Analysis of transformer model applications*

M.I. Cabrera-Bermejo[1][0009−0001−4749−0882], M.J. Del Jesus[1][0000−0002−7891−3059], A.J. Rivera[1][0000−0002−1062−3127], D. Elizondo[2][0000−0002−7398−5870], F. Charte,[1][0000−0002−3083−8942], and M.D. Pérez-Godoy[1][0000−0002−6670−564X]

[1] Universidad de Jaén, Campus Las Lagunillas s/n, Jaén
(mbermejo,mjjesus,arivera,fcharte,lperez)@ujaen.es
[2] De Montfort University, Gateway House, Leicester
elizondo@dmu.ac.uk

Abstract. Since the emergence of the Transformer, many variations of the original architecture have been created. Revisions and taxonomies have appeared that group these models from different points of view. However, no review studies the tasks faced according to the type of data used. In this paper, the modifications applied to Transformers to work with different input data (text, image, video, etc.) and to solve disparate problems are analysed. Building on the foundations of existing taxonomies, this work proposes a new one that relates input data types to applications. The study shows open challenges and can serve as a guideline for the development of Transformer networks for specific applications with different types of data by observing development trends.

Keywords: Neural Networks · Transformers · Applications · Survey.

1 Introduction

Transformer [59] was introduced in 2017 as a sequence-to-sequence network based only on attention mechanisms, removing recurrence and convolutions. Its use has spread, and numerous models have been developed that strictly follow the original architecture, make minor modifications to it or severally change it. They initially emerged as networks to carry out machine translation with text sequences. However, currently there are models that solve different tasks with any input data. This has resulted in the emergence of studies and taxonomies that categorize Transformers from different points of view. Nevertheless, there is no review of these at application level or according to the type of data used.

In this paper a survey of Transformers, according to applications and the type of data these networks work with, is carried out. This aims to provide insights into common modifications for the use of different types of data and

* The research carried out in this study is part of the project "ToSmartEADs: Towards intelligent, explainable and precise extraction of knowledge in complex problems of Data Science" financed by the Ministry of Science, Innovation and Universities with code PID2019-107793GB-I00 / AEI / 10.13039 / 501100011033.

their respective applications. Not all AI model users are Transformer experts, so choosing the right one for the task at hand is not always easy. This is where this paper would be helpful. In addition, knowing the most frequently used architecture and components in each type of application (with their possibilities and limitations) can help in the development of new proposals to address research challenges in the area.

The module-level and architecture-level taxonomy proposed in [42] is used to group the modifications and components used. This makes possible to observe the adaptations applied in architectures that work with a particular data type or achieving a certain goal.

This paper is organized as follows: Section 2 introduces the original Transformer [59] and the taxonomy at the module and architecture level [42]. Section 3 describes the methodology followed and our review from the point of view of the applications and data types used in Transformer networks. Finally, section 4 contains our conclusions.

2 Background

The Transformer [59] is composed of encoders and decoders with a similar structure: Multi-Head Attention (MHA) and Feed Forward Networks (FFN) as main elements, with residual connections and normalization after them. Positional encoding is included at the bottom of encoder and decoder stacks to add information about the position of the tokens in the sequence. Moreover, the decoder contains a Masked MHA to prevent positions from attending to subsequent positions and a MHA over the output of the encoder stack.

Transformer networks have become one of the most widely used and powerful architectures, with many variations and modifications to the original definition emerging. Taxonomies and studies are beginning to appear that group all Transformers models from different points of view. Given that, our aim is to present a review of Transformers applications based on the type of data these networks work with, we analyse them considering the module-level and architecture-level categorizations presented in [42].

