基于预训练深度学习模型的自然语言处理模式

牛奕童*

安阳学院航空工程学院，河南安阳 455000

摘要: 自然语言处理（NLP）是人工智能（AI）的一个领域，主要目的是让计算机具有以人类的方式理解书面和口头语言的能力，主要是通过创建能够以人类的方式阅读和响应信息的计算机，然后生成其文本或语音作为回应。但是在自然语言处理的过程中可能会遇到效率不高的问题，所以怎样使自然语言处理的过程更高效是目前人们所研究的一个方向。本文旨在通过计算语言学和统计学、机器学习和深度学习模型的结合，得出一套可编程的规则来帮助 NLP 描述人类语言。这样当文本和语音数据结合在一起时，计算机可以以文本或音频数据的形式理解人类语言，例如说话者或作者的意图和情感。本文介绍了用于 NLP 分析和研究方面的各类深度学习系统，并描述了一种基于预训练的自然语言处理方法。最终发现可以运用于改善运营效率，提高员工生产力，以及简化关键任务的业务操作等场景。

关键词: 深度学习；预训练学习模型；自然语言处理

DOI: 10.57237/j.cst.2022.01.007

A Natural Language Processing Model Based on Pre-trained Deep Learning Models

Yitong Niu

Aviation College, AnYang University, Anyang 455000, China

Abstract: Natural language processing (NLP) is a field of artificial intelligence (AI) whose primary purpose is to give computers the ability to understand written and spoken language in a human-like manner, mainly by creating computers that can read and respond to information in a human-like manner and then generate their text or speech as a response. However, the process of natural language processing may encounter the problem of inefficiency, so how to make the process of natural language processing more efficient is a direction that is currently being studied. This paper aims to derive a set of programmable rules from helping NLP describe human language by combining computational linguistics and statistics, machine learning and deep learning models. In this way, when text and speech data are combined, computers can understand human language in the form of text or audio data, such as the intent and emotion of the speaker or author. This paper introduces various types of deep learning systems used in NLP analysis and research and describes a pre-training-based approach to natural language processing. The final findings can improve operational efficiency, increase employee productivity, and streamline mission-critical business operations.

Keywords: Deep Learning; Pre-Trained Deep Learning Models; Natural Language Processing

*通信作者：牛奕童, itong_niu@163.com

收稿日期: 2022-09-24; 接受日期: 2022-10-20; 在线出版日期: 2023-01-05

1 引言

人工神经网络（A.N.N.）在深度学习中被用来模拟和模仿人脑的功能[1]。这是人工智能最重要的领域之一。为了从大量的数据中学习，这些神经网络试图模仿大脑的功能。隐蔽层可以提高单层神经网络预测的准确性，单层神经网络可能只能做出近似的预测。众多的人工智能应用和服务依靠深度学习来实现广泛的分析和物理任务的自动化[2-4]。当今社会最广泛使用的技术之一是深度学习。以下是深度学习在现实世界领域的几个主要应用:

1) 实时计算机视觉和图像分析
2) 股票交易和金融数据分析
3) 虚拟助理
4) 自动化制造
5) 数据科学和工程
6) 语音识别（声乐人工智能）。
7) 娱乐和音乐记谱
8) 电子商务中的购物模式分析
9) 社交媒体上的情绪分析
10) 客户关系管理系统
11) 自动驾驶汽车、自动驾驶汽车和无人机
12) 自然语言处理（NLP）。
13) 广告和促销活动
14) 情感情报
15) 欺诈检测和网络安全
16) 医疗保健和医疗诊断
17) 投资建模

图 1 深度学习的关键应用

图 1 描绘了与深度学习相关的各种关键应用，其中包括安全、隐私、数据科学、计算机视觉、医疗保健、医疗部门、自然语言处理等。如今，深度学习被用于许多需要高度完整性和准确性的高性能领域。

