ICS 国际标准分类号
CCS 中国标准文献分类号

团　　体　　标　　准

T/CES XXX-XXXX

电力人工智能样本增广技术架构要求

Technical architecture requirements for sample augmentation in electric power artificial intelligence

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　发布

目　　次

前　言

本文件按照 GB/T1.1—2009《标准化工作导则 第 1 部分 标准的结构与编写》给出的规则起草。

本文件由中国电工技术学会提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国家电网有限公司大数据中心、国网信息通信产业集团有限公司、中国电力科学研究院有限公司、国网智能电网研究院有限公司、安徽继远软件有限公司、国网福建省电力有限公司。

本文件主要起草人：李强、赵峰、邱镇、陈振宇、李博、刘识、李炳森、黄晓光、张琳瑜、秦余、王晓东、张国梁、周逸平、苏勇、朱署光、李小宁、徐凡、郑碧煌、李黎、余江斌、郭庆、浦正国、薛濛、黄旭东、聂文萍、刘晓飞、刘健、李扬笛、林晨翔、谢炜。

本文件为首次发布。

电力人工智能样本增广技术架构要求

1 范围

本文件规定了电力人工智能样本增广技术架构、策略制定、增广算子编排等方面做出规范性要求。
本文件适用于电力人工智能图像类、文本文档类、语音类等样本增广。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.29—2006 信息技术 词汇 第29部分：人工智能 语音识别与合成
GB/T 5271.31—2006 信息技术 词汇 第31部分：人工智能 机器学习
DA/T 77-2019 纸质档案数字复制件光学字符识别OCR工作规范

3 术语和定义

下列术语和定义适用于本文件。

3.1
人工智能 artificial intelligence
研究人类智能活动的规律，构造具有一定智能的人工系统，研究如何让计算机去完成以往需要人的智力才能胜任的工作，也就是研究如何应用计算机的软硬件来模拟人类某些智能行为的基本理论、方法和技术。

3.2
噪声 noise
真实标记与数据集中的实际标记间的偏差。

3.3
语音识别 automatic speech recognition
让机器通过识别和理解过程把语音信号转变为相应的文本或命令的技术。

3.4
信噪比 signal-noise ratio
是一种用于度量信号与噪声强度之间关系的指标。

3.5
峰值信噪比 peak signal-noise ratio
指用于表示信号的最大可能功率与影响其表示的保真度的破坏噪声的功率之间的比率。

3.6
语音清晰度 perceptual evaluation of speech quality
指语音质量的知觉评估方法。

3.7
语音质量指标 mean opinion score
是一种用工衡量语音质量的指标。

3.8
样本增广算子 sample augmentation operator
指在机器学习和深度学习中用于扩充训练数据集的技术。

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。

4.1 符号

无

4.2 代号

无

4.3 缩略语

OCR：光学字符识别（Optical Character Recognition）
GAN：生成对抗网络（Generative Adversarial Network）

5 样本增广技术总体架构

电力人工智能样本增广技术总体架构包括：

a）样本增广技术：包括图像、文本、音频三种类型样本的主流增广技术；

b）样本增广技术要求：包括样本增广效果评价要求、样本增广策略制定要求和样本增广算子编排技术和功能要求。样本增广效果评价要求部分包含通用评价要求和图像、文本、音频三种类型样本的增广效果评价要求。