To perform the module-level taxonomies, vanilla Transformer is divided into four different modules: Attention, Positional Encoding (PE), Normalization and FFN. The most extensively studied is the attention-level categorization with the improvements on attention mechanism divided into Sparse attention (Sparse), Linearized attention, Query Prototype, Memory Compression (MC), Low-Rank self-attention, attention With Prior and Improved Multi-Head Mechanism. The positional information is encoded as Absolute Sinusoidal Position Encodings (Absolute) in [59]. This is the first of the classes within PE of [42], along with Relative Position Representations (Relative), Implicit Representations and Other Representations (Other). With regard to the classification of the normalization, a distinction is made between those works in which the placement of these layers is modified (Placement), those that substitutes the normalization formula (Substitutes) and those that removes it (Removed). The last module to catego-

rize is FFN. Three categories are grouped: those studies that explore different Activation Functions (AF), those that replace FFN with similar structures with many more parameters, and those that remove these layers (Removed). In the taxonomy of Attention and FFN we have included two additional classes: Original and Other, as many models use original components or major modifications that do not fit into the other classes.

The architecture-level taxonomy proposed in [42] studies Transformers that modify the original one beyond the modules, making higher-level modifications. It differentiates those studies that adapt the Transformer to be lightweight (Lightweight) in terms of model size, memory, computation time or Floating Point Operations Per Second. Architectures that strengthen cross-block connectivity are grouped in a separate category. Another class includes those that adapt computation time conditioned to inputs. Finally, there are the Transformers with Divide-and-Conquer Strategies, among which are the recurrent (Recurrent), where a cache memory is maintained to incorporate the history information, and the hierarchical (Hierarchical), which decomposes inputs hierarchically into elements of finer granularity. In addition, several studies have explored alternative architectures for Transformer (Alternative). In the Normalization and Architecture, the Original class has also been included, as many models use the one proposed by Vaswani [59].

3 Analysis of transformer model applications

This section reviews the different applications that have been addressed with Transformers. To this end, first the review methodology is presented and then the different categories that have been defined.

3.1 Methodology

This review of Transformers applications based on the type of data analyse proposals with a significant contribution between 2017 and 2023. For this purpose, we have gathered relevant articles related to Transformer networks from different leading bibliographic databases such as Scopus, Google Scholar, Elsevier, Springer, IEEE, ACM or arXiv. The documents were analysed taking into account the publication site (journal, congress, conference, or other), the authors, the year of publication, the database where they were published, the number of citations, the task solved and the contribution made.

Subsequently, and given the limited number of pages present in a conference article, the most relevant ones were selected given the characteristics above for inclusion in this study. These are categorized according to the type of data used and the application to be performed. Furthermore, every element of these models is classified according to module-level and architecture taxonomies described in [42].

In the following subsections, tables have been included following the same structure to contain all the information in a more visual way. These tables include

a column for the target application of the model (the acronyms used in this column are described at the beginning of each subsection). Then, a column is included for each Transformer module: Attention, PE, Normalization (Norm) and FFN. Finally, a column is included for the Architecture (Arch) and a column for the Reference (Ref) of the corresponding paper. In addition, in cases where the type of data used may change slightly (such as multimodal), a first column has been included detailing the type of data used. The categories used in the modules and the architecture are described in Section 2.

3.2 Text

Transformers emerged as an architecture for handling text data, with a focus on machine translation, and most Transformer networks work with text data as input. All these are shown in this category (Table 1) describing the jobs they perform, among which we found: *Translation*, *Generation* and *Prediction*. In addition, there are those that solve multiple with a single architecture (*Multitask*).

Table 1. Taxonomy of Textual Data Processing Transformers

Application	Attention	PE	Norm	FFN	Arch	Ref
Translation	Original	Absolute	Original	Original	Lightweight	[41]
Translation	Original	Absolute	Original	Original	Lightweight	[18]
Translation	Original	Absolute	Original	Original	Lightweight	[61]
Translation	Original	Absolute	Original	Original	Lightweight	[60]
Translation	Original	Absolute	Original	Original	Original	[4]
Multitask	Original	Absolute	Original	Original	Lightweight	[35]
Multitask	Original	Absolute	Original	Original	Original	[14]
Multitask	Original	Absolute	Original	Original	Original	[46]
Multitask	Linearized	Relative	Original	Original	Lightweight	[51]
Generation	Original	Absolute	Original	Original	Original	[65]
Generation	Original	Absolute	Original	Original	Original	[62]
Generation	Original	Absolute	Placement	Original	Original	[66]
Generation	Sparse	Absolute	Substitutes	Original	Original	[52]
Prediction	Other	Absolute	Original	Original	Original	[1]
Prediction	Original	Absolute	Original	Original	Original	[49]
Prediction	Sparse	Relative	Original	Original	Original	[64]