针对不同种类的应用程序，深度学习模型也有很多类型。每个研究和应用领域都使用特定的深度学习模型，以获得更高程度的功效、性能和准确性，如表 1 所示[5-12]。

表 1 深度学习模型和用例。

深度学习模型	用例
经典神经网络（多层感知器）	表格数据分析、分类和基于回归的问题解决
Boltzmann Machines	基于监视和监视的应用程序
CNN	图像数据集、光学字符识别 (OCR) 智能
RNN	图像分类、图像字幕、情感分析、视频分类
SOM	降维、音乐、视频
Encoders	庞大的数据集分析、推荐引擎、降维

用于 NLP 的预训练模型（PTMs）是在大型数据集上训练的深度学习模型（如转化器），以执行特定的 NLP 任务。PTMs 在大型语料库上训练时，可以学习通用的语言表征，这对下游的 NLP 任务是有益的，可以避免从头开始训练一个新的模型。这样一来，预训练的模型可以被称为可重用的 NLP 模型，NLP 开发者可以用它来快速建立 NLP 应用。Transformers 提供了一套预训练的深度学习 NLP 模型，跨越不同的 NLP 任务，如文本分类、问题回答、机器翻译等。这些预训练的 NLP 任务是免费提供的，使用它们不需要 NLP 知识。第一代预训练模型被训练为学习良好的词嵌入。然而，最新的或第二代 PTM 的训练是为了学习上下文的词嵌入。预训练的模型可以很容易地加载到 NLP 库中，如 PyTorch、Tensorflow 等，并用于执行 NLP 任务，几乎不需要 NLP 开发者做额外的努力。预训练模型在 NLP 任务中的使用越来越频繁，因为与定制模型相比，它们更容易实现，具有较高的准确性，并且需要较少的训练时间。

2 方法

预训练的模型被用来以高精确度快速实现深度学习。预训练的模型具有权重，可以由研究人员导入，在特定领域快速部署深度学习应用，而无需从头建模[13-16]。

一般来说，在训练网络时，使用预训练模型有两个主要的优点。一是缩短训练时间。由于网络已经在类似的数据上进行了预训练，模型的收敛时间可以加快。例如，在 ImageNet 中，分类问题的预训练模型已经通过分类任务理解了图像的线条和纹理的含义，所以它可以快速适应其他图像任务。二是减少目标域中所需要的数据量。如上所述，预训练可以帮助模型学习一般的知识，所以有可能使用较少的目标域数据来训练一个具有良好泛化能力的模型。

表 2 预训练深度学习模型的各种应用。

深度学习模型	用例
自然语言处理	1) OpenAI GPT-3 2) Google BERT 3) Google ALBERT 4) Google Transformer-XL 5) ULMFiT 6) Facebook RoBERTa 7) Microsoft CodeBERT 8) ELMo 9) XLNet
音频和语音	1) Wavenet 2) Lip Reading 3) MusicGenreClassification 4) Audioset 5) DeepSpeech 6) Waveglow 7) Loop 8) TTS 9) ESPNET 10)MXNET-Audio

2.1 基于预训练模型库的主要优势

相较于没有经过预训练的数据集，一个在知道哪些参数有可能取得好的结果方面有先机的模型可以更快地被优化。此外，预训练的模型有一个好处，就是不需要像从头开始建立一个模型那样多的数据。以下就是经过预训练的模型库所带来的几种优势：

1) 包含预处理的微调功能
2) 易于使用的脚本和 API
3) 多语言支持，包括国际和地区语言
4) 与图形处理单元（GPU）的兼容性
5) 来自顶级公司的预编程算法

2.2 安装和使用预训练的自然语言处理的工作

如图 2 所示，HuggingFace（网址：https://huggingface.co/）是为自然语言处理（NLP）提

供预训练模型的重要平台之一。HuggingFace 是基于云的，可以与谷歌 Colab 集成，用于运行脚本。这个平台在多个基于研究的应用中整合了数百个预训练的模型和人工智能与机器学习的架构。

