Active learning and ExpBERT-based text classification for Crisis Scenarios

Introduction

Abstract

文本分类模型使用自然语言处理来分析预训练后的文本并分配标签。在洪水和地震等危机场景中，文本分类器可用于识别紧急信息，并将社交媒体上的信息类别转发给相关机构。然而，文本分类器的有效性在很大程度上依赖于大量的训练数据，这可能是稀缺的和难以获得的[1]。此外，训练大量的标记好的且不具代表性的数据会延迟模型的响应时间以及影响模型的准确性。识别可采取行动的信息类型，如搜索和救援请求，仍然具有挑战性。

将人类判断和解释集成到分类器训练过程中，使用针对多分类问题的主动学习策略来参与到人在循环中（HITL）过程，可以作为解决这些限制的有效方法。与之前的技术，即具有自然语言解释的表示工程（ExpBERT）相结合，该技术不仅依赖输入文本本身，而且还利用了自然语言解释（NLE）来改善表示工程，以提高分类器的性能[2]。此论文的主要目标是通过在分类过程中增加少量具有代表性的样例，与人类提供的解释结合，达到基于expbert的文本分类器的最大可实现精度[3]。

此论文探讨主动学习对基于ExpBERT文本分类问题的有效性。在迭代过程中，使用基于不确定性和多样性等抽样策略来查询具有代表意义的未的标注的实例，注释器接受并处理抽样后的这些实例。原有少量的标注过数据结合新处理的有效数据将用于模型的重训练，最终使用了少量数据在少量迭代后的模型，可以达到经过大量数据训练的后的模型的性能。本实验将建立比较组来观察有效性，比较不同抽样方法的影响，以及使用贝叶斯神经网络结合Monte Carlo的先进主动学习系统在 CrisisNLP dataset上的性能。

本文的主要结论是：

主动学习中的模型使用少量标注数据以及少量迭代就达到了全数据量下的模型性能。

多样性抽样以及不确定性抽样在收集解释并训练模型时有效性较高。

Bayesian Active Learning by Disagreement结合了MC dropout抽样比使用标准神经网络下的主动学习更有效。

1 Introduction

本章首先介绍了紧急响应系统的背景，根据背景生成了实验的动机，强调了此次研究的重要性。除此以外，将对工作进行概括并对比传统方式的局限。最后简要概括了主要目标以及挑战。

* 1. 背景

针对社交平台的紧急时间响应系统的重点主要集中在创建更好的文本分类算法来从少量数据中学习。然而，获得有用的注释数据集可能被证明是困难的[4]。针对文本分类器的预训练模型ExpBERT不仅接收分类标签，还接收对应的解释。解释包括哪些关键词导致了这种分类。ExpBERT使用这些信息融入到模型的训练中，使模型可以从更深层次的语义信息中学习来提高模型的泛化能力。因此，使用具有代表性的数据并了解关键词句后进行注释才可以引导模型精确分类。

但是，一般的社交平台数据的吞吐量无法完成对上万条数据注释。因此，许多能够低成本实现高信息价值数据抽样的主动学习形式在分类项目中得到了广泛的应用[5,6]。主动学习是解决这些问题的一种有效方法，它选择了少量具有高信息量的未注释样本，这些样本供在循环中的人类专家来注释[7,8]。得到有效注释后的新数据决定了下次迭代模型的性能，因此，使用不同的acquisition function查询信息最丰富的新实例可能是主动学习中最流行的方法。所以，查询策略自然成为了主动学习领域的研究重心。

同时，神经网络文本分类器往往不适应早期的不确定性抽样[9, 10]。因为，神经网络权重参数是固定的数值，导致了模型对正确以及错误数据的预测过于自信。贝叶斯神经网络通过引入概率分布，使得权重不再是确定的数值，因此权重参数是不确定的。使用变分推理和贝叶斯神经网络，模型的预测结果也将包含这种不确定性，展现了概率分布，这可以缓解模型的过分自信问题。

* 1. 动机

在上述背景下，本研究重点关注文本分类系统的准确性，代表性以及健全性。

* + 1. 准确性

由推特所代表的社交网络的统一特征是庞大的数据量结合互联网语法。最终目标是在捕捉紧急需求时，需要对紧急消息类别进行精准分类同时派遣相关部门来解决需求。然而，传统的文本分类系统是经过大量监督学习完成的性能提升，成本较高反应时间较慢。具有不同信息量的文本有着随机分布，而信息量的不同会极大地影响模型的准确性，捕获和识别重要应急信息的结果大多较差。因此，使用主动学习模型来应用少量数据并提高准确性是很重要的。

* + 1. 代表性

在大规模的无标记数据中提取丰富的特征表示的能力可以减少紧急事件多分类主动学习问题中的大部分偏差。通过主动选择样本进行标记，模型可以在每次迭代中选择最有价值和最有代表性的样本，以提高模型的性能。因此选择有效的acquisition function去提取具有代表性的数据对性能的影响是最大的，也是最值得改进的。

