19]. Each stack includes only attention mechanisms and feedforward neural networks. As this architecture does not include any recurrent or convolutional layer, information about the relative or absolute positions in the input sequence is given at the beginning of both encoder and decoder using positional encodings.

The calculations of self-attention are slightly different from the mechanisms described so far in this paper. It uses three vectors namely query, key and value for each word. These vectors are computed by multiplying the input with weight matrices $W_q$, $W_k$ and $W_v$ which are learned during training. In general, each value is weighted by a function of the query with the corresponding key. The output is computed as a weighted sum of the values. Based on this idea, two attention mechanisms are proposed: In the first one, called scaled dot-product attention, the dot products of the query with all keys are computed as given in the right side of Fig. 3. Each result is divided to the square root of the dimension of the keys to have more stable gradients. They pass into the softmax function, thus the weights for the values are obtained. Finally each softmax score is multiplied with the value as given in Eq. (8). The authors propose computing the attention on a set of queries...
simultaneously by taking queries and keys of dimension $d_k$, and values of dimension $d_v$ as inputs. The keys, queries and values are packed together into matrices $K$, $Q$ and $V$. Finally, the output matrix is obtained as follows [19]:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V$$

This calculation is performed by every word against the other words. This leads to having values of each word relative to each other. For instance, if the word $x_2$ is not relevant for the word $x_1$, then the softmax score gives low probability scores. As a result, the corresponding value is decreased. This leads to an increase in the value of relevant words, and those of others decrease. In the end, every word obtains a new value for itself.

As seen from Fig. 3, the Transformer model does not directly use scaled dot-product attention. But the attention mechanism it uses is based on these calculations. The second mechanism proposed, called the multi-head attention, linearly projects the queries, keys and values $h$ times with different, learned linear projections to $d_{q}$, $d_{k}$ and $d_{v}$ dimensions, respectively [19]. The attention function is performed in parallel on each of these projected versions of queries, keys and values, i.e., heads. By this way, $d_v$-dimensional output values are obtained. In order to get the final values, they are concatenated and projected one last time as shown in the center of Fig. 3. By this way, the self-attention is calculated multiple times using different sets of query, key and value vectors. Thus, the model can jointly attend to information at different positions [19]:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

In the decoder part of the Transformer, masked multi-head attention is applied first to ensure that only previous word embeddings are used when trying to predict the next word in the sentence. Therefore, the embeddings that should not be seen by the decoder are masked by multiplying with zero.

An interesting study examines the contribution made by individual attention heads in the encoder [76]. Also, there is an evaluation of the effects of self-attention on gradient propagation in recurrent networks [77]. For a deeper analysis of multi-head self-attention mechanism from a theoretical perspective see [78].

Self-attention has been used successfully in a variety of tasks including sentence embedding [79] and abstractive summarization [80]. It is shown that self-attention can lead to improvements to discriminative constituency parser [81], and speech recognition as well [82, 83]. Also, the listen-attend-spell model [84] has been improved with the self-attention for acoustic modeling [85].

As soon as these self-attention mechanisms were proposed, they have been incorporated with deep neural networks for a wide range of tasks. For instance, a deep learning model learned a number of large-scale tasks from multiple domains with the aid of self-attention mechanism [86]. Novel self-attention neural models are proposed for cross-target stance classification [87] and NMT [88]. Another study points out that a fully self-attentional model can reach competitive predictive performance on ImageNet.
classification and COCO object detection tasks [89]. Besides, developing novel attention mechanisms has been carried out such as area attention, a novel mechanism that can be used along multi-head attention [90]. It attends to areas in the memory by defining the key of an area as the mean vector of the key of each item, and defining the value as the sum of all value vectors in the area.

When a novel mechanism is proposed, it is inevitable to incorporate it into the GAN framework [91]. Self-Attention Generative Adversarial Networks (SAGANs) [92] introduce a self-attention mechanism into convolutional GANs. Different from the traditional convolutional GANs, SAGAN generates high-resolution details using cues from all feature locations. Similarly, Attentional Generative Adversarial Network (AttnGAN) is presented for text to image generation [93]. On the other side, a machine reading and question answering architecture called QANet [94] is proposed without any recurrent networks. It uses self-attention to learn the global interaction between each head's output by introducing gates [95]. Another interesting study introduces attentive group convolutions with a generalization of visual self-attention [96]. A deep transformer model is implemented for language modeling over long sequences [97].

5.1 Self-attention variants

In recent years, self-attention has become an important research direction within the deep learning community. Self-attention idea has been examined in different aspects. For example, self-attention is handled in a multi-instance learning framework [98]. The idea of Sparse Adaptive Connection (SAC) is presented for accelerating and structuring self-attention [99]. The research on improving self-attention continues as well [100–102]. Besides, based on the self-attention mechanisms proposed in the Transformer, important studies that modify the self-attention have been presented. Some of the most recent and prominent studies are summarized below.

