计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):141−147 [doi: 10.15888/j.cnki.csa.006900]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CNN 和 LSTM 的异构数据舆情分类方法①

黑富郁, 王景中, 赵林浩

(北方工业大学 计算机学院, 北京 100144)
通讯作者: 黑富郁, E-mail: 2413386315@qq.com

摘　要: 随着网络的发展, 网络舆情数据呈现出爆炸式增长的趋势. 使得数据类型越来越复杂, 这些网络数据相互结合, 构成了一个复杂的数据结构来表达数据的信息. 在舆情数据中, 通过单一类型的数据 (图片、文本、语音等) 越来越难以完整的表达数据信息. 对于一个包含多种类型数据的网络信息, 本文提出一种新的舆情分类模型, 通过神经网络模型分别去学习不同类型信息的数据特征, 对它们的特征融合后进行分类, 通过这种方法实现数据信息更好地分类. 在实验中, 本文分别使用 LSTM 和 CNN 神经网络提取文本和图像数据特征, 对二者特征融合后进行分类. 结果证明, 多种类型的数据特征进行融合后再分类, 可以更好地实现对网络舆情数据信息的分类, 提高了舆情信息分类的准确性.

关键词: 异构数据; 神经网络; CNN; LSTM; 特征提取; 特征融合; 舆情分类

引用格式: 黑富郁,王景中,赵林浩.基于 CNN 和 LSTM 的异构数据舆情分类方法.计算机系统应用,2019,28(6):141–147. http://www.c-s-a.org.cn/1003-3254/6900.html

Public Opinion Classification of Heterogeneous Data Based on CNN and LSTM

HEI Fu-Yu, WANG Jing-Zhong, ZHAO Lin-Hao

(School of Computer, North China University of Technology, Beijing 100144, China)

Abstract: With the development of the network, the public data which shows the trend of explosive growth, making the data type more and more complex. These network data combine with each other to form a complex network data structure to express the information of data. In this scenario, it is increasingly difficult to fully express data information through a single type of data (picture, text, voice, etc.). For the purpose of a network information that contains multiple types of data can be classified better, this study proposes a new public opinion classification model via neural network which is used to learn the data features respectively, and to classify their features after fusion. In the experiment, LSTM and CNN neural networks are used to extract text and image's features, fusing the two features to classified. The experimental results show that the reclassification after the fusion of various data features can better realize the classification and improve the accuracy of data information classification.

Key words: heterogeneous data; neural network; CNN; LSTM; feature extraction; feature fusion; public opinion classification

随着互联网的发展, 网络已经成为民众不可或缺的生活必备品. 根据第 41 次《中国互联网络发展状况统计报告》, 截至 2017 年 12 月, 我国网民规模达 7.72 亿, 手机网民规模达 7.53 亿, 网民使用手机上网人群的占比由 2016 年的 95.1% 提升至 97.5%[1]. 人们在获取多样化信息的同时, 过多的信息也造成了人们注意力的分散, 对舆情分析造成了极大的困难. 因此对舆情信息进行分类具有重要意义. 一方面, 可以按照类别

① 收稿时间: 2018-11-22; 修改时间: 2018-12-12; 采用时间: 2018-12-26; csa 在线出版时间: 2019-05-25

统计和查询各类事件信息, 统计形成相关的简报. 另一方面, 由于同一事件在网络上会有大量不同新闻报道, 对舆情进行分类可以快速查找定位相关的信息, 从技术上为判断不同来源的同一事件提供支持.

现在舆情分析主要是针对文本进行分类[2,3], 但是大数据[4]时代的到来使得网络上的舆情数据越来越多且复杂 (例如视频、声音、文本等), 这些不同类型的数据包括图片、视频、语音等都承载了越来越多的信息和内容. 网络舆情数据中包含的各种类型的信息, 它们在内容上和结构上相互之间有着密切的相关性, 只是通过网络舆情数据中的某一类型的数据进行分类, 这种忽视了不同数据之间的关联的传统分类方法渐渐不适用于当下的网络舆情数据信息.