* + 1. 健全性

原始模型在处理文本分类任务时，往往表现出过度自信的倾向。在模型输入未知内容时，导致其作出错误且不可靠的预测。而在许多实际应用中，如灾难响应、医疗判断等，这种错误的预测可能会带来严重的后果。因此，对模型本身的不确定性进行有效的估计以减少其过度自信的倾向，成为了一个重要的研究课题。另一方面，我们可以在不引入过多计算负担的情况下，提高模型的健全性。这对于神经网络在需要高度准确和可靠预测的场景中，具有非常高的价值。

* 1. 实验方法

结合背景以及动机，此论文最终设计了一个基于池的人在循环中的主动学习文本分类系统。同传统文本分类模型不同的是，我们用少量的标注数据去训练，用训练后的模型结合不同的acquisition functions筛选的用例。然后，使用人类注释者在循环中对提取的样例进行观察并分析关键特征，对提取的样例给予一个标签以及统一的解释。这些新用例将加入到原数据中，重复上面循环，直到性能达到全数据性能。

为了观察性能在不同查询策略下的表现，实现中将采用基于Least Confidence的不确定性抽样，sentiment diversity sampling以及作为基准的随机采样方式。具体的acquisition function将在第二节进行细节描述。除此以外，在每次循环中，拼接了解释的文本都将使用ExpBERT去预训练生成文本的深层次语义表示，并将预训练后的结果作为神经网络文本分类器的输入。

最后，将对传统神经网络模型进行结构上的改变，使用贝叶斯神经网络结合Variational Inference来学习权重的后验分布，来减少过分自信以及不健全性。由于原始Bayesian Active Learning by Disagreement （BALD）算法中求熵的计算成本过高，所以升级了BALD算法，利用MC dropout的同时使用Least Confidence来选择模型预测概率最均匀（即没有预测概率特别高的类别）的样本。

1.3.1目标

基于上述实验描述，我们的目标总结如下。

•开发一个基于池的人在循环中的具有解释的主动学习框架，适当地整合收购功能。

•构建注释器仿真流程，设计循环中止标准（例如达到全数据下模型性能）。

•生成基线（无主动学习下全数据的模型性能或无标注的随机抽样），研究主动学习是否可以提高文本分类模型的性能。比较不同查询算法抽样的影响。

•探索了先进的ExpBERT贝叶斯文本分类体系结构在嘈杂环境下的鲁棒性。例如，提供噪声解释来代替模拟系统产生的解释。比较贝叶斯神经网络是否在文本分类问题中比传统神经网络表现更好。

1.3.2 挑战

这个项目的主要挑战是不同的查询策略是否有很大的差别，以及如何改进原始ExpBERT文本多分类器的初始化框架以及训练框架来达到计算量的减少。

•值得考虑的是，主动学习文本分类系统存在多种查询策略，查询策略中不同算法的应用极大地影响了主动学习框架的性能。查询策略的选择需要从时间复杂性和代表性两方面来考虑。

•其次是贝叶斯神经网络在推断算法执行中以及利用MC dropout时的时间消耗可能会引起计算量的增加，因此如何在每次循环中减少计算量来提高反应效率时间是一个挑战。

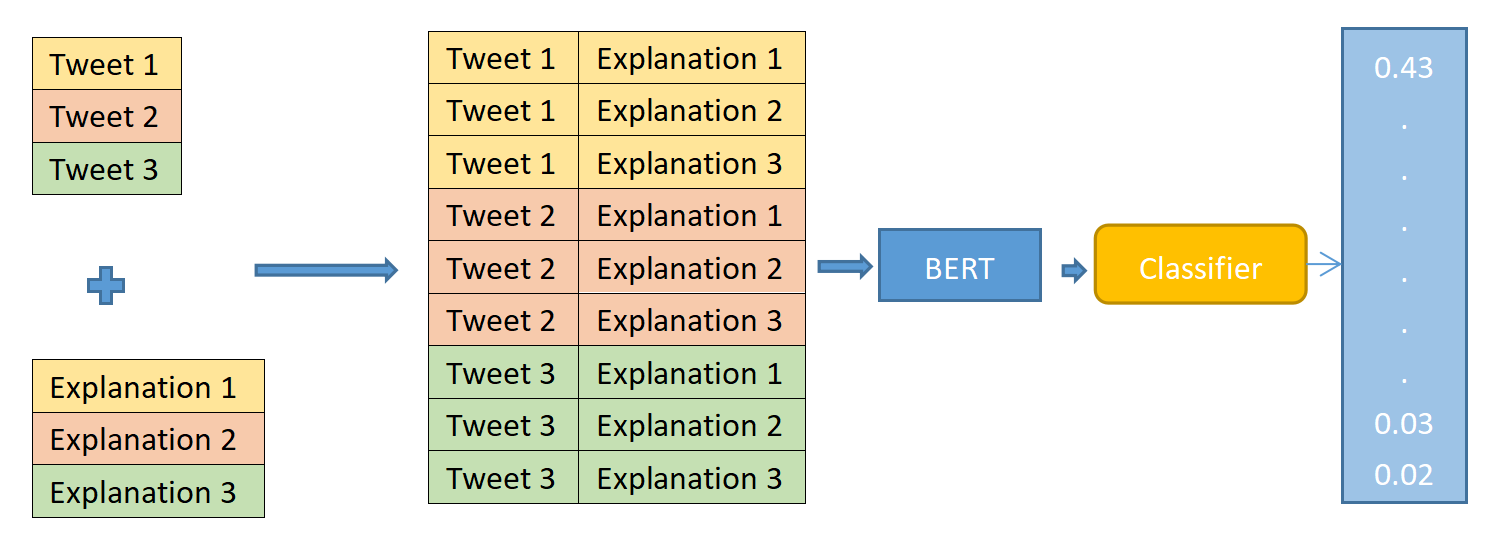
•最后对于框架设计，如何适当地设置采样的比例，迭代的次数以及每次迭代提供解释的数量等需要研究来达到满意的性能。