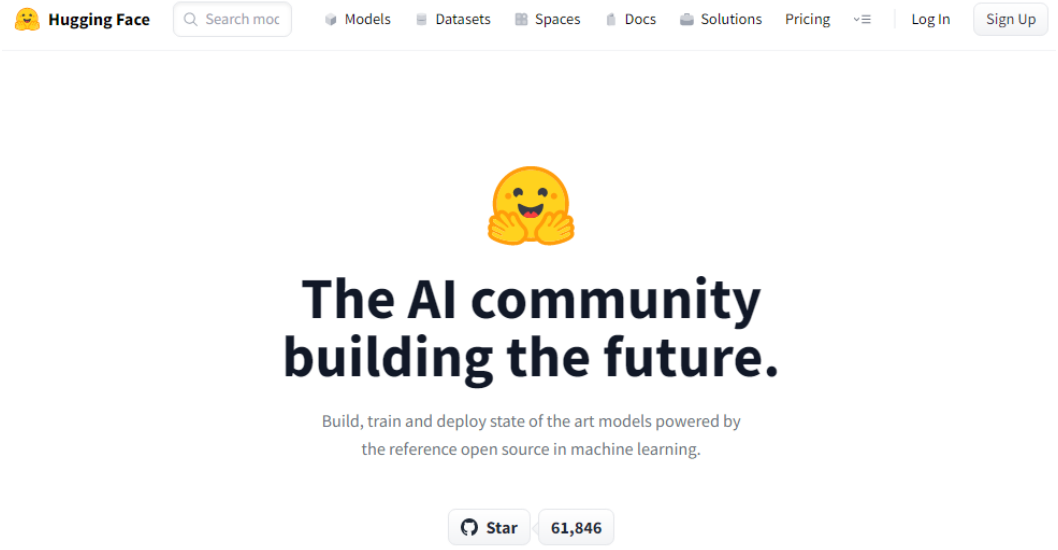

图 2 在线平台——HuggingFace 的预训练模型

要在 Google Colab 中安装基于 NLP 的预训练模型，请执行以下内容

```
! pip install pytorch-transformers
! pip install transformers
! pip install sentence piece
```

3 结果和分析

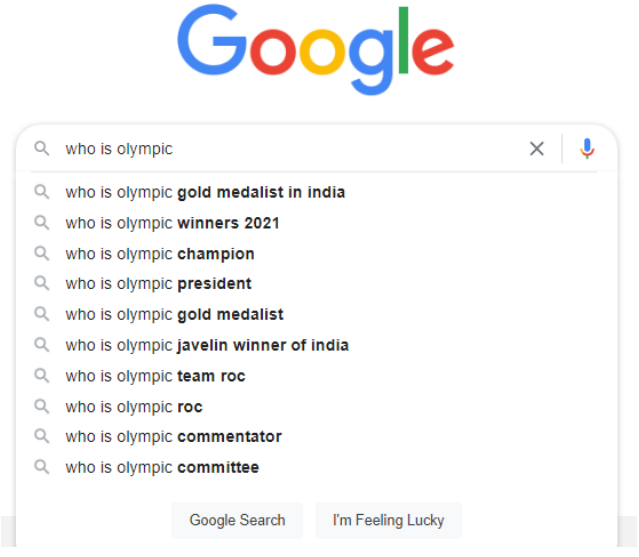

图 3 谷歌搜索中下一个序列的预测

当我们在谷歌搜索上写下一些文字时，下一个序列会由谷歌的后端库[17-19]建议，如图 3 所示。例如，如果我们想预测印度人的名字是什么之后的下一个词，可以使用下面的转化器[20]。使用自然语言处理的预训练模型，这种类型的应用可以进行自定义部署。

```
mytokenizer = Tokenizer.from_pretrained('2')
# Encode a ThisText inputs
this text = "what is the name of the Indian    "
IndexCurrented_tokens = mytokenizer.encode(ThisText)
tokens_tensor = torch.tensor([IndexCurrented_tokens])
model = LMHeadModel.from_pretrained('2')
model.eval()
tokens_tensor = tokens_tensor.to('cuda')
model.to('cuda')
with torch.no_grad():
      outs = model(tokens_tensor)
      preds = outs[0]
pred_IndexCurrent = torch.argmax(preds[0,-1, :]).item()
pred_ThisText = tokenizer.decode(IndexCurrented_tokens +
[pred_IndexCurrent])
print(pred_ThisText)
Output
The output from the execution of code will be predicted
depending on the following
    • flag
    • parliament
        ○ and many others depending upon the search
```