5.1.1 Relation-aware self-attention

It extends the self-attention mechanism by regarding representations of the relative positions, or distances between sequence elements [103]. Thus, it can consider the pairwise relationships between input elements. This type of attention mechanism defines vectors to represent the edge between two inputs. It provides learning two distinct edge representations that can be shared across attention heads without requiring additional linear transformations.

5.1.2 Directional self-attention (DiSA)

A novel neural network architecture for learning sentence embedding named Directional Self-Attention Network (DiSAN) [104] uses directional self-attention followed by a multi-dimensional attention mechanism. Instead of computing a single importance score for each word based on the word embedding, multi-dimensional attention computes a feature-wise score vector for each token. To extend this mechanism to the self-attention, two variants are presented: The first one, called multi-dimensional ‘token2token’ self-attention generates context-aware coding for each element. The second one, called multi-dimensional ‘source2token’ self-attention compresses the sequence into a vector [104]. On the other side, directional self-attention produces context-aware representations with temporal information encoded by using positional masks. By this way, directional information is encoded. First, the input sequence is transformed to a sequence of hidden states by a fully connected layer. Then, multi-dimensional token2token self-attention is applied to these hidden states. Hence, context-aware vector representations are generated for all elements from the input sequence.

5.1.3 Reinforced self-attention (ReSA)

A sentence-encoding model named Reinforced Self-Attention Network (ReSAN) uses reinforced self-attention (ReSA) that integrates soft and hard attention mechanisms into a single model. ReSA selects a subset of head tokens, and relates each head token to a small subset of dependent tokens to generate their context-aware representations [105]. For this purpose, a novel hard attention mechanism called reinforced sequence sampling (RSS), which selects tokens from an input sequence in parallel and trained via policy gradient, is proposed. Given an input sequence, RSS generates an equal-length sequence of binary random variables that indicates both the selected and discarded ones. On the other side, the soft attention provides reward signals back for training the hard attention. The proposed RSS provides a sparse mask to self-attention. ReSA uses two RSS modules to extract the sparse dependencies between each pair of selected tokens.

5.1.4 Outer product attention (OPA)

Self-Attentive Associative Memory (SAM) is a novel operator based upon outer product attention (OPA) [106]. This attention mechanism is an extension of dot-product attention [19]. OPA differs using element-wise multiplication, outer product, and tanh function instead of softmax.
5.1.5 Bidirectional block self-attention (Bi-BloSA)

Another mechanism, bidirectional block self-attention (Bi-BloSA) which is simply a masked block self-attention (mBloSA) with forward and backward masks to encode the temporal order information is presented [107]. Here, mBloSA is composed of three parts from its bottom to top namely intra-block self-attention, inter-block self-attention and the context fusion. It splits a sequence into several length-equal blocks, and applies an intra-block self-attention to each block independently. Then, inter-block self-attention processes the outputs for all blocks. This stacked self-attention model results a reduction in the amount of memory compared to a single one applied to the whole sequence. Finally, a feature fusion gate combines the outputs of intra-block and inter-block self-attention with the original input, to produce the final context-aware representations of all tokens.

5.1.6 Fixed multi-head attention

The fixed multi-head attention proposes fixing the head size of the Transformer in the aim of improving the representation power [108]. This study emphasizes its importance by setting the head size of attention units to input sequence length.

5.1.7 Sparse sinkhorn attention

It is based on the idea of differentiable sorting of internal representations within the self-attention module [109]. Instead of allowing tokens to only attend to tokens within the same block, it operates on block sorted sequences. Each token attends to tokens in the sorted block. Thus, tokens that may be far apart in the unsorted sequence can be considered. Additionally, a variant of this mechanism named SortCut sinkhorn attention applies a post-sorting truncation of the input sequence.

5.1.8 Adaptive attention span

Adaptive attention span is proposed as an alternative to self-attention [110]. It learns the attention span of each head independently. To this end, a masking function inspired by [111] is used to control the attention span for each head. The purpose of this novel mechanism is to reduce the computational burden of the Transformer. Additionally, dynamic attention span approach is presented to dynamically change the attention span based on the current input as an extension [51, 112].

5.2 Transformer variants

Different from developing novel self-attention mechanisms, several studies have been published in the aim of improving the performance of the Transformer. These studies mostly modify the model architecture. For instance, an additional recurrence encoder is preferred to model recurrence for Transformer directly [113]. In another study, a new weight initialization scheme is applied to improve Transformer optimization [114]. A novel positional encoding scheme is used to extend the Transformer to tree-structured data [115]. Investigating model size by handling Transformer width and depth for efficient training is also an active research area [116]. Transformer is used in reinforcement learning settings [117–119] and for time-series forecasting in adversarial training setting [120].