Most of the models that make machine translation try to improve on the vanilla Transformer. The one that is developed in [41] compresses the decoder sublayers into one for a higher degree of parallelism. In the case of [60], they propose to produce multiple successive words in parallel at each time step, keeping the autoregressive property and improving translation speed. One of the most popular works can be found in [18], where the architecture avoids this autoregressive property and produces its outputs in parallel. Furthermore, as a proposed improvement to the previous one in quality of decoder hidden representations is [61]. In contrast to the above that seek to improve the original, [4] proposes the incorporation of small adaptive layers to adapt the machine translation.

Many Transformer architectures are designed to face multiple tasks with text, an example is BERT [14], it has been used as a backbone for many studies like in [35, 46], which make minor modifications. In the former, memory consumption is reduced, while in the latter, clinical domain knowledge is integrated. Another

multitasking model is presented in [51], modifying the attention to be linear and using relative PE to improve the efficiency of the vanilla transformer.

With regard to the text generation task, the study in [66] shows a network that generates answers to questions with the incorporation of convolutions in the Vaswani encoder. The proposed in [52] learns dynamic patterns of sparse attention for language modelling (understanding this as a text generation task). Another example is the proposal in [62] of a generative dialogue system combining transfer learning and Transformers. In [65], they revisited triple extraction as a sequence generation job (which jointly extracts entities and relations).

About prediction task, in [49] an adaptation of BERT for rare diseases diagnosis is proposed. The proposal in [64] modifies the original architecture to perform Named Entity Recognition, using sparse attention and relative PE. The model described in [1] performs relation and event extraction tasks to test their multilingual transferability (understanding these as prediction), modifying attention and extracting syntactic distances.

As can be seen in the Table 1, most models that work with text use the original Transformer. In some studies minor modifications are made, especially with sparse attention, changes in normalization or the use of relative PE. Furthermore, most of those that perform machine translation use lightweight architectures, trying to improve the original, whose goal was sequences translation.

3.3 Image and Video

Table 2 lists all architectures included in this category based on Transformers for solving various tasks such as *Object Detection* (OD), *Image Generation*, *Image Captioning* and *Classification*. Also, there is a multitasking model (*Multitask*).

Table 2. Taxonomy of Image and Video Data Processing Transformers

Data	Application	Attention	PE	Norm	FFN	Arch	Ref
Image	OD	Sparse	Absolute	Original	Original	Original	[73]
Image	OD	Original	Absolute	Substitutes	AF (Leaky ReLU)	Original	[58]
Image	OD	Original	Absolute	Original	Original	Original	[2]
Image	OD	Original	Absolute	Original	Original	Original	[6]
Image	Generation	Other	Relative	Substitutes	AF (GELU)	Original	[33]
Image	Generation	Sparse	Absolute	Original	Original	Original	[48]
Image	Captioning	Original	Absolute	Original	Original	Original	[31]
Image	Captioning	Original	Abs / Rel	Original	Original	Original	[44]
Image	Captioning	Other	Absolute	Original	Original	Original	[11]
Image	Classification	Original	Absolute	Placement	AF (GELU)	Hierarchical	[16]
Image	Classification	Original	Other	Placement	AF (GELU)	Original	[23]
Image	Classification	Original	Absolute	Placement	Original	Original	[69]
Image	Classification	Other	Absolute	Placement	AF (GELU)	Hierarchical	[5]
Image	Classification	Other	Absolute	Placement	AF (GELU)	Hierarchical	[30]
Image	Classification	Original	Absolute	Original	Original	Original	[8]
Image	Multitask	Other	Absolute	Placement	AF (GELU)	Hierarchical	[43]
Video	Classification	Original	Absolute	Placement	AF (GELU)	Original	[68]
Video	Classification	Original	Absolute	Placement	Original	Original	[3]

For object detection, the proposal in [2] applies feature extraction before the Transformer encoder and a classifier after it. In the same way, [6] presents DETR (DEtection TRansformer) that uses a Transformer encoder-decoder with little changes, with a Convolutional Neural Network (CNN) ahead and a FFN

following it. Based on the latter, the model defined in [73] improves DETR with a deformable attention module. Finally, [58] faces place recognition and address as an OD task. It extracts lines for input images and uses a Transformer for line clustering. After that, they apply other Transformer for cluster description.

