

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):

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

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.

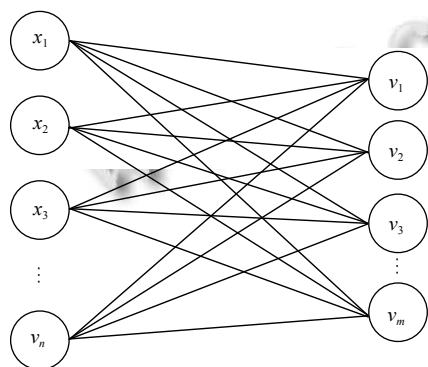


图3 全连接层操作

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:

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

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.

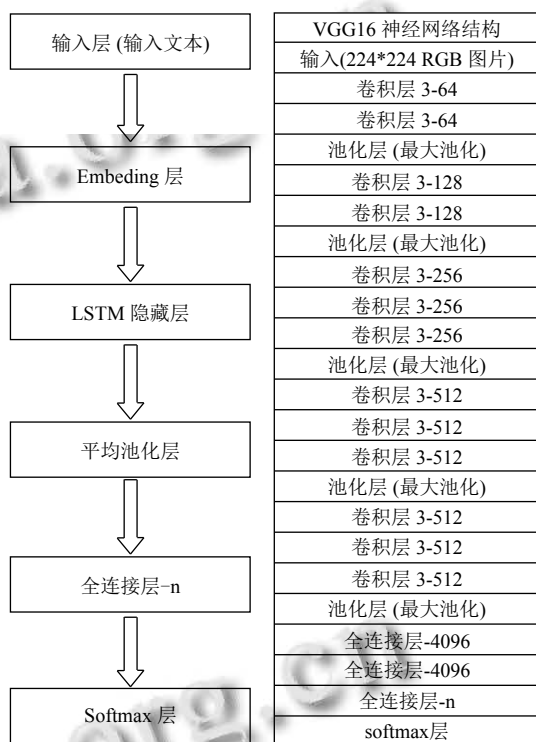


图4 FC-LSTM 模型和 CNN 模型

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.

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

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.

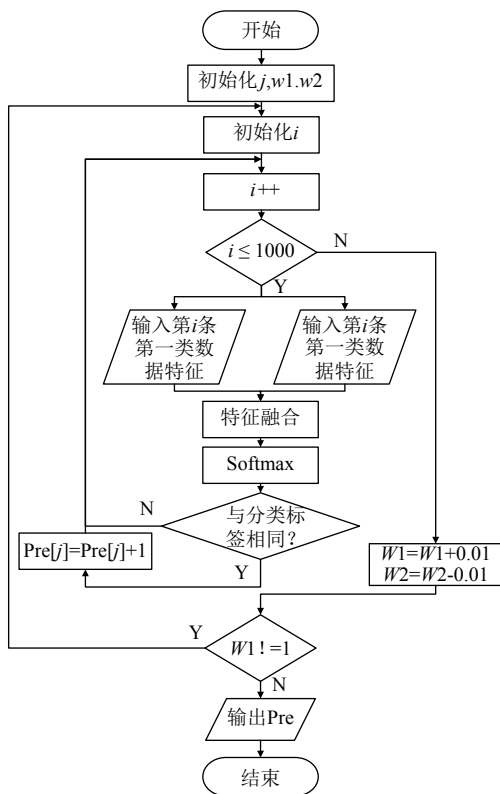


图5 权重获取流程图

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.

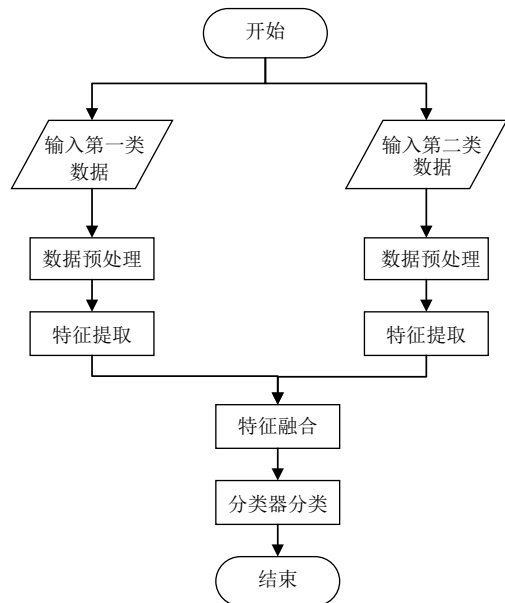


图6 舆情分类模型

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.

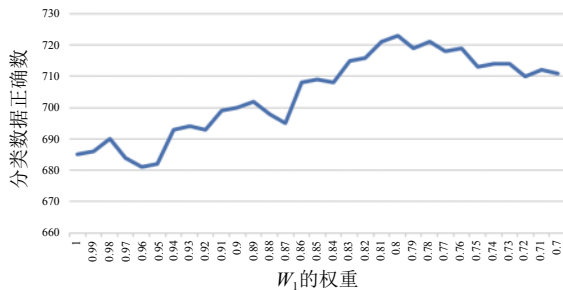


图7 不同权重下的分类结果

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 不同模型的分类型精度

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. 对于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

- 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.
- 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.
- 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.
- 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.
- 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.
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference On Learning Representations. San Diego, CA. 2015.
- Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
- 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.