Besides, many Transformer variants have been presented in the recent past. COMmonsEnse Transformer (COMET) is introduced for automatic construction of commonsense knowledge bases [121]. Evolved Transformer applies neural architecture search for a better Transformer model [122]. Transformer Autoencoder is a sequential autoencoder for conditional music generation [123]. CrossTransformer takes a small number of labeled images and an unlabeled query, and computes distances between spatially-corresponding features to infer class membership [124]. DEtection Transformer (DETR) is a new design for object detection systems [125], and Deformable DETR is an improved version that achieves better performance in less time [126]. FLOw-bAsed TransformER (FLOATER) emphasizes the importance of position encoding in the Transformer, and models the position information via a continuous dynamical model [127]. Disentangled Context (DisCo) Transformer simultaneously generates all tokens given different contexts by predicting every word in a sentence conditioned on an arbitrary subset of the rest of the words [128]. Generative Adversarial Transformer (GANformer) is presented for visual generative modeling [129].

Recent work has demonstrated significant performance on NLP tasks. In OpenAI GPT, there is a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer [130]. GPT-2 [131] and GPT-3 [18] models have improved the progress. In addition to these variants, some prominent Transformer-based models are summarized below.

5.2.1 Universal transformer

A generalization of the Transformer model named the Universal Transformer [132] iteratively computes representations $H^t$ at step $t$ for all positions in the sequence in
parallel. To this end, it uses the scaled dot-product attention in Eq. (8) where $d$ is the number of columns of $Q$, $K$, and $V$. In the Universal Transformer, the multi-head self-attention with $k$ heads is used. The representations $H'$ are mapped to queries, keys and values with affine projections using learned parameter matrices $W^Q \in \mathbb{R}^{d \times d/k}$, $W^K \in \mathbb{R}^{d \times d/k}$, $W^V \in \mathbb{R}^{d \times d/k}$ and $W^O \in \mathbb{R}^{d \times d}$ [132]:

$$\text{MultiHead}(H') = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(H'W_i^Q ; H'W_i^K ; H'W_i^V)$ (10)

### 5.2.2 Image transformer

Image Transformer [133] demonstrates that self-attention-based models can also be well-suited for images instead of text. This Transformer type restricts the self-attention mechanism to attend to local neighborhoods. Thus, the size of images that the model can process is increased. Its larger receptive fields allow the Image Transformer to significantly improve the model performance on image generation as well as image super-resolution.

### 5.2.3 Transformer-XL

This study aims to improve the fixed-length context of the Transformer [19] for language modeling. Transformer-XL [134] makes modeling very long-term dependency possible by reusing the hidden states obtained in previous segments. Hence, information can be propagated through the recurrent connections. In order to reuse the hidden states without causing temporal confusion, Transformer-XL uses relative positional encodings. Based on this architecture, a modified version named the Gated Transformer-XL (GTrXL) is presented in the reinforcement learning setting [135].

### 5.2.4 Tensorized Transformer

Tensorized Transformer [136] compresses the multi-head attention in Transformer. To this end, it uses a novel self-attention model multi-linear attention with Block-Term Tensor Decomposition (BTDT) [137]. It builds a single-block attention based on the Tucker decomposition [138]. Then, it uses a multi-linear attention constructed by a BTDT to compress the multi-head attention mechanism. In Tensorized Transformer, the factor matrices are shared across multiple blocks.

### 5.2.5 BERT

The Bidirectional Encoder Representations from Transformers (BERT) aims to pre-train deep bidirectional representations from unlabeled text [139]. BERT uses a multilayer bidirectional Transformer as the encoder. Besides, inspired by the Cloze task [140], it has a masked language model pre-training objective. BERT randomly masks some of the tokens from the input, and predicts the original vocabulary id of the masked word based only on its context. This model can pre-train a deep bidirectional Transformer. In all layers, the pre-training is carried out by jointly conditioning on both left and right context. BERT differs from the left-to-right language model pre-training from this aspect.

Recently, BERT model has been examined in detail. For instance, the behavior of attention heads are analysed [141]. Various methods have been investigated for compressing [142, 143], pruning [144], and quantization [145]. Also, BERT model has been considered for different tasks such as coreference resolution [146]. A novel method is proposed in order to accelerate BERT training [147].

Furthermore, various BERT variants have been presented. ALBERT aims to increase the training speed of BERT, and presents two parameter reduction techniques [148]. Similarly, PoWER-BERT [149] is developed to improve the inference time of BERT. This scheme is also used to accelerate ALBERT. Also, TinyBERT is proposed to accelerate inference and reduce model size while maintaining accuracy [150]. In order to obtain better representations, SpanBERT is proposed as a pre-training method [151]. As a robustly optimized BERT approach, RoBERTa shows that BERT was significantly undertrained [152]. Also, DeBERTa improves RoBERTa using the disentangled attention mechanism [153]. On the other side, DistilBERT shows that it is possible to reach similar performances using much smaller language models pretrained with knowledge distillation [154]. StructBERT proposes two novel linearization strategies [155]. Q-BERT is introduced for quantizing BERT models [156], BioBERT is for biomedical text mining [157], and RareBERT is for rare disease diagnosis [158].