The model described in [33], that builds a GAN completely free of convolutions using only pure Transformer-based architecture with many changes and a grid self-attention, is included on image generation. In addition, [48] uses sparse attention and the Vaswani's Transformer for generate images. One of the mentioned in text generation ([52]) is also carrying out image generation.

With respect to image captioning, in [44] both grid and region features are used to achieve complementary information of them inside the same image. In [11] they incorporate a priori knowledge in the attention with memory vectors and a meshed connectivity between encoding and decoding modules. The proposal in [31] introduces region and global features into the attention.

We distinguish those works that perform classification. With images: [16] proposes a model with a Transformer encoder that previously patches the original input image and applies a linear projection; [69] uses a set of reduction and normal cells containing Transformers; and [8] uses a hybrid architecture between CNN and Transformer for image matching. There are also models that perform image segmentation as classification [23, 5, 30]. In [23], they treat volumetric images using Transformers, while the last two make use of hierarchical frame-level Swin Transformer [43] to create two different models. Concerning classification with video, [68] apply pre-processing and the Token Shift Transformer that include the Token Shift Module. Also, [3] uses a Transformer whose attention factorizes the spatial and temporal dimensions of the input video.

Finally, in [43] they propose a general-purpose backbone for computer vision and run experiments in image classification, object recognition and semantic segmentation. This architecture has a shifted window based self-attention and minor changes with respect to the original Transformer.

For image and video processing on Transformers, Table 2 shows that attention, normalization and FFN are the most modified elements. In the case of the FFN, the use of the GELU function as an activation function stands out. So many modifications on the original are needed to adapt these networks to the use of pixels from an image and frames from a video instead of text sequences.

3.4 Audio

Transformers using audio spectrograms as input and/or output data accomplish various applications such as *Speech Enhancement* (SE), *Speech Recognition* (SR), *Text-to-Speech* (TTS), *Speech Separation* (SS) and *Generation* (Table 3).

Speech recognition is one of the most studied tasks when working with audio. Three similar proposals are found in [15, 20, 47], which apply preprocessing to data before the Transformer. Conformer [20] stands out by including CNN before the FFN. A different approach is [10], which faces speech recognition in real time with a Transducer network whose encoder contains a modified Conformer.

Table 3. Taxonomy of Audio Processing Transformers

Application	Attention	PE	Norm	FFN	Arch	Ref
Generation	MC	Relative	Original	Original	Lightweight	[27]
SS	Original	Absolute	Placement	Original	Recurrent	[55]
TTS	Original	Absolute	Original	Original	Original	[38]
TTS	Other	Absolute	Placement	Original	Lightweight	[29]
TTS	Original	Removed	Original	AF (non-linear)	Original	[70]
SR	Original	Other	Original	Original	Original	[47]
SR	Original	Relative	Placement	AF (swish)	Original	[20]
SR	Other	Absolute	Placement	Original	Original	[15]
SR	Original	Relative	Original	AF (non-linear)	Original	[10]
SE	Original	Removed	Original	Removed	Alternative	[67]
SE	Other	Removed	Original	Original	Alternative	[34]
SE	Original	Removed	Original	Other	Original	[13]
SE	Other	Removed	Placement	AF (GELU)	Lightweight	[32]

Text-to-speech is solved in [29] by using Locality-Sensitive Hashing Attention and Reversible Residual Network to reduce the memory used. Local LSTM before attention to encode PE locally, directly and differently to original Transformer is used in [70]. In [38] the vanilla Transformer is adapted to the specific task by applying pre- and post-processing to the data.