2 背景

本章介绍了实现利用主动学习提升基于ExpBERT文本分类器性能的关键技术，首先介绍了实验使用的预训练模型（2.1 section)，这个模型将应用在人在循环中（2.2 section and 2.4 section）。介绍了基于主动学习及其应用的查询策略，这是本实验的重点。2.5节介绍了贝叶斯理论下模型的升级以及主动学习方法的改进技术，探讨将贝叶斯理论应用在主动学习的可行性。最后描述了多模型整合方式（2.6 section），以及利用Openai模型模拟人在循环中的技术。

2.1 Representation Engineering with Natural Language Explanations (ExpBERT)

ExpBERT模型提出了一种增强语言模型可解释性和知识融合能力的方法。其将人类提供的固定解释与推文相结合，从这些解释中学习以改善模型的性能[2]。图 1 直观地展示了如何将带有解释的样本与 BERT 模型相结合。首先，将每条推文与一组预先编写的解释完全连接起来，生成一个3×3的集合。这些解释与特定推文无关，每个推文都会连接相同数量的解释。预处理后的文本和解释被输入微调后的BERT模型，该模型为每个输入样本生成一个特征向量，代表长度为786的整个输入。然后将推文和解释的特征向量连接起来，形成一个大小为768 \* E的模型，其中E是解释的数量，并将其作为分类器模型训练和预测的输入数据。本论文将使用该模型作为基础预训练模型，并将其集成和初始化到"人在回路"系统中，因为该模型可以处理带有解释的实例向量。



2.2 Human in the loop

传统的自然语言处理流程不是为了充分利用人类反馈而设计的。与此不同，人在回路（human-in-the-loop）作为交互系统的关键组成部分，可以通过模拟人在循环中的角色来揭示模型的缺陷，这些缺陷在真实世界测试之前可能并不显而易见[11]。Godbole等人（2004）[12]采用支持向量机（SVM）主动学习，扩展了一个文本分类器。他们巧妙地将人类在特征工程、术语选择、文档标签等方面的输入纳入模型中，以做出统计上合理的判断。这种新型的人与机器学习算法之间的互动方式被称为人在回路（human-in-the-loop）机器学习[13]。在此方法中，根据不同的需求和场景，人和机器之间不同类型的协作，可以确定不同的HITL机器学习方案[14]。

主动学习（AL） (Settles 2009)[15]：主动学习的关键是系统保持着控制模型学习的过程，虽然人类作为媒介来参与对未标记的数据的注释，但是人类无法根据偏好选择无标签数据。AL将作为优化框架应用在本论文中，在2.2节中将详细介绍主动学习的应用。

交互式机器学习（IML）（Amershi et al. 2014）[16]是一种人类与系统保持紧密交互的方法。在此背景下，主动学习（AL）和交互式机器学习（IML）之间存在一些关键区别。Dudley和Kristensson（2018）[17]指出，虽然AL和IML都着重于由用户选择新的数据点进行标记，它们的驱动方式却不同。具体来说，在AL中，选择是由模型本身推动的，而在IML中，选择则是由用户推动的。由于IML是在AL的基础上发展起来的，因此它们有一些共同的缺点。然而，IML还具有一些特有的问题，特别是与人机交互技术（HCI）的复杂性有关。这些特有的问题需要更多专门的研究来解决（Michael et al. (2020)）[18]。

机器教学（MT）（Ramos et al. 2020）[19]是一种通过人类教师的指导来训练机器学习模型的方法。具体来说，人类教师会界定他们打算传递给模型的知识，强调教师在学习过程中的主动参与和指导。Devidze et al. (2020)[20]指出，与主动学习相比，MT更依赖于教师的专业知识，同时在样本选择和处理复杂任务方面的灵活性较差。因此，根据具体的需求选择适当的学习方法显得尤为重要。

