Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(3), 219-225
Published Online August 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.123025

Hans汉斯

基于Transformer的自然语言处理模型综述

赖鸣姝

北京印刷学院，信息工程学院，北京

收稿日期：2023年6月25日；录用日期：2023年8月1日；发布日期：2023年8月9日

摘　要

自然语言处理是计算机科学中深度学习领域的一个分支，旨在使计算机能够理解、解析或生成人类语言(包括文字、音频等)。本文主要介绍了自然语言处理(Natural Language Processing, NLP)中基于Transformer结构所衍生出的多种类型的模型。近年，随着深度学习技术的快速发展，自然语言处理模型的性能也得到了极大的提升，更多的自然语言处理任务得到了更好的解决。这些进展主要得益于神经网络模型的不断发展。本文讲解了当前最为流行的基于Transformer的几类自然语言处理模型，包括BERT (Bidirectional Encoder Representations from Transformers)系列、GPT (Generative Pre-trained Transformer)系列和T5系列等。主要介绍了上述系列的模型各自的发展变化以及其在模型结构，设计思路等方面的区别与联系。同时，对于自然语言处理领域未来的发展方向进行了展望。

关键词

人工智能，深度学习，自然语言处理

A Survey of Transformer-Based Natural Language Processing Models

Mingshu Lai

Department of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Jun. 25th, 2023; accepted: Aug. 1st, 2023; published: Aug. 9th, 2023

Abstract

Natural language processing is a subfield of deep learning in computer science that aims to enable computers to understand, parse, or generate human language (text, audio, etc.). This paper mainly introduces various types of models derived from the Transformer structure in Natural Language Processing (NLP). In recent years, with the rapid development of deep learning technology, the

文章引用: 赖鸣姝. 基于 Transformer 的自然语言处理模型综述[J]. 人工智能与机器人研究, 2023, 12(3): 219-225.
DOI: 10.12677/airr.2023.123025

performance of natural language processing models has also been greatly improved, and more natural language processing tasks have been better solved. These advances are mainly due to the continuous development of neural network models. This article explains the most popular Transformer-based natural language processing models. These include BERT (Bidirectional Encoder Representations from Transformers) family, GPT (Generative Pre-trained Transformer) family, the T5 family, etc. This paper mainly introduces the development and changes of the above series of models, as well as their differences and connections in model structure, design ideas and other aspects. At the same time, the future development direction of natural language processing is prospected.

Keywords

Artificial Intelligence, Deep Learning, Natural Language Processing

1. 引言

自然语言处理(Natural Language Processing, NLP)这一领域的研究涉及到多个领域的知识，旨在让计算机能够像人类一样理解和处理文本信息，从而实现人机交互、信息检索、自然语言生成等多种应用[1]。近年来，随着深度学习技术的快速发展和大规模语料库的建立，对于自然语言任务的性能得到了很大的提升。这些进展主要得益于深度学习领域的不断发展，如循环神经网络、卷积神经网络和注意力机制等的提出和应用。

截至目前，自然语言处理领域仍然存在许多的问题亟待解决。例如，文本分类、情感分析、机器翻译、命名实体识别等。在这些任务中，计算机需要具备对文本的理解和处理的能力，同时还需要考虑到语言的多义性、歧义性、语境依赖性等因素。为了解决这些问题，学术界已有大量的相关工作，包括统计模型、规则模型、深度学习模型等。

综上所述，自然语言处理是一个不断发展和壮大的领域，它涉及到多个学科的交叉，具有广泛的应用前景。未来，随着技术的不断进步和发展，自然语言处理将会在人机交互、知识管理、智能搜索和自然语言生成等方面发挥越来越重要的作用。