Since 2017 when the Transformer was presented, research directions have generally focused on novel self-attention mechanisms, adapting the Transformer for various tasks, or making them more understandable. In one of the most recent studies, NLP becomes possible in the mobile setting with Lite Transformer. It applies long-short range attention where some heads specialize in the local context modeling while the others specialize in the long-distance relationship modeling [159]. A deep and lightweight Transformer DeLightT [160] and a hypernetwork-based model namely HyperGrid Transformers [161] perform with fewer parameters. Graph Transformer Network is introduced for learning node representations on heterogeneous graphs [162] and different applications are performed for molecular data [163] or textual graph representation [164]. Also, Transformer-XH applies eXtra
Hop attention for structured text data [165]. AttentionXML is a tree-based model for extreme multi-label text classification [166]. Besides, attention mechanism is handled in a Bayesian framework [167]. For a better understanding of Transformers, an identifiability analysis of self-attention weights is conducted in addition to presenting effective attention to improve explanatory interpretations [168]. Lastly, Vision Transformer (ViT) processes an image using a standard Transformer encoder as used in NLP by interpreting it as a sequence of patches, and performs well on image classification tasks [169].

5.3 What about complexity?

All these aforementioned studies undoubtedly demonstrate significant success. But success not make one great. The Transformer also brings a very high computational complexity and memory cost. The necessity of storing attention matrix to compute the gradients with respect to queries, keys and values causes a non-negligible quadratic computation and memory requirements. Training the Transformer is a slow process for very long sequences because of its quadratic complexity. There is also time complexity which is quadratic with respect to the sequence length. In order to improve the Transformer in this respect, recent studies have been conducted to improve this issue. One of them is Linear Transformer which expresses the self-attention as a linear dot-product of kernel feature maps [170]. Linear Transformer reduces both memory and time complexity by changing the self-attention from the softmax function in Eq. (8) to a feature map-based dot-product attention. Its performance is competitive with the vanilla Transformer architecture on image generation and automatic speech recognition tasks while being faster during inference. On the other side, FMMformers which use the idea of the fast multipole method (FMM) [171] outperform the linear Transformer by decomposing the attention matrix into near-field and far-field attention with linear time and memory complexity [172].

Another suggestion made in response to the Transformer’s quadratic nature is The Reformer that replaces dot-product attention by one that uses locality-sensitive hashing [173]. It reduces the complexity but one limitation of the Reformer is its requirement for the queries and keys to be identical. Set Transformer aims to reduce computation time of self-attention from quadratic to linear by using an attention mechanism based on sparse Gaussian process literature [174]. Routing Transformer aims to reduce the overall complexity of attention by learning dynamic sparse attention patterns by using routing attention with clustering [175]. It applies k-means clustering to model sparse attention matrices. At first, queries and keys are assigned to clusters. The attention scheme is determined by considering only queries and keys from the same cluster. Thus, queries are routed to keys belonging to the same cluster [175].

Sparse Transformer introduces sparse factorizations of the attention matrix by using factorized self-attention, and avoids the quadratic growth of computational burden [176]. It also shows the possibility of modeling sequences of length one million or more by using self-attention in theory. In the Transformer, all the attention heads with the softmax attention assign a nonzero weight to all context words. Adaptively Sparse Transformer replaces softmax with z-entmax which is a differentiable generalization of softmax allowing low-scoring words to receive precisely zero weight [177]. By means of context-dependent sparsity patterns, the attention heads become flexible in the Adaptively Sparse Transformer. Random feature attention approximates softmax attention with random feature methods [178]. Skyformer replaces softmax with a Gaussian kernel and adapts Nyström method [179]. A sparse attention mechanism named BIGBIRD aims to reduce the quadratic dependency of Transformer-based models to linear [180]. Different from the similar studies, BIGBIRD performs well for genomics data alongside NLP tasks such as question answering.

Music Transformer [181] shows that self-attention can also be useful for modeling music. This study emphasizes the infeasibility of the relative position representations introduced by [103] for long sequences because of the quadratic intermediate relative information in the sequence length. Therefore, this study presents an extended version of relative attention named relative local attention that improves the relative attention for longer musical compositions by reducing its intermediate memory requirement to linear in the sequence length. A softmax-free Transformer (SOFT) is presented to improve the computational efficiency of ViT. It uses Gaussian kernel function instead of the dot-product similarity [182].

Additionally, various approaches have been presented in Hierarchical Visual Transformer [183], Long-Short Transformer (Transformer-LS) [184], Perceiver [185], and Performer [186]. Image Transformer based on the cross-covariance matrix between keys and queries is applied [187], and a new vision Transformer is proposed [188]. Furthermore, a Bernoulli sampling attention mechanism decreases the quadratic complexity to linear [189]. A novel linearized attention mechanism performs well on object detection, instance segmentation, and stereo depth estimation [190]. A study shows that kernelized attention with relative positional encoding can be calculated using Fast Fourier Transform and it leads to get rid of the quadratic complexity for long sequences [191]. A linear unified nested attention mechanism namely Luna uses two nested attention functions to approximate the softmax attention in
Transformer to achieve linear time and space complexity [192].