In speech Enhancement, the proposal in [67] replaces FFN with 1D CNN and uses local LSTM as in [70] to remove original PE. In [34] they use self-attention with Gaussian-weighted and Short-Time Fourier Transform. The architecture defined in [13] combines Intra and Inter Transformers, as well as other elements, eliminates positional encoding and modifies FFN to use GRU, ReLU and linear transformations. There is also a model [32] that reduces the computational cost for this task by taking consecutive frames and treating them as a local window that computes attention by using hierarchical frame-level Swin Transformer [43] layers with an attention mechanism adapted to these frame windows.

In [55] they use a structure with two recurrently connected embedded Transformers for speech separation. An architecture that reduces the intermediate memory requirement by modifying attention and PE is described in [27]. It allows the generation of one-minute musical compositions.

Table 3 shows that PE is the most modified element in Transformers that use audio as input data, since the model must be adapted to work with spectrograms. The most common is use relative PE or remove it. We also find modifications in the placement of normalization, in the FFN activation function and, as the most notable change, the use of adapted attention to work with audio signals.

3.5 Tabular

All proposals analysed use structured or tabular data to perform the same overall task, *Prediction*, but in different domains (Table 4).

Two of these works perform molecular prediction [45, 7]. One of them run experiments on a large collection of datasets that represent typical tasks in molecule modelling, grouped like prediction: regression, binary classification, multiclass classification, etc. The other one performs molecular property prediction. In [63] a model that performs property prediction with polymers is defined.

Table 4. Taxonomy of Structured Data Processing Transformers

Application	Attention	PE	Norm	FFN	Arch	Ref
Prediction	Sparse	Absolute	Removed	Original	Original	[40]
Prediction	Original	Absolute	Original	Original	Original	[53]
Prediction	Other	Removed	Placement	Original	Original	[45]
Prediction	Original	Absolute	Original	Original	Original	[7]
Prediction	Sparse	Absolute	Original	Original	Lightweight	[71]
Prediction	Original	Other	Original	Other	Original	[36]
Prediction	Original	Absolute	Original	Original	Original	[25]
Prediction	Original	Relative	Original	Original	Original	[63]

Other works [40, 53, 71] carry out times series prediction, including elements for capturing temporal information. The study in [53] performs this job in medical field by adding dense interpolation and masked self-attention mechanism. Informer [71] is a lightweight Transformer for Long Sequence Time-Series Forecasting that uses its own sparse attention. In [40] a graph Transformer that captures spatial and time dependent data with graph structure for forecasting and prediction by applying sparsity to the whole architecture is defined.

Within prediction, some models make classification. The proposal in [36] suggests a new architecture with Gaussian range encoding and two-tower structure that captures time and channel stream for Human Activity Recognition. In [25] they create a Prior-Data Fitted Network with Transformers to perform supervised classification for small datasets in less than a second.

For Transformers working with tabular data, changes in attention to adapt it to this type of data are the most common, as shown in Table 4. However, modifications to the other elements are also made in some cases.

3.6 Multimodal

Most of the proposals working with several types of data (Table 5) consider linguistic and visual data. In addition, multitasking is one of the most studied applications of multimodal data.

Table 5. Taxonomy of Multimodal Data Processing Transformers

Data	Application	Attention	PE	Norm	FFN	Arch	Ref
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[37]
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[17]
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[57]
Text-Image	Multitask	Original	Abs / Rel	Original	Original	Original	[50]
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[39]
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[54]
Text-Image	Multitask	Original	Absolute	Original	Original	Original	[12]
Text-Image	Classification	Original	Absolute	Original	Original	Original	[72]
Text-Image	Classification	Original	Other	Original	Original	Original	[24]
Text-Image	Generation	Original	Absolute	Original	Original	Original	[26]
Text-Image	Generation	Original	Absolute	Original	Original	Original	[19]
Text-Video	VC	Original	Absolute	Original	Original	Original	[28]
Text-Video	Multitask	Original	Absolute	Original	Original	Original	[56]
Text-Audio	Multitask	Original	Absolute	Original	Original	Original	[21]
Text-Audio	Classification	Sparse	Absolute	Original	Original	Original	[9]
Text-Image-Audio	Classification	Original	Absolute	Original	Original	Original	[22]