为了应对这样的情况, 研究出更先进的技术是组织和管理这些数据的重要依据, 在这些技术中优秀的分类技术 (例如文本分类、图像分类等) 是其它技术的基础, 通过好的分类技术可以更好的管理这些信息. 近几年在数据处理技术方面的相关研究中, 神经网络的发展势头尤其迅猛. 在图像处理方面, 通过神经网络对图像的处理已经屡见不鲜, 例如人脸识别、物体识别、场景检测都已经有了长远的发展. 在图像处理、语音处理等领域取得的巨大进展的同时, 神经网络的焦点也开始汇集于自然语言处理方面的应用. 伴随着相关技术的日渐成熟, 为各类型数据的融合处理打下了良好的基础. 其实, 国外早在 19 世纪就已经开始信息融合的相关工作, 并且将信息融合技术列为 20 世纪开发和研究的关键技术之一. 然而我国展开对信息融合技术的研究时间较晚, 主要局限于军事相关的领域且发展缓慢. 通过三十多年的研究, 虽然现在信息融合方面的研究尚不成熟, 但是信息融合技术已经得到了非常广泛的关注和应用.

现在的信息融合技术从抽象的层次来分类, 可以分为数据层级融合、特征层级融合和决策层级融合. 本文主要从特征层级来考虑并实现对本文课题的研究. 基于舆情数据的分布情况、现行的概念和技术, 本文提出一种结合了不同类型的数据来进行综合考虑的舆情分类方法.

1 相关工作

1.1 神经网络

自 2012 年 Krizhevsky 等人在 ILSVRC-2012 大赛中, 利用深度卷积神经网络对 ImageNet 数据集进行分类, 取得优秀的结果并以此获得冠军[5]. 神经网络被学界和工业界越来越重视, 神经网络得以被广泛的应用于各领域. 2014 年, Simonyan 等人[6]提出一种名为 VGG16 的卷积神经网络, 该神经网络模型在 ILSVR2014 的比赛中获得冠军. Hochreiter 等人在 RNN 的基础上提出了长短时记忆网络 (Long Short-Term Memory, LSTM)[7], LSTM 很好的解决了语义的长距离依赖问题. 近年来, LSTM 模型被成功地应用于机器翻译[8]及信息检索[9]等方面.

1.2 异构数据的特征学习

Ngiam 等人提出了多模态深度学习模型, 通过玻尔兹曼机 (RBM) 分别独立地进行训练以提取视频和语音数据的特征, 在特征层对二者特征进行组合, 对多模态数据进行联合表示. 再通过多模态数据的联合表示的特征去学习数据的高层语义特征[10]. 2012 年, Srivastava 等人提出了一种新的与 Ngiam 等人的方法相似的训练过程, 同样是利用受限玻尔兹曼机独立学习不同数据的特征然后将二者的特征组合起来, 最后再通过监督标签对参数进行微调[11]. 除此之外与 Ngiam 等人不同的一点是, Srvastava 处理的是文本和图像数据. 冯方向通过自动编码机分别对不同模态信息进行特征抽取并通过典型关联分析学习共有信息以实现跨模态检索[12]. 异构数据特征学习方法还包括 Huiskes 提出的多模态支持向量机模型和 Guillaumin等人提出的多模特半监督学习方法等[13–15].

越来越多的神经网络模型被构建, 但是它们只是针对单一类型的数据来进行分类, 同时现在的多模态学习方法也主要是针对各类数据信息对称的异构数据, 而针对各类型数据信息不对称的网络舆情数据分类, 以上的方法难以适用.

2 相关技术

2.1 LSTM 神经网络模型

LSTM 神经网络是一种特别的 RNN 神经网络, 使用 LSTM 神经网络来对处理文本信息, 通过这种方法可以防止 RNN 神经网络常见的梯度爆炸问题, 同时 LSTM 的记忆机制在处理长文本信息方面也具有一定优势.

Embeding 层通过 Word2Vec 方法把文本信息表示到向量空间. 通过 LSTM 隐藏层提取文本特征, LSTM 隐藏层由一系列的 LSTM 基本单元组成.

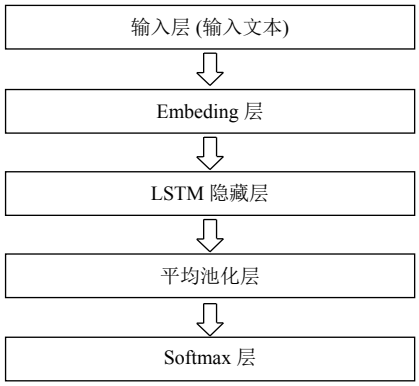

图 1 LSTM 模型

平均池化层: 通过对 LSTM 隐藏层的数据特征进行池化操作提取出新的特征, 实现特征的降维, 这样既可以降低计算复杂度又可以防止过拟合. 同时因为 LSTM 隐藏层的每一个特征向量都对分类结果有影响, 为了保证分类的准确度这里使用平均池化. 最后通过 Softmax 层对提取到的特征进行分类.