6 Concluding remarks: a new hope

Inspired by the human visual system, the attention mechanisms in neural networks have been developing for a long time. In this study, we examine this duration beginning with its roots up to the present time. Some mechanisms have been modified, or novel mechanisms have emerged in this period. Today, this journey has reached a very important stage. The idea of incorporating attention mechanisms into deep neural networks has led to state-of-the-art results for a large variety of tasks. Self-attention mechanisms and GPT-n family models have become a new hope for more advanced models. These promising progress bring the questions whether the attention could help further development, replace the popular neural network layers, or bring the questions whether the attention could help further development, replace the popular neural network layers, or bring the questions whether the existing attention mechanisms? It is still an active research area and much to learn we still have, but it is obvious that more powerful systems are awaiting when neural networks and attention mechanisms join forces.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press
2. Noton D, Stark L (1971) Eye movements and visual perception. Sci Am 224(6):34
3. Noton D, Stark L (1971) Scanpaths in saccadic eye movements while viewing and recognizing patterns. Vision Res 11:929
4. Alpaydín E (1995) Selective attention for handwritten digit recognition. Adv Neural Inf Process Syst 8:771–777
5. Ahmad S (1991) VISIT: a neural model of covert visual attention. Adv Neural Inf Process Syst 4:420–427
6. Posner M, Petersen S (1990) The attention system of the human brain. Annu Rev Neurosci 13(1):25
7. Bundesen C (1990) A theory of visual attention. Psychol Rev 97(4):523
8. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18(1):193
9. Corbetta M, Shulman G (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3(3):201
10. Petersen S, Posner M (2012) The attention system of the human brain: 20 years after. Annu Rev Neurosci 35:73
11. Rimey R, Brown C (1990) Selective attention as sequential behaviour: modeling eye movements with an augmented hidden markov model, Technical Report, University of Rochester
12. Sheliga B, Riggio L, Rizzolatti G (1994) Orienting of attention and eye movements. Exp Brain Res 98(3):507
13. Sheliga B, Riggio L, Rizzolatti G (1995) Spatial attention and eye movements. Exp Brain Res 105(2):261
14. Hoffman J, Subramaniam B (1995) The role of visual attention in saccadic eye movements. Percept Psychophys 57(6):787
15. Chaudhari S et al (2021) An atttive survey of attention models, ACM Transactions on Intelligent Systems and Technology (TIST) pp 1–32
16. Galassi A et al (2020) Attention in natural language processing. IEEE Trans Neural Netw Learn Syst 32:4291–4308
17. Lee J et al (2019) Attention models in graphs: a survey. ACM Trans Knowl Discov Data (TKDD) 13(6):11
18. Brown T et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
19. Vaswani A et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
20. Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193
21. Fukushima K (1987) Neural network model for selective attention in visual pattern recognition and associative recall. Appl Opt 26(23):4985
22. Fukushima K, Imagawa T (1993) Recognition and segmentation of connected characters with selective attention. Neural Netw 6(1):33
23. Posma E, den Herik HV, Hudson P (1997) SCAN: a scalable model of attentional selection. Neural Netw 10(6):993
24. Schmidhuber J, Huber R (1991) Learning to generate artificial fovea trajectories for target detection. Int J Neural Syst 2:125–134
25. Milanese R et al (1994) Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In: IEEE computer society conference on computer vision and pattern recognition, Seattle, WA, pp 781–785
26. Tsotsos J et al (1995) Modeling visual attention via selective tuning. Artif Intell 78(1–2):507
27. Culhane S, Tsotsos J (1992) A prototype for data-driven visual attention. In: Proceedings of the 11th IAPR international conference on pattern recognition, The Hague, pp 36–40
28. Reifeld D, Wolfson H (1995) Yeshurun Y, Context-free attentional operators; the generalized symmetry transform. Int J Comput Vis 14(2):119
29. Rybak I et al (1998) A model of attention-guided visual perception and recognition. Vis Res 38(15–16):2387
30. Keller J et al (1999) Object recognition based on human saccadic eye movements. Pattern Anal Appl 2(3):251–263
31. Miao F, Itti L (2001) A neural model combining attentional orienting to object recognition: preliminary explorations on the interplay between where and what. In: Proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society, Istanbul, pp 789–792
32. Zhang W et al (2006) A computational model of eye movements during object class detection. Adv Neural Inf Process Syst 19:1609–1616
33. Salah A, Alpaydín E, Akarun L (2002) A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. IEEE Trans Pattern Anal Mach Intell 24(3):420
34. Walther D et al (2002) Attentional selection for object recognition—A gentle way. In: Bültthoff HH, Wallraven C, Lee SW, Poggio TA (eds) International workshop on biologically motivated computer vision. Springer, Berlin, Heidelberg, pp 472–479
35. Schill K et al (2001) Scene analysis with saccadic eye movements: top-down and bottom-up modeling. J Electron Imaging 10(1):152
36. Paletta L, Fritz G, Seifert C (2005) Q-learning of sequential attention for visual object recognition from informative local descriptors. In: International conference on machine learning
37. Meur O.L (2006) A coherent computational approach to model bottom-up visual attention. IEEE Trans Pattern Anal Mach Intell 28(5):802
38. Gould S et al (2007) Peripheral-foveal vision for real-time object recognition and tracking in video. In: International joint conference on artificial intelligence (IJCAI) pp 2115–2121