HuggingFace 正在为巨大的应用提供模型，并被众多企业巨头使用，包括微软、谷歌、Grammarly、SpeechBrain、Facebook、亚马逊网络服务和其他许多企业，如图 4 所示。这些组织开发和部署的算法可以被从业者和研究人员用于巨大领域的研究应用。

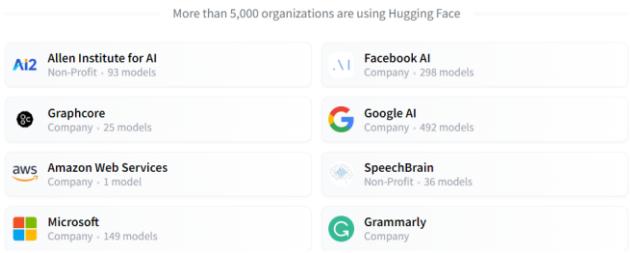

图 4 谷歌搜索中下一个序列的预测

使用 NLP [21]的预训练模型可以解决经典的实时搜索填空问题。以下是预测单词代替[MASK]的案例示例。

```
myprediction = pipeline('bert specifications')
myprediction ("This is a [MASK].")
Output
 [{'CurrentScore': 0.03235777094960213,
   'sequence': 'this is a dream.',
   'KeyToken': 3959,
   'KeyToken_str': 'dream'},
  {'CurrentScore': 0.030467838048934937,
   'sequence': 'this is a mistake.',
   'KeyToken': 6707,
   'KeyToken_str': 'mistake'},
  {'CurrentScore': 0.028352534398436546,
   'sequence': 'this is a test.',
   'KeyToken': 3231,
   'KeyToken_str': 'test'},
  {'CurrentScore': 0.025175178423523903,
   'sequence': 'this is a game.',
   'KeyToken': 2208,
   'KeyToken_str': 'game'},
  {'CurrentScore': 0.024909017607569695,
   'sequence': 'this is a lie.',
   'KeyToken': 4682,
   'KeyToken_str': 'lie'}]
unmasker = pipeline('bert model')
myprediction ("He is a [MASK].")
Output
[{'CurrentScore': 0.17371997237205505,
   'sequence': 'he is a christian.',
   'KeyToken': 3017,
   'KeyToken_str': 'christian'},
  {'CurrentScore': 0.08878538012504578,
   'sequence': 'he is a democrat.',
   'KeyToken': 7672,
   'KeyToken_str': 'democrat'},
  {'CurrentScore': 0.06659623980522156,
   'sequence': 'he is a republican.',
   'KeyToken': 3951,
   'KeyToken_str': 'republican'},
  {'CurrentScore': 0.03911091387271881,
   'sequence': 'he is a vegetarian.',
   'KeyToken': 23566,
   'KeyToken_str': 'vegetarian'},
  {'CurrentScore': 0.036758508533239365,
   'sequence': 'he is a catholic.',
   'KeyToken': 3234,
   'KeyToken_str': 'catholic'}]
```

文本准备是每个 NLP 工作的最初步骤。这只是一个把数据变成可以分析的形式的问题[22]。创建一个优秀的 NLP 应用程序是一个必不可少的阶段。标记化是其中最关键的。它的内容是将文本材料流划分为有意义的标记，称为代币的过程。这是一种分解[23, 24]。标记化可以使用各种开源技术来完成。本节将探讨标记化的必要性和标记化的多种形式，以及执行标记化的几种工具和它们面临的困难[25, 26]。一个文档的标记出现可以作为一个矢量来表示该文档。此外，它们可以被计算机利用来激活有用的活动和反应。然后，它们可以作为机器学习管道中的特征被利用，以启动更复杂的判断或行动[27-36]。

从模拟尝试和执行的一致性方面获取的结果。算法和预训练的模型来自 huggingface，并与 Google Colab 集成，以便有效评估结果，如表 3 和图 5 所示。