2.2 CNN 卷积神经网络模型

卷积神经网络采用权值共享工作方式, 相邻两层只有部分节点相连, 这种模式显著降低了神经网络模型的复杂度, 减少了权值的数量, 因而成为了现在众多领域研究的热点. 由于 CNN 神经网络可以绕过复杂的预处理过程直接输入原始图像, 而得到了学术界和工业界的青睐. 其中有代表性的 VGG16 模型, 它是由 16 层卷积层和全连接层组合而成, 其中前 13 层为卷积层, 后 3 层为全连接层. 整个模型如图 2 所示.

卷积神经网络通过卷积层和池化层来完成特征提取. 卷积层使输入的特征图 (或原始图像) 与卷积核进行卷积操作, 最终通过非线性的激活函数得到新的特征图. 池化层进行下采样操作, 通过激活函数得到一个更小的特征图, 以此减少训练参数降低神经网络的复杂度, 并防止过拟合现象. 通过全连接层来将特征映射到特征空间, 全连接层的每一个神经元与前一层的所有神经元进行全连接, 全连接层可以整合池化层中具有类别区分性的局部信息. 最后一层全连接层的输出值, 通过 Softmax 层进行分类.

3 舆情分类

随着大数据时代的到来和网络技术的不断提升, 不同类型的数据开始越来越多出现在网络上, 这些不同类型的数据在网络上构成了一个复杂的集合. 与以往不同, 单一类型的数据难以完整表达舆情数据的信息. 通过对舆情数据的多种类型数据综合考虑进行分类, 以便能够在舆情数据中挖掘出更多有价值的信息和知识, 更好地利用舆情数据.

VGG16 神经网络结构
输入(224*224 RGB 图片)
卷积层 3-64
卷积层 3-64
池化层 (最大池化)
卷积层 3-128
卷积层 3-128
池化层 (最大池化)
卷积层 3-256
卷积层 3-256
卷积层 3-256
池化层 (最大池化)
卷积层 3-512
卷积层 3-512
卷积层 3-512
池化层 (最大池化)
卷积层 3-512
卷积层 3-512
卷积层 3-512
池化层 (最大池化)
全连接层-4096
全连接层-4096
全连接层-1000
softmax层

图 2 VGG16 网络模型

3.1 异构数据的特征提取

不同类型数据的底层信息存在明显的差异, 本文考虑到不同类型的数据, 例如图像数据和文本数据, 文本数据的表示通常是离散的, 而图像数据的表示则是连续的, 因此很难在底层数据表示上建立不同类型数据之间的关联. 神经网络适用于不同类型数据信息的特征提取, 考虑到各类数据信息的特点选择更加适合的神经网络模型并通过全连接层来将各类信息特征表达到相同的特征空间.

神经网络的全连接层的结点与上一层的每一个结点相连, 用来将前面提取到的特征综合起来. 由于其全连接的特性, 一般的全连接层的参数也是最多的. 全连接层的核心就是矩阵的乘积操作, 具体过程如下:

矩阵表示 (其中 W_{ij} 表示权重系数, b_i 表示偏置系数):

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ w_{31} & w_{32} & w_{33} & & w_{3n} \\ & \vdots & & \ddots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \tag{1}$$

通过全连接层能将特征空间中的特征映射到另一个特征空间. 在 CNN 神经网络中, 全连接层一般出现在整个神经网络的最后几层, 对前面提取的特征做加权和, 起到将提取到的特征映射到样本标记空间的作用. 在 RNN 等神经网络中, 全连接层也可以用来将 embedding 空间映射到隐层空间, 再将其映射到样本标记空间.

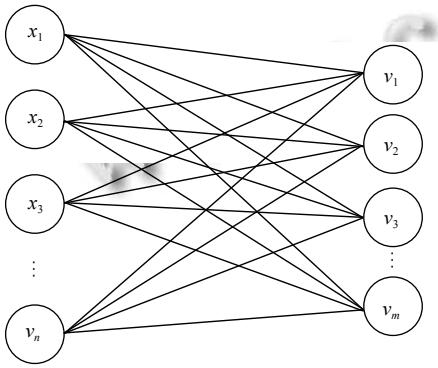

图 3 全连接层操作

基于神经网络对不同类型数据的良好适用性, 本文通过神经网络来实现特征的提取. 在现有神经网络模型的基础上, 在最后几层构建全连接层将不同类型的信息表示到同一特征空间, 以便对各类数据特征进行融合.