39. Larochelle H, Hinton G (2010) Learning to combine foveal glimpses with a third-order Boltzmann machine. Adv Neural Inf Process Syst 23:1243–1251
40. Bazzani L et al (2011) Learning attentional policies for tracking and recognition in video with deep networks
41. Mnih V et al (2014) Recurrent models of visual attention. Adv Neural Inf Process Syst 27:2204–2212
42. Stollenga M et al (2014) Deep networks with internal selective attention through feedback connections. Adv Neural Inf Process Syst 27:3545–3553
43. Tang Y, Srivastava N, Salakhutdinov R (2014) Learning generative models with visual attention. Advances in Neural Information Processing Systems, 27
44. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate
45. Sutskever I, Vinyals O, Le Q (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 27:3104–3112
46. Cho K et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) pp 1724–1734
47. Schuster M, Paliwal K (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673
48. Xu K et al (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057
49. Vinyals O et al (2015) Show and tell: a neural image caption generator. In: In proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
50. Williams R (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8(3–4):229
51. Luong MT, Manning HPC (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, pp 1412–1421
52. Lu J et al (2016) Hierarchical question-image co-attention for visual question answering. Advances in Neural Information Processing Systems, 29
53. Weston J, Chopra S, Bordes A (2014) Memory networks
54. Graves A, Wayne G, Danihelka I (2014) Neural Turing Machines, arXiv preprint arXiv:1410.5401
55. Sukhbaatar S et al (2015) End-to-end memory networks. Adv Neural Inf Process Syst 28:2440–2448
56. Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 551–561
57. Parikh A et al (2016) A decomposable attention model for natural language inference. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp 2249–2255
58. You Q et al (2016) Image captioning with semantic attention. In: In proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, pp 4651–4659
59. Rush A, Chopra S, Weston J (2015) A neural attention model for sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, pp 379–389
60. Yu D et al (2016) Deep convolutional neural networks with layer-wise context expansion and attention, Interspeech pp 17–21
61. Chorowski J et al (2015) Attention-based models for speech recognition. Adv Neural Inf Process Syst 28:577–585
62. Zanfir M, Marinou E, Smintchiscu C (2016) Spatio-temporal attention models for grounded video captioning. In: Asian conference on computer vision. Springer, Cham, pp 104–119
63. Cheng Y et al (2016) Agreement-based joint training for bidirectional attention-based neural machine translation. In: Proceedings of the 25th international joint conference on artificial intelligence
64.ock T (2016) Reasoning about entailment with neural attention
65. Y. Zhu, et al., (2016) Visual7W: Grounded question answering in images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4995–5004
66. Chen K et al (2015) ABC-CNN: An attention based convolutional neural network for visual question answering, arXiv preprint arXiv:1511.05960
67. Xu H, Saenko K (2016) Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: In European conference on computer vision, pp 451–466
68. Yin W et al (2016) ABCNN: Attention-based convolutional neural network for modeling sentence pairs, Trans Assoc Comput Linguist 4:259
69. Sharma S, Kiros R, Salakhutdinov R (2016) Action recognition using visual attention
70. Yang Z et al (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
71. Sorokin I et al (2015) Deep attention recurrent Q-network, arXiv preprint arXiv:1512.01693
72. Ba J et al (2015) Learning wake-sleep recurrent attention models. Adv Neural Inf Process Syst 28:2593–2601
73. Gregor K et al (2015) DRAW: a recurrent neural network for image generation. In: International conference on machine learning, pp 1462–1471
74. Mansimov E et al (2016) Generating images from captions with attention. In: International conference on learning representations
75. Reed S et al (2016) Learning what and where to draw. Adv Neural Inf Process Syst 29:217–225
76. Voita E et al (2019) Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. In: In proceedings of the 57th annual meeting of the association for computational linguistics, florence, pp 5797–5808
77. Kerg G et al (2020) Constituency parsing with a self-attentive encoder. In: In proceedings of the 56th annual meeting
of the association for computational linguistics (Long papers) pp 2676–2686
82. Povey D et al (2018) A time-restricted self-attention layer for ASR. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5874–5878
83. Vyas A et al (2020) Fast transformers with clustered attention. Advances in Neural Information Processing Systems, 33
84. Chan W et al (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Shanghai, pp 4960–4964
85. Sperber M et al (2018) Self-attentional acoustic models. In: In proceedings of annual conference of the international speech communication association (Interspeech), pp 3723–3727
86. Kaiser L et al (2017) One model to learn them all. arXiv preprint arXiv:1706.05137
87. Xu C et al (2018) Cross-target stance classification with self-attention networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Short papers), Melbourne, pp 778–783
88. Maruf S, Martins A, Haffari G (2019) Selective attention for context-aware neural machine translation. In: Proceedings of NAACL-HLT, Minneapolis, Minnesota, pp 3092–3102
89. Ramachandran P et al (2019) Stand-alone self-attention in vision models. Adv Neural Inf Process Syst 32:68–80
90. Li Y et al (2019) Area attention