2.3 Active Learning

作为HITL中最热门的学习方案，主动学习系统通过向未标记的实例提出问题并由专家（例如人工标注者）进行标记，试图克服标记瓶颈[21]。简单来说，主动学习致力于寻找最具信息价值的未标记原始数据，并将其交给注释者进行标记，这一过程更贴近实际从源数据中提取信息的场景。注释者标记这些源数据，将已经标记的实例加入到模型的训练过程。这样，主动学习能够通过使用更少的标注数据来实现全量数据下的性能表现，从而克服了标签瓶颈问题。在此项目中实例数量庞大、噪声较多、全量训练任务沉重，且准确率不高，主动学习方法就显得尤为适用。因此主动学习适用于此项目去专注于那些可能增加准确率的实例，减少无相关实例，提高紧急反应系统效率以及准确率。

2.3.1 The AL process and scenarios

主动学习的进程如图一所示，图一为基于池的主动学习的进程。首先，利用已标记的实例集L，对机器学习模型进行初始化训练。然后，模型对未标记的样本集进行特征提取后，根据查询策略选择具有代表性的未标记样本，提供给人类标注者。人工标注者对所选样本进行标注，标注后的样本将从未标记样本集U移除，加入到已标记样本集L中。通过反复上述步骤，已标记样本的数量逐渐增加，模型性能不断提高。最后一旦达到终止标准（例如达到全数据量下的模型性能），主动学习过程将中止。

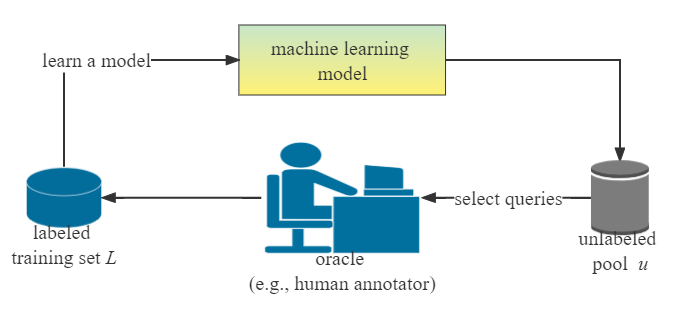


Figure 2: 基于池的主动学习过程

根据Settles (2009) [15]的分类，主动学习主要有三种场景设置：membership query synthesis[22, 23]，stream-based selective sampling[24]，以及pool-based active learning[25]。

Membership query synthesis: 这种场景的查询策略是由模型生成的。即机器学习模型可以请求标记任何未标记的实例，其并未在潜在的自然分布中取样[15]。对于比较绝对的问题领域，例如对绝对坐标采用回归预测的任务中，有效的查询合成通常是有效的。因为它可以解析简单d的数据分布，构造合理数据供人类标注。但是, Baum and Lang (1992)[26]的研究中可以看出对于自然语言处理这类复杂任务中，模型可能会产生不易理解的文本串，导致人类无法对这些混乱文字进行判断。在深度主动学习背景中，可以通过生成对抗网络（GANs）用于数据增强来解决成员查询合成的场景，因为GANs能够生成合理度高的实例[27]。

Stream-Based Selective Sampling：此场景设置相比于Membership query synthesis的不同之处在于，此方式可以不受输入分布的影响。基于流的抽样是一次从实际的分布中抽取一个未标记实例，模型通过“信息性度量”或“查询策略“来决定是否请求实例的标签，相当于一种带有偏见的随机抽样[28]。

pool-based AL: 作为普遍用于主动学习的设置，pool-based AL与基于流的采样的区别在于，pool-based AL在选择最佳查询之前采取贪婪机制，对整个数据集进行比较，但是后者单独接收数据进行评估。但是，当数据集非常大时，由于需要对全部数据进行评估，可能导致计算耗时。除此以外，此方法适用于人工标注成本较大的场景，通过选择最有价值的数据进行标注，从而最大化标注投入的价值。

2.4 Acquisition Functions

本节考虑主动学习中的各种acquisition function，在论文实现部分中将使用基于ExpBERT的文本多分类模型结合不同acquisition functions实现性能的提升。

随机抽样

随机抽样随机选择实例，既不基于预测，也不基于数据以及模型，通常被用作任务的基准。因为，它可能忽略了潜在地信息，从而降低了学习效率。在这种情况下，随机抽样将作为底线与下面提到的更复杂的策略对比，特别是当标记池过于庞大时[29]。

不确定性抽样

不确定性抽样是一种特殊的主动学习策略，通过查询最难以判定其分类的实例，进行标注后提高模型判定准确度。在用概率模型做二分类问题中，这种实例的后验正概率最接近0.5[30]。但是对于更复杂多标签分类问题，将使用基于熵的方法。概率分布越均匀，熵越大，随机变量的不确定性越高，信息量越大。当概率集中在几个数据点，表示不确定性越低，信息量越小。



表示在分类的概率分布。

在文本分类领域，对于不确定性的测量，常用一种替代方式，即least confident[31]：



表示可能性最大类标签。对于二值文本分类，该方法等价于基于熵的算法有效性。考虑那些模型预测概率最大但是可信度较低的样本。具体方式是通过选择最大概率最小的样本进行标注。