据此, 本文已构建了以下两个特征提取模型. 在上文提到的 CNN 和 LSTM 模型的基础上增加或调整全连接层构建出新的 CNN 模型和 FC-LSTM 模型如图 4 所示.

神经网络分别单独通过不同类型的数据训练后, 去掉神经网络的 Softmax 分类器即可得到对应的特征提取模型. 通过调整的神经网络模型, 它们抽取的特征已经表示在了同一特征空间上, 在此基础上可以直接对特征进行融合.

3.2 异构数据的特征融合分类

由于舆情信息的各类型数据包含的内容并不对称, 只是简单地将数据特征进行融合, 难以达到预期的效果. 考虑到不同类型的信息的重要性, 具体的融合过程如下:

$$\sum_{i=1}^{n} V_i = \sum_{i=1}^{n} (V1_i \times W_1 + V2_i \times W_2) \tag{2}$$

其中, $V1_i$、$V2_i$ 表示不同类型信息的特征向量, V_i 表示融合后的特征向量, W_1、W_2 分别表示不同类型信息的权重, 这里通过对若干条数据测试来确定 W_1、W_2, 测试过程如图 5 所示.

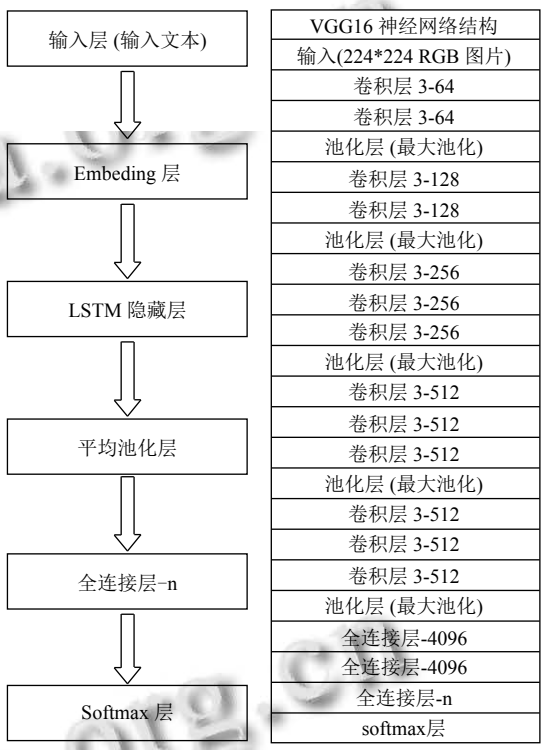

图 4 FC-LSTM 模型和 CNN 模型

通过特征提取模型之后, 在对整个特征融合过程中, 让特征 $V1_i$、$V2_i$ 分别和权重 W_1、W_2 求积, 将它们的结果相加得到融合后的特征.

$$\vec{z} = \frac{\exp(V)}{\exp(V_i)} \tag{3}$$

最后, 使用了 Softmax 分类器 (3) 对融合后的特征进行分类.

3.3 异构数据的舆情分类模型

根据上文可以架构出整个模型. 如图 6 所示.

以此 (图 6), 通过不同的神经网络分别去提取不同类型网络数据的特征, 将他们表达到同一特征空间, 并通过特征融合获取更加全面的数据信息来对网络数据进行分类.