Most studies that use a stream of linguistic and visual data use text and images. In many studies like [37, 17, 57, 50, 39, 54], models are multitasking. Most of them slightly modify a BERT architecture [14], except for [39] which employs Vaswani's original Transformer [59]. In [12] they introduce a new task of Edited Media Understanding that consists in answering questions on image manipulation by multitasking, with classification and generation of responses. Other works carry out only one task, like in [26, 19] where they generate responses to questions asked about images and the text of the images. Other case is [24] that performs different types of classification on user interfaces. Classification is carried out in [72] for the skin lesion diagnosis using self-attention and guided-attention.

There are also articles dealing with linguistic and visual data flows through text and video: using a BERT Large [56] for multitasking or two stream encoder, cross-modal attention and a text only decoder for Video Captioning (VC) [28].

The processing of audio and text signals together incorporates the study described in [21], whose model conducts Machine Translation and Speech Translation with a shared semantic memory network between encoder and decoder of the original Transformer. In [9] a model for emotion recognition using cascaded cross-attention block to fuse text and audio modalities is proposed.

A Transformer that works with language, acoustic, and vision features is defined in [22], encoding all features separately and using a bimodal cross-attention layer to exchange multimodal information for predicting whether the input is humorous or not.

The most common in architectures dealing with multimodal data is to work with different Transformers for the different types of data, and then to unify them or pre-process the data in order to feed them together into one Transformer. As shown in Table 5, except some modifications made to the PE, vanilla Transformer is widely used.

4 Conclusion

In this survey, we review Transformers solving different tasks and group them according to the type of data used as input. Most of the existing works use text sequences to resolve multiple tasks, but Transformers are increasingly being applied to image, video, audio, multimodal or tabular data processing, with structured data being the least studied.

This review shows that text-based models typically use the original architecture or BERT [14] without changes. Many improve the vanilla Transformer in memory, computation time, etc. Those working with visual data such as image or video commonly modify the attention to adapt to working with pixels (spatial information) or with frames (spatio-temporal information). Furthermore, these also modify normalization and often use GELU as an activation function in FFN. For those working with spectrograms and audio signals, it is most usual to change the PE to encode the time and frequency information of this type of data and the attention to attend to it. In addition, they also make changes to the normalization and the FFN activation function. Transformers working with

tabular data mostly transform the attention to deal with this data. A unified semantic space is used to work with multimodal data. Some of these models apply one Transformer for each data type and unify the outputs into another. Other works unify input data first and feed it into a Transformer. Except for certain modifications, the vanilla Transformer or BERT is commonly used.

References

1. Ahmad, W.U., et al.: Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. Proc. of AAAI **35**(14), 12462–12470 (2021)
2. Alamri, F., et al.: Transformer-encoder detector module: Using context to improve robustness to adversarial attacks on object detection. Proc. of ICPR pp. 9577–9584 (2021)
3. Arnab, A., et al.: Vivit: A video vision transformer. Proc. of ICCV pp. 6836–6846 (2021)
4. Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation. Proc. of EMNLP IJCNLP pp. 1538–1548 (2019)
5. Cao, H., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. Proc. of ECCV pp. 205–218 (2023)
6. Carion, N., et al.: End-to-end object detection with transformers. Proc. of ECCV pp. 213–229 (2020)
7. Chen, B., et al.: Path-augmented graph transformer network. arXiv:1905.12712 (2019)
8. Chen, J., et al.: Shape-former: Bridging cnn and transformer via shapeconv for multimodal image matching. Information Fusion **91**, 445–457 (2023)
9. Chen, W., et al.: Key-sparse transformer for multimodal speech emotion recognition. Proc. of ICASSP pp. 6897–6901 (2022)
10. Chen, X., et al.: Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. Proc. of IEEE ICASSP pp. 5904–5908 (2021)
11. Cornia, M., et al.: Meshed-memory transformer for image captioning. Proc. of CVPR pp. 10575–10584 (2020)
12. Da, J., et al.: Edited Media Understanding Frames: Reasoning About the Intent and Implications of Visual Misinformation. Proc. of ACL IJCNLP pp. 2026–2039 (2020)
13. Dang, F., et al.: Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. Proc. of ICASSP pp. 6857–6861 (2022)
14. Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. of NAACL pp. 4171–4186 (2019)
15. Dong, L., et al.: Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. Proc. of IEEE ICASSP pp. 5884–5888 (2018)
16. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
17. Gao, D., et al.: Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. Proc. of ACM SIGIR pp. 2251–2260 (2020)
18. Gu, J., et al.: Non-autoregressive neural machine translation. Proc. of ICLR (2018)
19. Gui, L., et al.: KAT: A knowledge augmented transformer for vision-and-language. Proc. of NAACL pp. 956–968 (2022)
20. Gulati, A., et al.: Conformer: Convolution-augmented Transformer for Speech Recognition. Proc. of Interspeech pp. 5036–5040 (2020)