除了上面两种较为常用的测量方式之外，（Munro 2020）[13]提到了margin of confidence 以及ratio of confidence这两种方式，前者是两个最自信的预测之间的差异，后者则是这两个预测的比率。



是最自信，是次自信。

Semantic-based diversity sampling：

Peng, Hao, et al.（2023）[30]提出了一种基于语义的多样抽样方式，该方式可应用在文本分类中。与采用不确定性抽样的测量置信度的方式的不同之处在于，基于语义的多样抽样的方式是利用欧氏距离在语义上消除文本样本的冗余。确保在后续过程中提供给模型（学习者）更丰富的，重复度更低的样本。此抽象方式使用（Sener和Saveravarese，2017）[31]的贪婪k中心算法进行聚类操作。数据集包含个无标签的文本，并将分成个批次，每个样本集包含m个实例。是数据集编码的结果。首先选择vectors从中去初始化聚类，这里examples将被认为是聚类中心。然后适用k中心算法searches from ,这是一个包括不在中的成员的集合。寻找距离现有聚类中心最远的样本。通过这种方式，可以逐渐构建一个代表各种不同文本类型的样本集合，从而确保模型在学习过程中接触到尽可能多样化的内容。算法公式如下：



Where：



随后，通过将现有的聚类更新为。经过循环执并将输出结果合并到中。中的所有文本实例会收敛成一个核心集合，在语义空间中最好地体现并概括数据集。

Bayesian active learning by Disagreement：

BALD（Bayesian Active Learning by Disagreement）是文本分类中的一种主动学习策略，它结合了贝叶斯推断和不确定性度量来选择信息丰富的样本。在BALD中，使用如贝叶斯神经网络等贝叶斯模型来估算每个样本的类别概率分布，通过计算预测间的不一致性来衡量不确定性（Houlsby, N（2011））[33]。一个关键方面是BALD采用了MC（Monte Carlo）dropout技术。MC dropout是一种训练过程中的正则化方法，通过在训练过程中随机关闭一些神经元，来模拟贝叶斯网络的后验分布[这个地方加上文献]。从而可以通过对每个样本的多次预测并计算模型的预测之间的不一致性来选择具有最高不确定性的样本进行标注。最后，BALD通过比较每个样本的BALD值，选出具有最大信息增益的样本进行标注，并将其加入已标注的训练集。除此以外，BALD可以和前文提到的Least Confidence结合来选择模型预测概率最均匀（即没有预测概率特别高的类别）的样本。

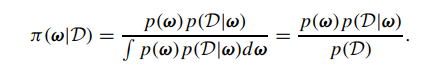
2.5 Active Learning for (Bayesian) Neural Networks

在现代的机器学习和深度学习应用中，贝叶斯神经网络以其灵活的推理方式成为一个关键组件。通过引入权重和预测的不确定性，贝叶斯神经网络提供了一种丰富和灵活的建模框架。这不仅增强了模型的可靠性，还使模型能够更好地泛化、适应噪声和异常值，并在诸如主动学习和增量学习等先进场景中工作。在主动学习方面，贝叶斯推理允许模型从不确定性中受益，优先选择信息量最大的样本进行标注，从而实现更有效的学习。本部分将探讨贝叶斯方法下对模型框架，包括主动学习框架的优化。

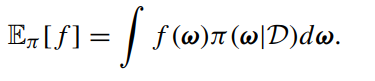
2.5.1 贝叶斯神经网络

与传统的神经网络不同之处在于，贝叶斯神经网络中的权值不是固定不变的。该模型的思想是将神经网络和随机建模的优势结合起来，可以为预测提供概率上的保证[文献]。这里的权值是基于我们观察到的来推断的，是一个逆概率问题，需要通过贝叶斯定理来解决[文献]。这里权重w代表的是无法直接观察分布的不可见变量，无法直接观察真实分布，贝叶斯允许我们得到模型参数在已观察到的数据条件下的分布p(ω|D)，即后验分布。

同时，p(D|w) 似然function在多分类问题里将神经网络的预测（经过Log Softmax转换）与实际观测类别标签相结合，来量化他们之间的一致性。在确定似然项以及p(w)先验分布后，便可根据贝叶斯理论计算后验分布，



得到后验分布后，将使用边缘化方法进行预测：



2.5.2 推理方法：

推理方法

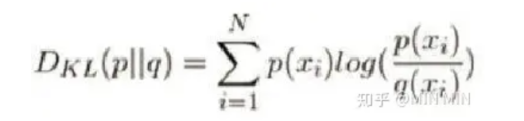
适用于解决这个归一化因子即证据项在高维中难以处理的问题。因此需要采用近似技术来获得后验分布。除此以外，贝叶斯推理方法还可以根据模型预测的不确定性来寻找最有可能提高模型性能的样本。主流的BNN有有两种方式，一种是基于采样的方式（马尔可夫链蒙特卡洛），或者是变分推断[文献]。