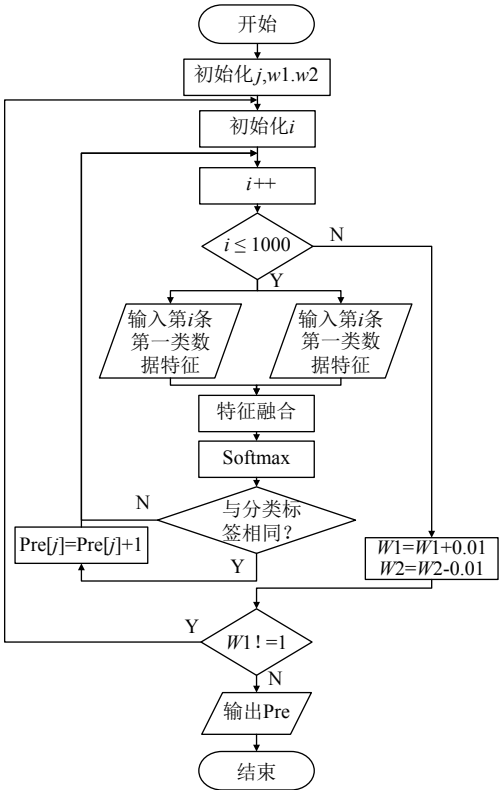

图 5 权重获取流程图

的异构舆情数据库. 为此, 本文收集了搜狐、腾讯网站上的图像和文本数据信息, 采用图像和文本这两种类型的数据信息来进行实验验证. 它们的内容如表 1 所示.

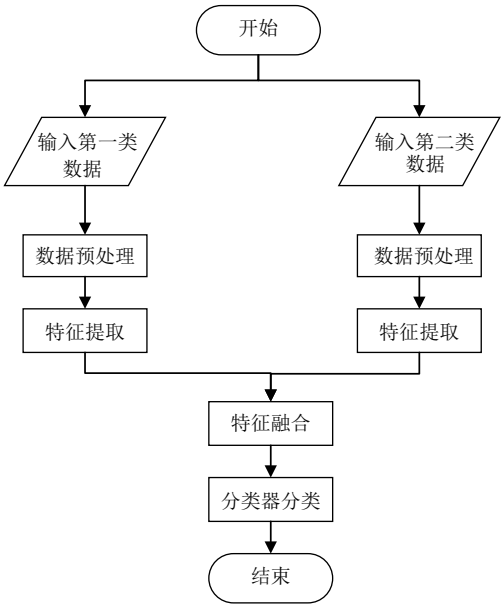

图 6 舆情分类模型

4 实验与讨论

4.1 数据集

在数据集上, 当前缺少一个公开的具有一定标准

本文选取内容较多的军事、历史、旅游、财经、房产、科技、体育、娱乐八个类别进行分类, 一共收集了 9000 条数据, 各类别数据一千多条, 将其中的 8000 条作为训练数据集, 剩下 1000 条作为测试数据集.

表 1 数据集

来源	主题分类数	主题内容
搜狗新闻	15	国内、国际、社会、军事、娱乐、财经、体育、教育、房产、女人、汽车、科技、互联网、游戏、公益
腾讯新闻	18	国际、军事、历史、文化、公益、旅游、财经、娱乐、体育、房产、科技、汽车、游戏、文化、教育、数码、时尚

4.2 实验设置

通过上文构建的 CNN 神经网络和 FC-LSTM 神经网络特征提取模型构建分别提取图像和文本信息的特征, 实现特征融合并进行舆情分类. 采用 CNN、LSTM、FC-LSTM 神经网络模型和 LSTM-CNN 神经网络模型进行对比实验.

CNN 神经网络模型: 通过 CNN 神经网络模型仅对图片进行分类.

LSTM 神经网络模型: 通过 LSTM 神经网络模型对文本进行分类.

FC-LSTM 神经网络模型: 通过 FC-LSTM 神经网络模型对文本进行分类.

LSTM-CNN 多模态深度学习模型: 对图像和文本进行特征提取, 并对提取到的特征融合后再进行分类.

4.3 实验结果

分析图 7 可知, 随着文本权重 W_1 的变小和图像权重 W_2 的变大, 分类的准确率开始上升, 当文本和图像的权重分别为 W_1=0.81, W_2=0.19 时, LSTM-CNN 可以得到准确性最好的分类结果, 之后随着图像权重 W_2 的增加, 准确率开始出现下降. 当分类结果达到最优时, 图像权重 W_2 远远小于文本权重 W_1, 经分析对比图像和文本数据具备以下特点:

1) 信息承载量: 在图片中可以包含的信息量少于文本信息. 文本信息可以承载更多的信息.

2) 信息可靠度: 文本信息与图像信息相比可靠性更高. 在一些相对数据质量不高网络数据中, 相对应的图像质量要更低.