21. Han, C., et al.: Learning shared semantic space for speech-to-text translation. Proc. of ACL IJCNLP pp. 2214–2225 (2021)
22. Hasan, M.K., et al.: Humor knowledge enriched transformer for understanding multimodal humor. Proc. of AAAI **14B**, 12972–12980 (2021)
23. Hatamizadeh, A., et al.: Unetr: Transformers for 3d medical image segmentation. Proc. of IEEE/CVF WACV pp. 1748–1758 (2022)
24. He, Z., et al.: Actionbert: Leveraging user actions for semantic understanding of user interfaces. Proc. of AAAI **7**, 5931–5938 (2021)
25. Hollmann, N., et al.: Tabpfn: A transformer that solves small tabular classification problems in a second. arXiv:2207.01848 (2022)
26. Hu, R., et al.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. Proc. of CVPR pp. 9989–9999 (2020)
27. Huang, C.Z.A., et al.: Music transformer. arXiv:1809.04281 (2018)
28. Huang, G., et al.: Multimodal pretraining for dense video captioning. Proc. of AACL pp. 470–490 (2020)
29. Ihm, H.R., et al.: Reformer-TTS: Neural Speech Synthesis with Reformer Network. Proc. of Interspeech pp. 2012–2016 (2020)
30. Iqbal, A., Sharif, M.: Bts-st: Swin transformer network for segmentation and classification of multimodality breast cancer images. KBS **267**, 110393 (2023)
31. Ji, J., et al.: Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. Proc. of AAAI **35**(2), 1655–1663 (2021)
32. Jiang, W., et al.: Low complexity speech enhancement network based on frame-level swin transformer. Electronics **12**(6) (2023)
33. Jiang, Y., et al.: Transgan: Two pure transformers can make one strong gan, and that can scale up. Proc. of NIPS **34**, 14745–14758 (2021)
34. Kim, J., et al.: T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. Proc. of IEEE ICASSP pp. 6649–6653 (2020)
35. Lan, Z., et al.: Albert: A lite bert for self-supervised learning of language representations. Proc. of ICLR pp. 344–350 (2020)
36. Li, B., et al.: Two-stream convolution augmented transformer for human activity recognition. Proc. of AAAI **35**(1), 286–293 (2021)
37. Li, L.H., et al.: Visualbert: A simple and performant baseline for vision and language. arXiv:1908.03557 (2019)
38. Li, N., et al.: Neural speech synthesis with transformer network. Proc. of AAAI **33**, 6706–6713 (2019)
39. Li, W., et al.: UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. Proc. of ACL IJCNLP pp. 2592–2607 (2021)
40. Li, Y., Moura, J.M.F.: Forecaster: A graph transformer for forecasting spatial and time-dependent data. Frontiers in Artificial Intelligence and Applications **325**, 1293–1300 (2020)
41. Li, Y., et al.: An efficient transformer decoder with compressed sub-layers. Proc. of AAAI **35**(15), 13315–13323 (2021)
42. Lin, T., et al.: A survey of transformers. AI Open **3**, 111–132 (2022)
43. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. Proc. of ICCV pp. 10012–10022 (2021)
44. Luo, Y., et al.: Dual-level collaborative transformer for image captioning. Proc. of AAAI **35**(3), 2286–2293 (2021)
45. Maziarka, L., et al.: Molecule attention transformer. arXiv:2002.08264 (2020)
46. Michalopoulos, G., et al.: UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. Proc. of NAACL pp. 1744–1753 (2021)