Markov Chain Monte Carlo (MCMC)

MCMC是目前最流行的基于采样来估计后验分布的方式，主要思想是构建前后样本的依赖关系[78]，进而模拟分布。但是由于连续样本可能出现自相关的情况，则需要更加庞大的样本集来抽样去获得近似独立样本，这样做会增加计算成本。因此，对于BNN来说，Metropolis-Hasting算法更加有效[文献]。通过与分布成比例的函数，省去获取精确概率分布的过程，并使用接受-拒绝机制来完成对候选样本的选取。如果样本比前一个数据点出现概率低则拒绝样本。除此以外，汉密尔顿蒙特卡罗算法（HMC）[81]是更加改进的方式，通过更新提议来拒绝少量样本的同时减少相关性来缩短burn-in时间。

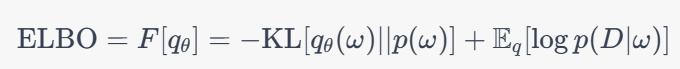
Variational Inference

变分推理是相较于MCMC的一种提高计算速度的方式，通过优化参数，在VI定义的参数化数族中找到最优近似分布。主要思想是结合KL散度来定义分布的差距，当两个分布的差异越大时，KL散度就越大：



P(x)是目标分布，q(x)是去匹配的分布，如果这两个分布完全匹配则。

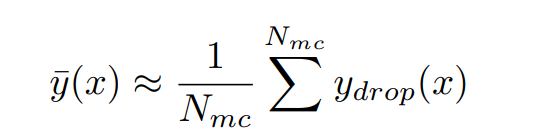
但是，真实后验分布是未知的，直接优化KL散度并不简单，在此论文中将KL公式进行转化为计算Evidence Lower Bound（ELBO），最小化KL散度等于最大化ELBO[文献]。



这里的ELBO（证据下界）包括：变分分布和真实先验之间的KL散度的负值，以及给定模型参数的数据似然的期望。

MC Dropout

MC dropout本质是通过概率性停用部分神经元以防止过拟合的正则化方法，在推断过程中，通过有限次（T次）对神经网络进行输入，使用dropout来在预测阶段产生不同的结果[文献]。因此，这些不同预测结果进行不确定性量化后，可以用来估计模型对未标记样本的不确定性[文献]，并应用在主动学习的Bayesian Active Learning by Disagreement(BALD)抽样策略中，来选择不确定性最大的未标注数据。MC dropout的推断结果通过以下方式得出：



其中，ydrop是dropout网络的输出的不同结果，x是网络的输入，Nmc是获取分布所需的采样次数。

Small-Text Library

本项目中的文本分类器倾向于专注于一个模型，因此可能错过其他可行模型的应用。然而，切换模型和主动学习策略的时间成本，以及代码的冗余性，将会显著影响实验的进展。小型文本库整合了scikit-learn、transformers和PyTorch等可以在Python环境中应用的常用库[37]。基于池的主动学习的文本分类架构连接了查询策略、分类器和中止策略的接口。它不仅为文本分类工作提供了先进的主动学习框架，而且提供了一系列的分类器和查询策略组件，以便于可以混合搭配用于实验和应用的主动学习任务，使主动学习在Python生态系统中易于实施。与最常用的ModAL库[38]相比，Small-Text提供了更灵活的定制服务，前者更专注于模型集成和查询策略的选择。然而，这种包罗万象的方法并不一定会改善基于ExpBERT的项目的性能，因为性能差距会因情况而异。

ChatGPT annotator simulation explanation generation

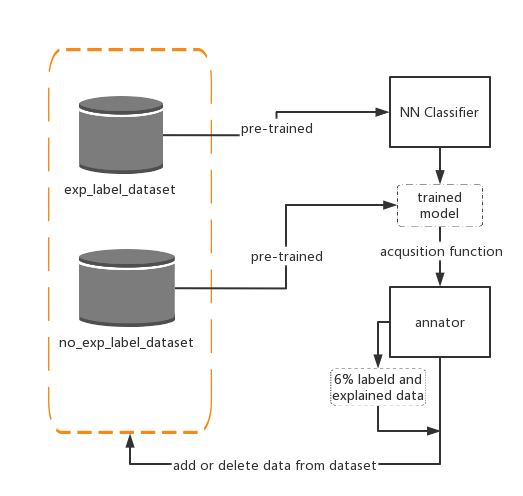
ChatGPT（生成预训练变换器）[39]是OpenAI团队在2018年开发的一种预训练语言模型。基本算法是Transformer，这是一种基于自我注意机制的深度神经网络结构，具有强大的序列建模和表示学习过程。通过预训练和微调，该模型可以分析和生成自然语言文本，并在多个场景中有所帮助，例如自动回答、智能客户支持、语言翻译等。因此，作为一种节省时间和高效的文本解释系统，该项目将使用此模型作为基于ExpBERT的主动学习注释器。主要任务是为ExpBERT生成合适的解释，并将它们提供给模型以帮助提高分类性能。在此配置中的主动学习通过利用ChatGPT强大的生成模型[40]根据输入文本生成人类可读的解释。常见的方法是使用ChatGPT创建多个解释，然后选择其中一个或多个作为解释实例传递给ExpBERT模型，这样每个标记的实例不再对应于相同的全连接解释，而是对应于专门针对该实例量身定制的解释，这样更为相关。