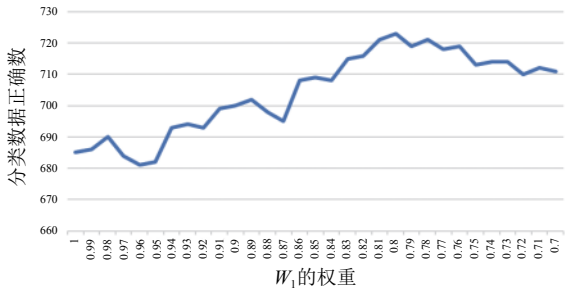

图 7 不同权重下的分类结果

虽然图像数据有这些不足, 但是图像数据作为整个数据的一部分, 仍然有着不容忽视的作用. 当文本信息内容出现缺失或两个类别特征出现冲突时, 将图像信息作为辅助信息可以得到正确的分类结果.

训练好的模型的精度如表 2 所示.

表 2 不同模型的分类精度

模型	平均分类精度 (%)
CNN 神经网络模型 (仅图片)	61
LSTM 神经网络模型 (仅文本)	68
FC-LSTM 神经网络模型 (仅文本)	68
LSTM-CNN 多模态学习模型 (图像和文本)	72

根据表 2 比较各神经网络模型. CNN 模型和 LSTM 模型对比可知, 文本信息的可靠度和质量要高于图像信息. 对比 LSTM 和 FC-LSTM 可知, FC-LSTM 的全连接层并不会对分类结果构成影响. 结合文本信息和图像信息的 LSTM-CNN 与 LSTM 模型 (文本) 对比准确率提高了 4%, 与 CNN 模型 (图像) 对比准确率提高了 11%.

综上所述, 结合文本和图片信息的特征对网络数据信息进行分类, 较原来只是通过单一类型的数据进行分类, 准确率有了一定的提高. 对一个含有图像和文本的舆情信息而言, 根据数据集包含不同类型数据的特点, 图像和文本信息扮演的角色和重要性也各不相同. 实验结果证明在本文数据集中, 文本数据相比图像数据无论是信息承载量或信息质量都更为出色. 但是文本和图像数据都是不可或缺的一部分. 本文通过根据它们的重要性, 实现数据特征的融合及整体数据的分类. 一方面, 考虑到了文本信息的重要性, 尽量减小图像对文本信息分类结果造成的影响. 另一方面, 当文本信息出现不足时, 通过图像数据来对文本信息进行补充, 最终达到了更好的分类效果.

5 总结与展望

本文针对现在网络上舆情数据信息分布的特点和状况, 提出了基于异构数据的舆情分类方法. 与传统的只是针对单一类型数据进行分类的方法不同, 本文考虑到舆情数据的特点对不同类型的网络舆情数据进行特征提取, 通过融合后的特征进行分类, 同时这种方法最大限度的考虑到了各类数据中的有效信息和各类数据的不同特性, 据此可以使用不同的神经网络模型来完成特征提取, 使得数据分类的结果更加准确.

随着网络的发展例如像微博、微信等新媒体已经渐渐兴起并壮大, 包含多种类型数据的舆情信息已经成为一种常态, 网络上的数据随之必然更为复杂. 如何更好地利用不同类型的数据, 并针对这样的数据进行综合的处理和考虑, 必然是未来的趋势.

参考文献

1 第 41 次《中国互联网络发展状况统计报告》发布. 中国广播, 2018, (3): 96.
2 钮成明, 詹国华, 李志华. 基于深度神经网络的微博文本情感倾向性分析. 计算机系统应用, 2018, 27(11): 205–210.
3 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究. 计算机系统应用, 2018, 27(5): 209–215.
4 梁吉业, 冯晨娇, 宋鹏. 大数据相关分析综述. 计算机学报, 2016, 39(1): 1–18.
5 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.
6 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference On Learning Representations. San Diego, CA. 2015.
7 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
8 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada. 2014. 3104–3112.

9 Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015. 1556–1566.

10 Ngiam J, Khosla A, Kim M, *et al*. Multimodal deep learning. Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, WA, USA. 2009. 689–696.

11 Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1–9.

12 冯方向. 基于深度学习的跨模态检索研究[博士学位论文]. 北京: 北京邮电大学, 2015

13 Huiskes MJ, Thomee B, Lew MS. New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative. International Conference on Multimedia Information Retrieval. Philadelphia, PA, USA. 2010. 527–536.

14 Guillaumin M, Verbeek J, Schmid C. Multimodal semi-supervised learning for image classification. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 902–909.

15 Xing EP, Yan R, Hauptmann AG. Mining associated text and images with dual-wing harmoniums. Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence. Edinburgh, Scotland. 2005. 633–641.