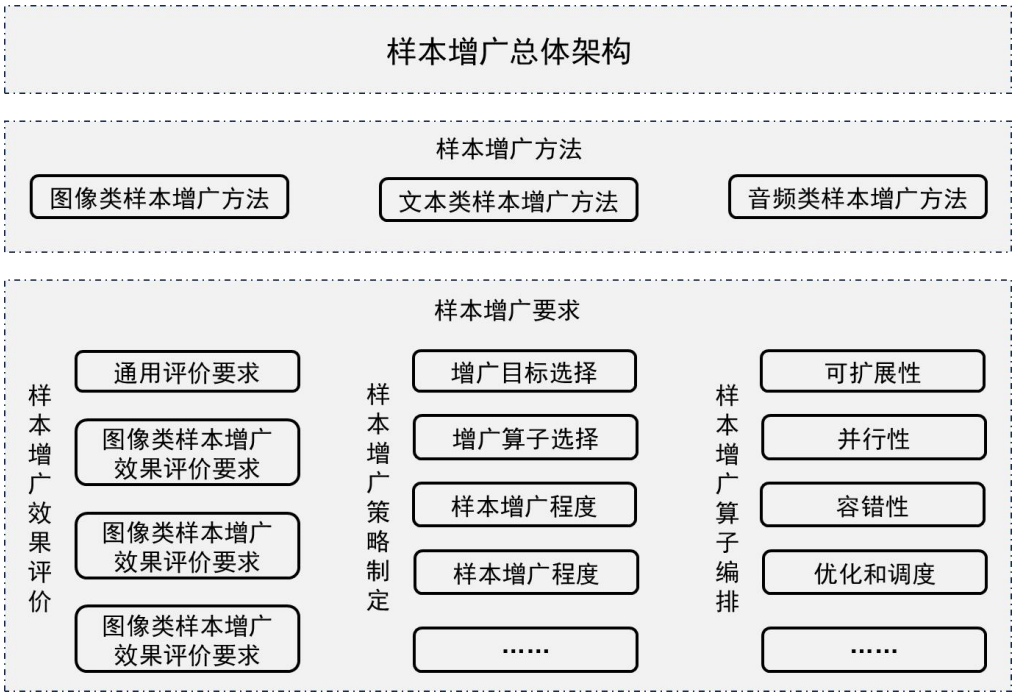

图 1 电力人工智能样本增广技术总体架构图

6 图像类样本增广技术

6.1 基本图像增广

基本图像增广是指在原有样本的基础上，通过进行一些较小的几何变换（例如翻转、平移、旋转或添加噪声等）或者色彩变换（例如亮度、对比度、饱和度或通道混洗等），以此来增加训练集的多样性。以下是一些常用的基本图像增广方法：

a）镜像翻转：将图像水平或垂直翻转，生成镜像的图像。这种增广方式可以在保持图像语义不变的情况下增加数据的多样性。

b）随机裁剪：随机从图像中裁剪出一个子区域作为新的图像或改变图像的尺寸和位置，增加样本的多样性。

c）旋转：对图像进行旋转操作，以一定的角度将图像进行旋转，增加样本在不同角度下的多样性。

d）缩放：对图像进行缩放操作，将图像放大或缩小，改变图像的尺寸，增加样本的尺度不变性。

e）平移：对图像进行平移操作，将图像沿水平或垂直方向移动，增加样本的平移不变性。

f）亮度调整：对图像的亮度进行调整，增加样本在不同亮度条件下的多样性。

对于一些小尺寸或小粒度的目标，例如杆塔上的螺母缺失识别，其缺少样本时可以采用上述方式。图像变换增广的主要特征是面向训练数据集的图像样本执行特定的图像变换操作，产生新的图像样本的标签信息与原始图像样本的标签信息保持一致。

6.2 混合图像增广

图像混合增广方法通过使用数据中的多个图像样本进行混合以合成新的图像样本。图像混合增广方法具备以下特点：

（1）增广过程中需要两个或两个以上图像样本参与；

（2）混合增广后生成的新的图像样本，其语义信息取决于多个参与增广样本的语义；

（3）增广后生成的图像样本往往不具备人眼视觉理解特性。

6.3 虚拟图像生成

虚拟图像生成增广是通过生成模型(主要以生成对抗网络为主)直接生成图像样本，并将生成的样本加入到训练集中，从而达到数据集增广的目标。虚拟图像生成增广通常使用生成对抗网络及其衍生网络作为图像样本的生成模型。