设计（7页）

1. 整体流程

整体实验框架基于第二章提到的基于池的主动学习技术而搭建的，从固定的未标记的样本池中进行抽样，利用训练好的模型来评估样本。与传统主动学习不同的是，我们改变了对数据集的要求，引入了带标签及解释（exp\_label\_dataset）和无标签及解释(no\_exp\_label\_dataset)的两个数据集。因为为了充分利用ExpBERT预训练模型，每轮标注不仅提供标签，还会增加解释到默认解释集中。

初始训练阶段使用占全数据量百分之二十的带标签数据集，每个样本连接了默认的9个解释（标签描述）进行第一轮训练。第一轮训练后的模型将对no\_exp\_label\_dataset数据集进行评估，并通过不确定性或多样性计算每个样本的信息量，然后选择信息量最大的数据进行标注。整体的主动学习的架构设计及流程如下图所示。



图中的标注过程中注释者会总结出抽样文本的关键句作为解释，并提供每个文本的分类标签。增加后的解释数量将大于默认解释数量。使用acquisition function进行量化后，我们将抽样出的数据进行全标注后加入到exp\_label\_dataset中，形成下一轮迭代所需的数据集，这些数据集均会被ExpBERT模型预处理后输入到神经网络文本分类器中。然后把标注了的6%的数据从no\_exp\_label\_dataset数据集中移除。每轮exp\_label\_dataset的数据量会增加百分之六，解释会增加一个或多个，反复执行上述过程，直到验证集的评估分数接近或超过全数据量下的评估分数。

1. Embedding设计（增加未标注实例时的embedding）
2. 分类器模型设计

根据上面的框架图，我们采用神经网络作为分类器模型。评估阶段将对比传统神经网络分类器以及贝叶斯神经网络分类器结合主动学习技术的表现，此实验将对传统的神经网络进行结构上的调整，以及对训练和预测方式的改变。最终的模型结构将采用PyTorch[到时候插入文献]以及Pyro [到时候插入文献]library 来搭建。Pyro库是更简洁的扩展版PyTorch，更专注复杂模型的推断任务例如贝叶斯推断，不仅可以与PyTorch集成，还可以声明多种的概率分布来简化深度学习流程[加上文献]。

普通神经网络

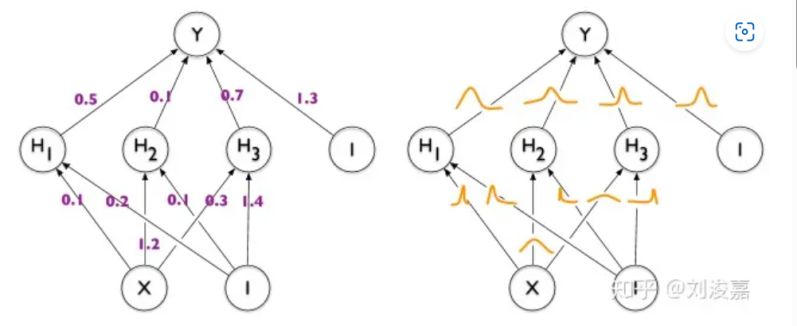
在模型结构上，普通神经网络使用了一层的隐藏层以及100个神经单元，输出特征数量即‘output\_size’等于9（标签种类），向前传播的设计可以表示为下面的这个公式，wi和bi分别是线性层i的权重和偏置，中间经过激活函数后得到模型的输出即y。



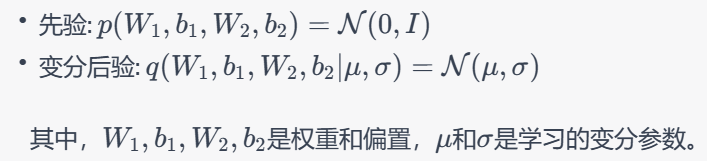
在训练方法上，普通神经网络使用了标准的损失计算方法（例如交叉熵[文献]），除此之外，前向传播、损失计算、反向传播和优化的这些步骤是分开执行的，仅使用了PyTorch的标准方法。

贝叶斯神经网络

在模型结构上，与普通神经网络结构相同的是他们有相同数量的隐藏层，以及激活函数，并具有大小为’input\_size’的输入。不同之处是贝叶斯神经网络使用Dropout正则化，在训练期间用’dropout\_prob’的概率来随机关闭部分输入。不同于传统神经网络的权重，在这里权重被视为有概率分布的随机变量，如下图所示。



结合第二章的描述，贝叶斯神经网络通过‘model’方式定义权重和偏置的先验，并描述了似然分布，公式如下所示。通过‘guide’方式找到逼近真实后验分布的变分分布。