2. 近年进展

2.1. Transformer

2017 年，Vaswani 等人提出了 Transformer [2]，这个模型起初针对自然语言处理中序列到序列(Seq2Seq)中的机器翻译任务，采用了处理序列任务中常见的编码器－解码器架构。后期也有许多研究表明并将其应用到许多自然语言处理任务当中。和之前的自然语言处理任务不同，Transformer 不再使用循环神经层(recurrent neural layer)，而是仅依赖注意力机制进行信息的编码和解码操作，相比之前的模型也有着较高的并行度，缩短了训练时长。为了解决循环神经网络中存在的遗忘前序信息和必须串行这两个问题，Transformer 采用了自注意力机制(self-attention)中的多头注意力机制(muti-head attention)和编解码器结构(encoder-decoder)。之所以采用多头注意力机制，是因为其中的多个注意力头(muti-head)能够达到和常用的卷积操作相类似的输出多个通道的效果。

目前，学界已经存在许多的研究成果表明，Transformer 这一架构可以在图像、视频、音频等多个领域均达到优良的性能，因其是对不同类型的数据的信息提取过程更加的一般化。但这也就造成了其提取信息的能力不够精准，需要大量的数据进行训练才能够挖掘出模型自身最优的性能。在自然语言处理等领域中，许多的模型也都是基于 Transformer 中编码器和解码器的结构进行设计的。

2.2. BERTs

最初的 Transformer 模型被广泛应用于序列到序列的机器翻译任务，但随着研究的深入，学者们开始探索其在其他自然语言处理任务中的应用。其中最著名的变体是 BERT (Bidirectional Encoder Representations from Transformers) [3]，它是一种预训练语言模型，可以十分便捷地用于许多其他自然语言处理任务当中。

BERT 针对的任务是更加广义的语言理解任务，是一个双向的网络。也就是说，和 GPT (Generative Pre-trained Transformer) [4]仅通过左侧的信息来预测之后的信息(在后文进行详细介绍)不同的是，BERT 利用了左右两侧的信息进行目标位置的相关预测。

自然语言处理领域主要包括两种类型的任务，第一个类型是句子层面的任务，例如，文本情感分类。第二个类型是词语级别的任务，例如，实体识别。第二类任务中需要一些细粒度的信息。其下游任务迁移一般有两种方式：一种是基于特征的迁移方式，另一种是基于微调的迁移方式。基于特征的迁移方法就是针对每一个下游任务设计一个新的网络，基于微调的方式就是使用预训练模型，再使用下游任务的数据集对模型进行微调。BERT 主要使用一种带掩码的语言模型，这个模型会随机遮盖住句子的一些词，损失函数负责预测被遮盖住的词。这样就需要参考被预测位置两侧的信息，形成了双向的 Transformer 预训练模型。在 BERT 的模型预训练过程中，在未标记的数据集上进行无监督训练，向下游任务迁移时，在带标记的下游任务数据集上做微调，这一做法原本在深度学习的计算机视觉领域十分常见，但是相比于已经存在大规模含标注数据集且单个样本已经蕴含了丰富信息的计算机视觉领域，自然语言处理领域中的模型很难达到这一效果。因此，BERT 选择在无监督数据集上进行预训练。BERT 的预训练有两个步骤：预训练和微调，预训练使用没有标签的数据集，进行无监督训练；微调步骤使用的是有标注的数据。进行有监督训练。BERT 更加适合处理文本序列的数据而不是单个的句子。由于 BERT 在训练过程中能够十分方便的适应到下游任务中，因此解决下游任务问题的重点就转变为了对于下游任务的转换和数据的组织，而非模型的设计工作，这大大方便了下游任务实际落地的过程。

在 BERT 模型发布之后，谷歌公司也在不断地对其进行改进和升级。2019 年，提出了 BERT 的多语言版本 MBERT [5]，通过在多种语言的大规模语料库上进行预训练，可以处理 104 种语言。这个版本的发布，在全球范围内的应用得到了进一步的扩展。RoBERTa [6]利用更多的数据、更大的训练批次和更长的训练时间，去掉下一句预测目标，较长序列的训练以及动态掩码(Dynamic Masking)机制，来优化模型性能。ELECTRA [7]引入了一种类似于生成对抗网络(Generative Adversarial Network, GAN)的训练方式。首先，使用一个参数量较少的模型作为生成器，针对随机遮盖的词元进行预测，然后再将重新修复后的句子交给判别器进行判断，主要判断输入的句子当中每个单词是否经过生成器的替换。而 BERT 的训练过程则是先对一部分词元进行随机遮盖，再利用上下文信息预测被遮盖的词，预测的样本空间是整个词表。ELECTRA 进一步提升了自然语言处理任务上的性能。