表 3 执行时间的一致性。

模型	执行时间（秒）
1	0.00049
2	0.00042
3	0.00053
4	0.00041
5	0.00042

图 5 执行时间的一致性分析

从结果和图形描述中可以看出，基于 NLP 的预训练模型是相当有效的，并且在执行时间方面也是非常少的，这表明这些算法和模型可以在没有任何特定要求的情况下被集成到多个应用中，并且具有完整性的结果[37-39]。这些模型可以与 Google Colab 或甚至包括 Raspberry Pi、Arduino 或任何其他的开源硬件联系起来，这样就可以为多个部分开发和部署实时应用。

4 结论

本文以期望对于学术界和相关从业人员在音频、

语音和自然语言处理领域的研究，可以使用免费和开源的预训练模型，以便在真实世界的数据集上实现更好的调整、准确性和性能。音频取证、语音识别、语音到文本的转换、音频翻译和其他动态应用从基于云的系统上的模型中获益匪浅。语言学和认知计算是处理计算机和自然语言如何互动的两个领域，特别是如何构建能够处理和分析大量自然语言的计算机。最终，我们的目标是创造出不仅能理解文件内容，而且能理解文件写作背景的软件。届时，人类将不需要再去手工整理和分类文件，因为这项技术将自动为他们做这件事。

参考文献

[1] Haque N I, Islam M, Ahsan M M. A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning [J]. Technologies, 2022, 10 (3): 57.

[2] Lauriola A L, Aiolli F. An introduction to deep learning in natural language processing: Models, techniques, and tools [J]. Neurocomputing, 2022, 470: 443-456.

[3] Alaidi M, Salim H, Aljazaery I A, Abbood S H. Dark web illegal activities crawling and classifying using data mining techniques [J]. International Journal of Interactive Mobile Technologies, 2022, 16 (10).

[4] Aljazaery A, ALRikabi H T S, Alaidi A H M. Encryption of Color Image Based on DNA Strand and Exponential Factor [J]. International Journal of Online Biomedical Engineering, 2022, 18 (3): 101-113.

[5] ALRikabi S, Qateef J S, Al_airaji R M. Face Patterns Analysis and recognition System based on Quantum Neural Network QNN [J]. International Journal of Interactive Mobile Technologies, 2022, 16 (9).

[6] Alseelawi H T H, ALRikabi H T S. A Novel Method of Multimodal Medical Image Fusion Based on Hybrid Approach of NSCT and DTCWT [J]. International Journal of Online Biomedical Engineering, 2022, 18 (3).

[7] Azeez A, Abdul-Hussein M K, Mahdi M S. Design a system for an approved video copyright over cloud based on biometric iris and random walk generator using watermark technique [J]. Periodicals of Engineering Natural Sciences, 2022, 10 (1): 178-187.

[8] Charniak. Parsing as language modeling [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

[9] Goodfellow Y B, Courville A. Deep learning [M]. MIT press, 2016.

[10] Goldberg. A primer on neural network models for natural language processing [Z].

[11] Guida, Mauri G. Evaluation of natural language processing systems: Issues and approaches [J]. Proceedings of the IEEE, 1986, 74 (7): 1026-1035.

[12] Joshi K, Weinstein S. Control of Inference: Role of Some Aspects of Discourse Structure-Centering [M] // IJCAI.

[13] Koskenniemi. Two-level morphology: A general computational model for word-form recognition and production [M]. Finland: University of Helsinki, Department of General Linguistics Helsinki, 1983.

[14] Jozefowicz O V, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling [Z].

[15] Mezaal S, Abdulkareem S F. Affine cipher cryptanalysis using genetic algorithms [J]. JP Journal of Algebra, Number Theory Applications, 2017, 39 (5): 785-802.

[16] Mezaal S, Hammood D A, Ali M H. OTP encryption enhancement based on logical operations [C] // 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC. IEEE, 2016: 109–112.

[17] Salah, Khairy R. The Detection of Counterfeit Banknotes Using Ensemble Learning Techniques of AdaBoost and Voting [J]. International Journal of Intelligent Engineering and Systems, 2021, 14 (1): 326-339.