47. Mohamed, A., et al.: Transformers with convolutional context for asr. arXiv:1904.11660 (2019)
48. Parmar, N., et al.: Image transformer. Proc. of ICML **80**, 4055–4064 (2018)
49. Prakash, P., et al.: Rarebert: Transformer architecture for rare disease patient identification using administrative claims. Proc. of AAAI **35**(1), 453–460 (2021)
50. Qi, D., et al.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv:2001.07966 (2020)
51. Qin, Z., et al.: cosformer: Rethinking softmax in attention. arXiv:2202.08791 (2022)
52. Roy, A., et al.: Efficient content-based sparse attention with routing transformers. TACL **9**, 53–68 (2021)
53. Song, H., et al.: Attend and diagnose: Clinical time series analysis using attention models. Proc. of AAAI pp. 4091–4098 (2018)
54. Su, W., et al.: VL-BERT: pre-training of generic visual-linguistic representations. arXiv:1908.08530 (2019)
55. Subakan, C., et al.: Attention is all you need in speech separation. Proc. of IEEE ICASSP pp. 21–25 (2021)
56. Sun, C., et al.: Videobert: A joint model for video and language representation learning. Proc. of ICCV pp. 7463–7472 (2019)
57. Sun, L., et al.: Rpbert: A text-image relation propagation-based bert model for multimodal ner. Proc. of AAAI **15**, 13860–13868 (2021)
58. Taubner, F., et al.: LCD - Line Clustering and Description for Place Recognition. Proc. of 3DV pp. 908–917 (2020)
59. Vaswani, A., et al.: Attention is all you need. Proc. of NIPS **30**, 5999–6009 (2017)
60. Wang, C., et al.: Semi-autoregressive neural machine translation. Proc. of EMNLP pp. 479–488 (2018)
61. Wang, Y., et al.: Non-autoregressive machine translation with auxiliary regularization. Proc. of AAAI pp. 5377–5384 (2019)
62. Wolf, T., et al.: Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv:1901.08149 (2019)
63. Xu, C., et al.: Transpolymer: a transformer-based language model for polymer property predictions. Npj Comput. Mater. **9**, 1–14 (2023)
64. Yan, H., et al.: Tener: Adapting transformer encoder for named entity recognition. arXiv:1911.04474 (2019)
65. Ye, H., et al.: Contrastive triple extraction with generative transformer. Proc. of AAAI **35**(16), 14257–14265 (2021)
66. Yu, A.W., et al.: Fast and accurate reading comprehension by combining self-attention and convolution. Proc. of ICLR (2018)
67. Yu, W., et al.: Setransformer: speech enhancement transformer. Cognitive Computation **14**, 1152–1158 (2022)
68. Zhang, H., et al.: Token shift transformer for video classification. Proc. of ACM MM p. 917–925 (2021)
69. Zhang, Q., et al.: Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. IJCV **131**, 1141–1162 (2023)
70. Zheng, Y., et al.: Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer. Proc. of IEEE ICASSP pp. 6734–6738 (2020)
71. Zhou, H., et al.: Informer: Beyond efficient transformer for long sequence time-series forecasting. Proc. of AAAI **35**, 11106–11115 (2021)
72. Zhou, L., Luo, Y.: Deep features fusion with mutual attention transformer for skin lesion diagnosis. Proc. of ICIP pp. 3797–3801 (2021)
73. Zhu, X., et al.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. Proc. of ICLR pp. 1–16 (2021)