在训练过程中对于数据遮盖方式的研究，也是 BERT 系列中的一个研究重点。BERT 使用了较为简单的随机遮盖词元的方式对于数据进行处理，但这种方式对于语料的信息有所损失。BERT-WWM [8]中所使用的方式为将被遮盖词元所属单词的其他词元也进行遮盖。随后提出的 ERNIE [9]则引入命名实体(Named Entity)这类外部知识，对实体单元进行遮盖。SpanBERT [10]中根据几何分布，先随机选择一段

语句的长度，之后再根据均匀分布随机选择这一段的起始位置，最后按照长度遮盖，达到了更好的效果。

另外，学界许多学者针对于使用更少的参数达到和 BERT 相似的性能并加快训练速度这一目的进行研究，ALBERT [11]、Q8BERT [12]、DistilBERT [13]和 TinyBERT [14]等相关工作先后被提出。四个工作分别采用了结构优化、量化、知识蒸馏等方式进行模型的压缩，使其能够在更多的场景和设备中进行应用和部署。

总体而言，BERT 模型基于 Transformer 中的编码器结构，有众多优化方向，主要有前文详述的模型压缩和微调与任务优化。BERT 系列模型的优点包括：可以处理各种自然语言处理任务，可以使用多种语言进行预训练和微调的模式，使得模型可以更好地应用到下游任务当中。由于使用了 Transformer 架构，BERT 系列模型在处理较长文本时具有优势。然而，BERT 系列模型的缺点也很明显，首先，BERT 需要大量的计算资源和时间进行训练和微调。其次，BERT 并不擅长处理所有的自然语言处理任务，比如生成类任务。再次，因为 BERT 是基于词而不是句子级别的编码，且 BERT 的段嵌入(segment embedding)只能包含两种句子类型，没有办法直接用于输入存在多个句子的摘要任务当中。最后，BERT 不擅长处理一些专业领域(如医疗、金融)用词或中文偏僻词相关的问题。

2.3. GPTs

由 Transformer 还发展出的另一个系列的模型，GPT 系列。其目标任务是使得模型可以变为一个可以解决所有自然语言处理问题的通用型模型，相比于 Transformer 和 BERT 系列针对的文本分类和机器翻译任务而言，是一个更加难以达到的目标，因此其在单个任务方面的性能相比 BERT 而言也就有所下降。和 BERT 系列使用 Transformer 中的编码器部分进行堆叠不同，GPT 系列的模型使用了 Transformer 的解码器进行堆叠。和 BERT 的训练过程不同，GPT [4]的微调过程把带标签的数据全部送到预训练模型中，使用最后一个输入对应的输出的特征，乘以对应的权重。另外，在微调过程中还需要考虑无监督训练过程的标准语言模型的损失计算方式的设计，以求达到更好的效果。

2019 年，Alec Radford 等人提出的 GPT-2 [15]所面临的一个问题是：当数据集数量和模型参数量都进行增大的情况下，GPT 的性能并没有优于 BERT。因此，GPT-2 为了提高模型的泛化性，使得预训练模型在迁移到下游任务的时候完全不用进行微调，提出了一个基于零样本学习的方法，使用自然语言的方式来描述问题，作为提示(Prompt)，也就是零样本学习的核心思想，也是提示工程(Prompt project)所需要做的工作。