7 文本文档类样本增广技术

7.1 标签无关增广方法

标签无关增广方法是指不需要提供数据标签、任务需求等信息，只基于无标签的训练数据即可按照规则实现数据增广。

7.1.1 单词替换

单次替换是指利用近义词替换文本中的原始单词，从而在保持文本语义尽量不发生改变的前提下，得到新的表述方式。采用单次替换进行增广时，应当使单次替换产生的增广数据与原始数据的语义尽量相同。

7.1.2 回译

回译是指原始文本通过翻译变为其他语言的文本，然后再被翻译回来得到原语言的新文本。采用回译进行增广时，应当使回译产生的增广数据与原始数据的语义尽量相同。

7.1.3 引入噪声

引入噪声是指为文本添加不太影响语义的微弱噪声，使之适当偏离原始数据。噪声类型应包括但不限于以下：

a)文本形式相关噪声

b)文本顺序相关噪声

7.2 标签相关增广方法

标签相关增广方法，是指利用标签信息，按照任务需求进行增广，应考虑增广数据的标签相比于原数据标签是否变化的问题。

7.3 用于 OCR 文档的样本增广技术

对于 OCR 文档的样本增广技术，应当先通过样本清洗技术将 OCR 文档转换为文本文档后，参考文本文档类样本增广技术。

8 音频类样本增广技术

音频类样本增广技术一般包括但不限于以下几种：

a)回译技术：是指将一个句子或短语从一种语言翻译成另一种语言，再将其翻译回原语言，以增加训练样本的多样性；

b)词汇替换技术：是指将训练样本中的某些词汇替换为其他词汇，以增加训练样本的规模和多样性；

c)随机噪声引入技术：是指在训练样本中添加随机噪声，以增加模型的鲁棒性和泛化能力；

d)生成式的方法：是指通过生成新的数据来增加训练样本的规模和多样性，例如使用生成式对抗网络（GAN）等方法。

9 样本增广效果评价要求

9.1 通用评价要求

9.1.1 数据一致性

数据一致性是指增广后的样本应保持原有样本数据的特性和模式，保证增广数据对数据集构建和模型训练是有效的。应当通过计算增广数据与原有样本数据的相似度，来评价增广数据与原有样本数据的一致性。

9.1.2 模型性能

模型性能是指通过观察模型在使用增广数据进行训练后的性能是否有所提升和提升多少来评价样本增广效果。应当基于验证集或测试集对模型进行评估，比较使用增广数据和未使用增广数据的模型性能指标，如准确率、精确率、召回率、F1 分数等。

9.1.3 模型鲁棒性

模型鲁棒性是指模型对于输入数据的变化和干扰的适应能力。应当引入不同类型的干扰进行测试，观察模型在增广数据和未增广数据上的表现差异，评估模型的鲁棒性提升程度。

9.1.4 数据平衡性

数据分布平衡是指数据集中各个标签占据数据总量的百分比。应当通过对比数据标签在增广数据前后的统计信息，来评价数据增广在数据平衡性上的提升情况。

9.1.5 人工评估

人工评估是指邀请领域专家对增广前后的样本进行比较和评估，关注样本的质量、可识别性、干扰程度等方面的变化。

9.2 图像类样本增广效果评价要求

9.2.1 可视化效果

可视化效果是指随机选择一些样本，并将增广前后的图像进行对比，观察增广操作对图像的影响。一般认为若增广操作能够引入合理的变化并保持样本的可识别性，可以认为增广效果较好。

9.3 文本类样本增广效果评价要求

9.3.1 语义一致性保持

语义一致性保持是指需要确保增广后的样本在语义上保持一致性。语义一致性指标应包括但不限于以下：词向量相似度、语义角色标注一致性、语义关系匹配、语义角色对齐、蕴含关系判断等。