[18] Shareef S, Abd T, Mezaal Y S. Gender voice classification with huge accuracy rate [J]. TEL KOMNIKA, 2020, 18(5): 2612-2617.

[19] Sundaram S, Gurajada S, Fisichella M, Abraham S S. Why are NLP Models Fumbling at Elementary Math? A Survey of Deep Learning based Word Problem Solvers [Z].

[20] TH., Hazim H T. Enhanced Data Security of Communication System Using Combined Encryption and Steganography [J]. International Journal of Interactive Mobile Technologies, 2021, 15 (16).

[21] Vinyals Ł K, Koo T, Petrov S, Sutskever I, Hinton G. Grammar as a foreign language [J]. Advances in neural information processing systems, 2015, 28.

[22] Abbas F, Yasmin M, Fayyaz M, Elaziz M A, Lu S, El-Latif A A A. Gender Classification Using Proposed CNN-Based Model and Ant Colony Optimization [J]. Mathematics, 2021, 9 (19): 2499.

[23] Abd-El-Atty A M I, Alaskar H, El-Latif A A A. A Robust Quasi-Quantum Walks-based Steganography Protocol for Secure Transmission of Images on Cloud-based E-healthcare Platforms [J]. Sensors, 2020, 20 (11): 3108.

[24] Abdullah S, Abed M A, Barazanchi I A. Improving face recognition by elman neural network using curvelet transform and HSI color space [J]. Period. Eng. Nat. Sci, 2019, 7 (2): 430-437.

[25] Ahmed S M, Ahmad M, El-Latif A A A. Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling [J]. IEEE Trans. Intell. Transp. Syst, 2021: 1-10.

[26] Elgendy A, Zhang W-Z, He H, Gupta B B, El-Latif A A A. Joint computation offloading and task caching for multi-user and multi-task MEC systems: reinforcement learning-based algorithms [J]. Wirel. Networks, 2021, 27 (3): 2023-2038.

[27] Hammad. Deep Learning Models for Arrhythmia Detection in IoT Healthcare Applications [J]. Comput. Electr. Eng, 2022, 100: 108011.

[28] Johnson. How the statistical revolution changes (computational) linguistics [C] // Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?

[29] Rashid A A, Barazanchi I A, Jaaz Z A. Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set [J]. Period. Eng. Nat. Sci, 2019, 7 (2): 448-457.

[30] Resnik. Four revolutions [J]. Language Log, February, 2011, 5.

[31] Sabir F S. An Automated Real-Time Face Mask Detection System Using Transfer Learning with Faster-RCNN in the Era of the COVID-19 Pandemic [J]. Comput. Mater. Contin, 2022, 71 (2): 4151-4166.

[32] Salih K, See O H, Yussof S, Iqbal A, Salih S Q M. A proactive fuzzy-guided link labeling algorithm based on MIH framework in heterogeneous wireless networks [J]. Wirel. Pers. Commun, 2014, 75 (4): 2495-2511.

[33] Salih Q, Alsewari A R A, Yaseen Z M. Pressure vessel design simulation: Implementing of multi-swarm particle swarm optimization [Z] 10.1145/3316615.3316643.

[34] Schank C, Abelson R P. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures [M]. Psychology Press, 2013.

[35] Tawalbeh F M, Tawalbeh M, Quwaider M, El-Latif A A A. Edge enabled IoT system model for secure healthcare [J]. Measurement, 2022, 191: 110792.

[36] Trabelsi P K, Komurcugil H. Mitigation of grid voltage disturbances using quasi-Z-source based dynamic voltage restorer [C] //2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG 2018. IEEE, 2018: 1–6.

[37] Turchin, Builes L F F. Using natural language processing to measure and improve quality of diabetes care: a systematic review [J]. Journal of Diabetes Science Technology, 2021, 15 (3): 553-560.

[38] Winograd. Procedures as a representation for data in a computer program for understanding natural language [M] // MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC1971.

[39] Niu Y, Korneev A. Identification Method of Power Internet Attack Information Based on Machine Learning [J]. Iraqi Journal for Computer Science and Mathematics, 2022, 3 (2): 1–7.