2020 年，Tom B. Brown 等人提出的 GPT-3 [16]改为采用小样本学习的思路，弱化了极致的零样本学习，尝试解决 GPT2 中存在的有效性低的问题。但是，GPT-3 作为一个拥有大量参数(175 亿)的非稀疏模型在应用到下游任务时选择了不更新梯度和微调。GPT-3 虽然已经在很多任务中达到了很好的效果，但其仍然存在一定的限制。首先，对于长文本的生成有一定的限制；其次，这一系列的论文由于仅使用 Transformer 的解码器部分导致其对于样本只能“从左往右”看，并不能像 BERT 一样总揽全局的信息；最后，每一个词元(token)在预测下一个词元的时候发挥的效率均等，导致了学习效率较低，样本有效性不足，导致可解释性较差且模型的训练成本偏高。

2022 年 3 月，Long Ouyang 等人提出了 InstructGPT [17]模型，利用了基于人类反馈的强化学习方法(Reinforcement Learning from Human Feedback, RLHF)对 GPT-3 进行微调，使得该模型的输出更加符合人类偏好。在 InstrcutGPT 中，输入序列是一段自然语言文本和一条给定的程序指令，模型的任务是生成与给定指令相对应的程序代码。InstrcutGPT 的预训练过程主要分为两个阶段：第一个阶段是基于代码库的预训练，第二个阶段是基于程序指令的预训练。在基于代码库的预训练阶段，模型主要学习代码库中的代码结构和语法规则；在基于程序指令的预训练阶段，模型主要学习如何将自然语言指令转换为程序代

码。与 GPT-2 [15]相比，InstrcutGPT 在预训练过程中引入了一些技巧，例如，基于代码库的语言模型微调、自适应词汇表等，从而提高了模型在程序指令生成任务上的性能。

GPT 系列中还有几种针对不同使用场景的文本生成模型：WebGPT [18]、ChatGPT [19]和 Toolformer [20]。WebGPT 是专为处理 Web 文本而设计的预训练模型，通过辨别关键词进行搜索并抽取相关段落以生成回答。与 GPT-3 相比，WebGPT 更适应 Web 文本的多样性和复杂性。ChatGPT 则主要用于对话场景，其可模拟人类语言习惯生成符合语法规则和语境的自然语言文本，已在多个对话场景中得到了应用。和 InstructGPT 之间，除了使用场景有所不同之外，训练数据也存在差异。InstructGPT 主要使用包含指令或约束条件的数据来微调模型，而 ChatGPT 则使用了大规模的对话数据来进行训练。Toolformer 以自监督的方式微调语言模型，让模型学会如何自动调用接口。可在自动化写作、智能客服、智能文档生成等方面提高文本生成的效率和质量。但 Toolformer 并不适合处理较长的文本。

综上所述，GPT 系列主要基于 Transformer 中的解码器结构进行堆叠，旨在成为一个能够解决所有自然语言处理任务的通用型的模型。其中，主要变化的部分在于对提示样本的使用方式中探索了多种方式，产生了适用于不同场景下的 GPT 模型。

2.4. T5s

Google 在 2019 年提出了一种基于 Transformer 结构的生成式预训练语言模型 T5 (Text-to-Text Transfer Transformer) [21]，与之前提到的 BERT 系列和 GPT 系列不同的是，T5 系列同时使用了 Transformer 总的编码器和解码器结构。T5 与 Transformer 之间的最大区别在于，T5 不仅包含了 Transformer 编码器和解码器，还包括了一种称为任务规范(task specification)的方法，该方法将各种自然语言处理任务转化为文本的输入及输出的任务。例如，在问答任务中，任务规范将问题和答案拼接成一个文本序列，输入到 T5 模型中进行处理。这使得 T5 可以在多种自然语言处理任务中进行端到端的学习，而无需针对每个任务都进行单独的微调。此外，T5 还使用了大规模的无监督学习来预训练模型，使用大小约为 800 GB 的 C4 (Colossal Clean Crawled Corpus)数据集进行训练。而在有监督训练部分，则使用自然语言处理领域的公开数据集，两个阶段均统一转化为序列到序列的任务来训练。