在训练过程中，我们采用了与普通神经网络不同的方法。通过使用Pyro库和随机变分推断（SVI），我们合并了优化的各个步骤，而不是分开执行。SVI的目标是通过迭代最大化证据下界（ELBO）来逼近真实的后验分布。

在对验证集和测试集的预测方面，贝叶斯神经网络与传统网络的工作方式不同。它通过多次采样来生成多个预测结果，而不是仅依赖一个固定模型。这种方式增加了预测的多样性，并反映了模型的不确定性。最后，在确定最可能的类别标签时，我们采取了一种整合方法，即取所有预测结果的均值，然后使用`np.argmax`。这种方法可能提供更鲁棒的预测。

1. 抽样策略

我们在神经网络分类器中采用了四种不同的抽样策略来提升模型性能：随机抽样作为基准；基于"Least Confidence"的不确定性抽样；基于情绪的多样性抽样；以及基于MC dropout的贝叶斯主动学习异议抽样（Bayesian Active Learning by Disagreement, BALD）。这些策略各自通过不同的途径挖掘和利用数据的独特属性，作为主动学习框架的核心共同促进了模型学习的效率和准确性。通过对不同角度的抽样来发掘对模型性能提升最好的方式。

随机抽样：

抽样过程中，acquisition function将对接近占全数据集75%的未标注数据集进行评估，每次评估给出6%的未标注数据集交给注释者来进行标注。随机抽样来确保没有任何偏见的样本选择。通过设置随机种子直接对未标注数据集抽样，并未将初训练后的模型纳入考虑，在其影响下的模型性能将作为基线。

基于"Least Confidence"的不确定性抽样：

此论文采用了Least Confidence方式来计算不确定性，As described in第二章，对多标签文本分类问题Least Confidence常常作为计算熵的替代方式。在此实验中，通过将模型设置为评估模式，在模型给出的概率分布‘prob\_dist’中寻找最大的概率值作为不确定得分，计算标签数量num\_labels并将最有信心的预测进行标准化，标准化公式如下，normalized\_lc得到标准化的最小信心值。

normalized\_lc= (1 - max(prob\_dist)) \* (num\_labels / (num\_labels - 1))

通过最小信心值排序，选择前6%的的样本作为此不确定方法的输出。

基于情绪的多样性抽样：

BALD：

1. 注释过程

4.2 注释者

实现：

1. 实验环境
2. 数据集描述（train val test 怎么分呀等等）
3. 注释者模拟
4. 评估标准
5. 运行模块（使用了学校的那个计算模式）

@article{mullachery2018bayesian,

title={Bayesian neural networks},

author={Mullachery, Vikram and Khera, Aniruddh and Husain, Amir},

journal={arXiv preprint arXiv:1801.07710},

year={2018}

}

@article{goan2020bayesian,

title={Bayesian neural networks: An introduction and survey},

author={Goan, Ethan and Fookes, Clinton},

journal={Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018},

pages={45--87},

year={2020},

publisher={Springer}

}

@inproceedings{zhao2018empirical,

title={An empirical evaluation of Bayesian inference methods for Bayesian neural networks},

author={Zhao, Rui and Ji, Qiang},

booktitle={NIPS Bayesian Deep Learning (BDL) Workshop},

year={2018}

}

@article{valentin2020hands,

title={Hands-on Bayesian neural networks--a tutorial for deep learning users},

author={Valentin Jospin, Laurent and Buntine, Wray and Boussaid, Farid and Laga, Hamid and Bennamoun, Mohammed},

journal={arXiv e-prints},

pages={arXiv--2007},

year={2020}

}

1. R. Bardenet, A. Doucet, and C. Holmes, “On Markov Chain Monte Carlo methods for tall data,” J. Mach. Learn. Res., vol. 18, no. 1, pp. 1515–1557, Jan. 2017.

@article{chib1995understanding,

title={Understanding the metropolis-hastings algorithm},

author={Chib, Siddhartha and Greenberg, Edward},

journal={The american statistician},

volume={49},

number={4},

pages={327--335},

year={1995},

publisher={Taylor \& Francis}

}

@article{thomas2021learning,

title={Learning hamiltonian monte carlo in r},

author={Thomas, Samuel and Tu, Wanzhu},

journal={The American Statistician},

volume={75},

number={4},

pages={403--413},

year={2021},

publisher={Taylor \& Francis}

}

Hoffman, M. D., & Johnson, M. J. (2016, December). Elbo surgery: yet another way to carve up the variational evidence lower bound. In Workshop in Advances in Approximate Bayesian Inference, NIPS (Vol. 1, No. 2).

Myojin, T., Hashimoto, S., & Ishihama, N. (2020). Detecting uncertain BNN outputs on FPGA using Monte Carlo dropout sampling. In Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part II 29 (pp. 27-38). Springer International Publishing.

Tsymbalov, E., Panov, M., & Shapeev, A. (2018). Dropout-based active learning for regression. In Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7 (pp. 247-258). Springer International Publishing.