91. Goodfellow I et al (2014) Generative adversarial networks. Adv Neural Inf Process Syst 27:2672–2680
92. Zhang H et al (2019) Self-attention generative adversarial networks. In: International conference on machine learning, pp 7354–7363
93. Xu T et al (2018) Attn GAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) pp 1316–1324
94. Yu A et al (2018) QANet: combining local convolution with global self-attention for reading comprehension. In: International conference on learning representations
95. Zhang J et al (2018) GaAN: gated attention networks for learning on large and spatiotemporal graphs. In: Conference on uncertainty in artificial intelligence
96. Romero D et al (2020) Attentive group equivariant convolutional networks on trees. In: Conference on learning representations
97. Al-Rfou R et al (2019) Character-level language modeling with deeper self-attention. AAAI Conf Artif Intell 33:3159
98. Du J et al (2018) Multi-level structured self-attentions for distantly supervised relation extraction. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 2216–2225
99. Li X et al (2020) SAC: accelerating and structuring self-attention via sparse adaptive connection. Advances in Neural Information Processing Systems, 33
100. Yang B et al (2019) Context-aware self-attention networks. AAAI Conf Artif Intell 33:387
101. Yang B et al (2018) Modeling localness for self-attention networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, pp 4449–4458
102. Attention augmented convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 3286–3295
103. Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. In: Proceedings of NAACL-HLT, New Orleans, Louisiana, pp 464–468
104. Shen T et al (2018) Di SAN: directional self-attention network for RNN/CNN-free language understanding. In: AAAI Conference on artificial intelligence, pp 5446–5455
105. Shen T et al (2018) Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In: Proceedings of the 27th international joint conference on artificial intelligence (IJCAI-18), pp 4345–4352
106. Le H, Tran T, Venkatesh S (2020) Self-attentive associative memory
107. Shen T et al (2018) Bi-directional block self-attention for fast and memory-efficient sequence modeling. In: International conference on learning representations
108. Bhogranapalli S et al (2020) Low-rank bottleneck in multi-head attention models
109. Tay Y et al (2020) Sparse sinkhorn attention
110. Sukhbaatar S et al (2019) Adaptive attention span in transformers. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, pp 331–335
111. Jernite Y et al (2017) Variable computation in recurrent neural networks. In: International conference on learning representations
112. Shu R, Nakayama H (2017) An empirical study of adequate vision span for attention-based neural machine translation. In: Proceedings of the first workshop on neural machine translation, Vancouver, pp 1–10
113. Hao J et al (2019) Modeling recurrence for transformer. In: Proceedings of NAACL-HLT, Minneapolis, Minnesota, pp 1198–1207
114. Hu S et al (2020) Improving transformer optimization through better initialization
115. Shiv V, Quirk C (2019) Novel positional encodings to enable tree-based transformers. Adv Neural Inf Process Syst 33:12081–12091
116. Li Z et al (2020) Train large, then compress: rethinking model size for efficient training and inference of transformers
117. Hoshen Y (2017) VAIN: Attentional Multi-agent predictive modeling, Advances in Neural Information Processing Systems, 30, Long Beach, CA
118. Hu S et al (2021) UPDeT: universal multi-agent reinforcement learning via policy decoupling with transformers. In: International conference on learning representations
119. Parsi S, Salakhutdinov R (2021) Efficient transformers in reinforcement learning using actor-learner distillation. In: ICML
120. Wu S et al (2020) Adversarial sparse transformer for time series forecasting. Advances in Neural Information Processing Systems, 33
121. Bosselut A et al (2019) COMET: commonsense transformers for RNN/CNN-free language understanding. In: AAAI Conference on artificial intelligence, pp 5446–5455
122. So D, Liang C, Le Q (2019) The evolved transformer
123. Choi K et al (2020) Encoding musical style with transformer autoencoders
124. Doersch C, Gupta A, Zisserman A (2020) Cross Transformers: spatially-aware few-shot transfer. Adv Neural Inf Process Syst 33:21981–21993
125. Carion N et al (2020) End-to-end object detection with transformers. In: Proceedings of the 27th Annual Conference on Computer Vision and Pattern Recognition (CVPR) pp 1316–1324
126. Bhojanapalli S et al (2020) Low-rank bottleneck in multi-head attention models
127. Liu X et al (2020) Learning to encode position for transformer with continuous dynamical model. In: International conference on machine learning, pp 6327–6335
128. Kasai J et al (2020) Non-autoregressive machine translation with disentangled context transformer.
129. Hudson D, Zitnick L (2021) Generative adversarial transformers. In: International conference on machine learning, pp 4487–4499
130. Radford A et al (2018) Improving language understanding by generative pre-training. Technical Report, OpenAI
131. Radford A et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
132. Dehghani M et al (2019) Universal transformers. In: International conference on learning representations
133. Parmar N (2018) Image transformer
134. Dai Z et al (2019) Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2978–2986
135. Parisotto E (2020) Stabilizing transformers for reinforcement learning
136. Ma X et al (2019) A Tensorized Transformer for Language Modeling. Adv Neural Inf Process Syst 32:2232–2242
137. Lathauwer L (2008) Decompositions of a higher-order tensor in block terms-part ii: definitions and uniqueness. SIAM J Matrix Anal Appl 30(3):1033