2021 年，Linting Xue 等人提出了能够支持多种语言的 mT5 [22]，其继承了 T5 所有的优点的同时，在 T5 的基础上进行了一些模型结构的细微调整，在多语言版本的 C4 数据集——mC4 上进行训练，得到了支持多种语言的 mT5 模型。同年，Jianmo Ni,等人提出了 sentence-T5 [23]，探索了 T5 模型在文本表示任务方面的能力，作者发现编码器部分输出的特征做平均池化后，对于模型的效果有所提升。

该系列模型的提出给自然语言处理领域增添了一种新的范式，且提升了各个自然语言处理任务的性能。对于各个任务的处理有了更低的使用门槛，但是其仍需要耗费相比于其他系列的模型而言更加巨大的资源和时间，

2.5. XLs

2019 年，Zihang Dai 等人提出 Transformer XL [24]开创新的自然语言处理模式，主要针对的是模型能够处理的序列长度固定且不能处理过长文本的问题。同时，加快了对于长序列的处理速度。与 Transformer 不同的是，Transformer XL 在编码器和解码器中均使用了相对位置编码，从而在处理长序列时能够更好地捕捉语句之间的关系。此外，Transformer XL 还使用了一种称为循环机制(recurrence mechanism)的方法，使得模型可以在多个时间段中重复使用前一段的信息，从而进一步提高了模型的效率和性能。并且针对循环机制使用了相对位置编码(relative position encoding)。

同年，Zhilin Yang 等人提出 XLNet [25]，其针对于 BERT 中存在的忽略了被遮盖词与其他词之间的

关系以及训练和测试过程面对不同的数据这两点缺陷进行解决，提出一种基于 Transformer 结构的自回归模型。XLNet 使用了两种不同的训练方式，一种是单向语言模型训练，另一种是双向语言模型训练。在单向语言模型训练中，模型的训练数据是从左到右的，而在双向语言模型训练中，模型同时考虑了左右两侧的信息。这种训练方式使得 XLNet 能够更好地理解文本中的语境和关系，从而在多项自然语言处理任务中取得了最好的效果。

3. 结论

综上，Transformer 结构作为一种革命性的神经网络结构，对于自然语言处理领域的发展起到了重要的作用，后续的大量模型和结构的设计都是在 Transformer 的基础上进行的。本文主要总结了由 Transformer 衍生出的不同类型的模型各自的发展变化以及其在模型结构，设计思路等方面的区别与联系。在基于 Transformer 的模型中，BERT 系列模型基于 Transformer 的编码器部分主要用于语言理解类任务，GPT 系列模型基于 Transformer 中的解码器部分用于语义生成类任务，T5 系列模型同时使用了 Transformer 中的编码器和解码器结构，更适合应用于多种自然语言处理任务，XL 系列则在 Transformer 的基础上重新使用了自回归编码方式，主要解决了其他模型难以处理长序列的问题，适合处理自然语言处理中的语言建模、文本分类等一些复杂任务。这些模型的结构在本质上都是在 Transformer 中的编码器－解码器结构上的一些组合和变化。主要的变化和创新更加侧重对于数据的处理和训练的方式有着更多的变化和调整。

本文中给出的模型虽然已经在大多数自然语言处理任务中获取了优良的性能，但是仍然存在一些问题。首先，大多数模型并没有给出完整清晰的模型选择过程，或者只是公开了使用的配置，没有提及搜索策略和测试值。众所周知，这些体系结构可能对超参数的选择非常敏感，这对于模型应用是较为不利的。另外，编码器和解码器的架构的输出几乎是不可解释的。而可解释性是自然语言处理在医学等复杂领域应用的一个关键方面，且对于后续的性能提升及模型针对性的改进是必需的。最后，自然语言处理仍然由许多问题亟待解决，包括网络安全，如假新闻检测，工业应用，如虚拟助手，以及基于最近变分自编码器或生成对抗网络的文本生成等。

致　谢

在此感谢各位学界前辈对于学术做出的贡献，让后继之人能够学到如此精妙的知识和结构设计。

参考文献

[1] Lauriola, I., Lavelli, A. and Aiolli, F. (2022) An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools. *Neurocomputing*, **470**, 443-456. https://doi.org/10.1016/j.neucom.2021.05.103

[2] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30: *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 6000-6010.