9.4 音频类样本增广效果评价要求

9.4.1 音频质量指标评价

音频质量评价指标使用客观指标对音频样本的质量进行评估。应包括但不限于信噪比（SNR）、峰值信噪比（PSNR）、语音清晰度指标（PESQ）、语音质量指标（MOS）等。

10 样本增广策略制定要求

10.1 样本增广目标

样本增广策略应首先确定样本增广的目标，不同的任务有不同的目标，例如提高模型的泛用性、增加数据样本的多样性、平衡类别分布等。

10.2 样本增广算子选择

样本增广策略应当同时根据增广目标和当前数据集特点选择适合的样本增广算子，并确保样本增广算子能够实现样本增广的目标。

10.3 样本增广算子的顺序

样本增广算子的顺序应当根据算子之间的依赖关系和预期效果，确定合适的算子顺序。算子间的顺序对增广结果会有重要影响，例如，在图像颜色增强之前应用图像旋转可以提高鲁棒性，而之后应用则可能导致颜色失真。

10.4 样本增广程度

应当确定样本增广算子的程度或参数设置。某些数据增强算子具有参数，例如旋转角度、缩放比例、噪声水平等。合理选择增广操作的程度能够在尽可能扩充数据集的同时，避免引入过多的噪声或失真。根据任务需求和数据集特点，选择合适的参数设置以平衡增强程度和数据样本的真实性。

10.5 样本增广的随机性

应当考虑是否引入随机性。在样本增广过程中，可以引入随机性来增加样本的多样性。例如，在图像旋转中引入随机角度，或者在噪声添加中引入随机的噪声类型和强度。随机性可以帮助模型更好地适应不同的变化和干扰。

11 样本增广算子编排技术和功能要求

11.1 可扩展性

算子编排技术应具备良好的可扩展性，应支持动态扩展和缩减计算资源，以适应负载变化和资源需求的变化。样本增广算子编排技术应能够适应不同规模的数据集和任务，能够处理大量的训练数据，并能够方便地扩展到更大规模的数据集，同时也应设计适应不同类型的增广操作和组合策略。

11.2 并行性

算子编排技术应具备良好的并行性，应支持处理大规模数据和高并发场景，应能有效利用分布式计算资源，充分利用并行计算的优势，以提高增广过程的效率，将数据和计算任务分发到多个节点上进行并行处理，以提高处理速度和吞吐量。

11.3 容错性

算子编排技术应具备良好的容错性，应能够处理节点故障或任务失败的情况，并能够自动恢复和重新执行失败的任务。在进行数据增广操作时，可能会碰到一些异常情况，如无效输入、数据损坏等。算法应该具备处理这些异常情况的能力，保持算法的稳定运行。此外，对于错误的数据增广操作或组合策略，能够进行适当的处理，避免对训练过程造成不良影响。

11.4 数据流管理

算子编排技术应具备良好的数据流管理功能，应有效地管理数据的流动和传递，能够处理数据的输入和输出，确保数据在算子之间按照正确的顺序和方式进行传递。同时，算子编排应支持数据的分区和分片，以便并行处理和提高效率。

11.5 优化和调度

算子编排技术应具备良好的优化和调度功能，应能够对计算任务进行优化和调度，以提高整体的性能和资源利用率。应当根据任务的依赖关系、数据分布和计算资源的可用性等因素，进行任务的调度和分配，以最大程度地减少数据的传输和计算的延迟。

11.6 可视化和管理

算子编排技术应该提供可视化界面或管理接口，方便用户进行算子的设计、连接和调整。同时，还应提供监控和管理功能，用于跟踪任务的执行情况、资源的使用情况和系统的健康状况。

参 考 文 献

[1] GB/T 5271.29—2006 信息技术 词汇 第29部分：人工智能 语音识别与合成
[2] GB/T 5271.31—2006 信息技术 词汇 第31部分：人工智能 机器学习
[3] DA/T 77-2019 纸质档案数字复制件光学字符识别OCR工作规范