138. Tucker L (1966) Some mathematical notes on three-mode factor block terms-part ii: definitions and uniqueness. SIAM J Matrix Anal Appl 30(3):1033
139. Clark K et al (2019) What does BERT look at? An analysis of
140. Devlin J et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proc of NAACL-HLT 2019:4171–4186
141. Tucker L (1966) Some mathematical notes on three-mode factor analysis. Psychometrica 31(3):279
142. Sun S et al (2019) Patient knowledge distillation for BERT inference via progressive word-vector elimination. In: Empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, pp 4323–4332
143. Wang W et al (2020) MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems, 33
144. Dehghani M et al (2019) Universal transformers. In: International conference on learning representations
145. Clark K et al (2019) What does BERT look at? An analysis of BERT’s attention, arXiv preprint arXiv:1906.04341
146. Sun S et al (2019) Patient knowledge distillation for BERT model compression. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong, pp 4323–4332
147. Joshi M et al (2019) BERT for coreference resolution: baselines and analysis. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 5803–5808
148. Gong L et al (2019) Efficient training of BERT by progressively stacking. In: International conference on machine learning, pp 2337–2346
149. Goyal S et al (2020) Po WER-BERT: accelerating BERT inference via progressive word-vector elimination
150. Zhao C et al (2020) Transformer-XH: multi-evidence reasoning with extra hop attention. In: International conference on learning representations
151. You R et al (2019) AttentionXML: label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems, 32
152. Fan X et al (2020) Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 33
153. He P et al (2021) DeBERTa: decoding-enhanced BERT with disentangled attention. In: International conference on learning representations
154. Sanh V et al (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: The 5th workshop on energy efficient machine learning and cognitive computing—NeurIPS
155. Wang W et al (2020) Struct BERT: incorporating language structures into pre-training for deep language understanding. In: International conference on learning representations
156. Shen S et al (2020) Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. AAAI Conf Artif Intell 34:8815
157. Lee J et al (2020) Bio BERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234
158. Prakash P et al (2021) Rare BERT: transformer architecture for rare disease patient identification using administrative claims. AAAI Conf Artif Intell 35:453
159. Wu Z et al (2020) Lite transformer with long-short range attention. In: International conference on learning representations
160. Mehta S et al (2021) DeLighT: deep and light-weight transformer. In: International conference on learning representations
161. Tat Y et al (2021) HyperGrid transformers: towards a single model for multiple tasks. In: International conference on learning representations
162. Yun S et al (2018) Graph transformer networks. In: International conference on learning representations
163. Jiang Y et al (2020) Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 33
164. You R et al (2019) AttentionXML: label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems, 32
165. Fan X et al (2020) Bayesian attention modules. Advances in Neural Information Processing Systems, 33
166. You R et al (2019) AttentionXML: label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems, 32
167. Fan X et al (2020) Bayesian attention modules. Advances in Neural Information Processing Systems, 33
168. Brunner G et al (2020) On identifiability in transformers. In: International conference on learning representations
169. Dosovitskiy A et al (2021) An image is worth 16 words:
170. Katharopoulos A et al (2020) Transformers are RNNs: fast and lightweight transformer via decomposed near-field and far-field attention. In: International conference on learning representations
171. Tang C et al (2020) Transformer-XH: multi-evidence reasoning with extra hop attention. In: International conference on learning representations
172. You R et al (2019) AttentionXML: label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems, 32
173. You R et al (2019) AttentionXML: label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems, 32
174. Lee J et al (2020) Efficient content-based sparse attention with routing transformers. Trans Assoc Comput Linguist 9:53–68
175. Child R et al (2019) Generating long sequences with sparse transformers, arXiv preprint arXiv:1904.10509
176. Correia G, Niculae V, Martins A (2019) Adaptively sparse transformers. In: Proceedings of the 2019 conference on...
empirical methods in natural language processing and the 9th international joint conference on natural language processing pp 2174–2184

178. Peng H et al (2021) Random feature attention. In: International conference on learning representations UK

179. Chen Y et al (2021) Skyformer: remodel self-attention with Gaussian kernel and Nyström method. Advances in Neural Information Processing Systems, 34

180. Zaheer M et al (2020) Big bird: transformers for longer sequences. Advances in Neural Information Processing Systems, 33

181. Huang CZ et al (2019) Music transformer: generating music with long-term structure. In: International conference on learning representations UK

182. Lu J et al (2021) SOFT: softmax-free transformer with linear complexity. Advances in Neural Information Processing Systems, 34

183. Pan Z et al (2021) Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/cvf international conference on computer vision, pp 377–386

184. Zhu C et al (2021) Long-short transformer: efficient transformers for language and vision. Advances in Neural Information Processing Systems, 34

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.