[3] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[4] Radford, A., Narasimhan, K., Salimans, T., *et al*. (2018) Improving Language Understanding by Generative Pre-Training.

[5] Liu, Y., Gu, J., Goyal, N., *et al*. (2020) Multilingual Denoising Pre-Training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726-742. https://doi.org/10.1162/tacl_a_00343

[6] Liu, Y., Ott, M., Goyal, N., *et al*. (2019) Roberta: A Robustly Optimized BERT Pretraining Approach.

[7] Clark, K., Luong, M.T., Le, Q.V., *et al*. (2020) Electra: Pre-Training Text Encoders as Discriminators Rather than Generators.

[8] Cui, Y., Che, W., Liu, T., *et al*. (2021) Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transac-*

tions on Audio, Speech, and Language Processing, **29**, 3504-3514. https://doi.org/10.1109/TASLP.2021.3124365

[9] Zhang, Z., Han, X., Liu, Z., *et al*. (2019) ERNIE: Enhanced Language Representation with Informative Entities. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 1441-1451. https://doi.org/10.18653/v1/P19-1139

[10] Joshi, M., Chen, D., Liu, Y., *et al*. (2020) Spanbert: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, **8**, 64-77. https://doi.org/10.1162/tacl_a_00300

[11] Lan, Z., Chen, M., Goodman, S., *et al*. (2019) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations.

[12] Zafrir, O., Boudoukh, G., Izsak, P., *et al*. (2019) Q8BERT: Quantized 8Bit BERT. 2019 5*th Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition* (*EMC2-NIPS*), Vancouver, 13 December 2019, 36-39. https://doi.org/10.1109/EMC2-NIPS53020.2019.00016

[13] Sanh, V., Debut, L., Chaumond, J., *et al*. (2019) DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.

[14] Jiao, X., Yin, Y., Shang, L., *et al*. (2019) TinyBERT: Distilling BERT for Natural Language Understanding. *Findings of the Association for Computational Linguistics*: *EMNLP*, 16-20 November 2020, 4163-4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

[15] Radford, A., Wu, J., Child, R., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, **1**, 9.

[16] Brown, T., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners. 34*th Conference on Neural Information Processing Systems* (*NeurIPS* 2020), Vancouver, 6-12 December 2020, 1877-1901.

[17] Ouyang, L., Wu, J., Jiang, X., *et al*. (2022) Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35: *Annual Conference on Neural Information Processing Systems NeurIPS* 2022, New Orleans, 28 November-9 December 2022, 27730-27744.

[18] Nakano, R., Hilton, J., Balaji, S., *et al*. (2021) Webgpt: Browser-Assisted Question-Answering with Human Feedback.

[19] Bubeck, S., Chandrasekaran, V., Eldan, R., *et al*. (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4.

[20] Schick, T., Dwivedi-Yu, J., Dessì, R., *et al*. (2023) Toolformer: Language Models Can Teach Themselves to Use Tools.

[21] Raffel, C., Shazeer, N., Roberts, A., *et al*. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, **21**, 5485-5551.

[22] Xue, L., Constant, N., Roberts, A., *et al*. (2020) mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, June 2021, 483-498. https://doi.org/10.18653/v1/2021.naacl-main.41

[23] Ni, J., Ábrego, G.H., Constant, N., *et al*. (2021) Sentence-T5: Scalable Sentence Encoders from Pre-Trained Text-to-Text Models. *Findings of the Association for Computational Linguistics*: *ACL* 2022, Dublin, May 2022, 1864-1874. https://doi.org/10.18653/v1/2022.findings-acl.146

[24] Dai, Z., Yang, Z., Yang, Y., *et al*. (2019) Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 2978-2988. https://doi.org/10.18653/v1/P19-1285

[25] Yang, Z., Dai, Z., Yang, Y., *et al*. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019. https://papers.nips.cc/paper_files/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html