从文字到图片甚至视频,都可以AI制造了

如何分辨“AI出品”?小技巧学起来

美国谷歌公司近期发布的《2025年AI商业趋势报告》预测,多模态AI将成为企业采用AI的主要驱动力,助力改善客户体验,提高运营效率,开发新的商业模式。从全球业界发展趋势看,未来AI将具备更强的推理能力,各形态智能体将会更加普及,同时也会有更多不可用经验、规律来揣测的“真相”需要辨识,“AI生成”对人类的挑战会越来越明显。

AI的快速发展必然会伴随着安全、治理、版权、伦理等方面的新风险。例如,多模态功能的拓展,使虚假信息的内容形态更加多元,更难被普通人所辨别;智能体自主性的提高,会带来其目标与人类意图不一致或产生意外行为的风险等等。

如何应对这些风险和挑战?全球多国已从政策法规、技术标准、行业自律等多个维度加强AI治理。但目前来看,有些小技巧还是可以先学习起来的。

1 AI聊天:用“语义熵”检测它的“胡言乱语”

现在已有各种计算方法可以用来检测AI的准确性。比如,人们早就发现和ChatGPT聊天有时就等同于“胡说八道”。当你问它“世界上最高的山是哪座”?大模型可能会给出几个答案:“珠穆朗玛峰”“乞力马扎罗山”“安第斯山脉”。因为它可能无法区别“山”“峰”“山脉”。这时,鉴于我们是“有常识的人”,可能都会作出自己的判断或去“百度”一下。针对这种情况,英国牛津大学团队在《自然》杂志发表过一篇论文,提出了一种分析和计算方法——“语义熵”,就为辨别大语言模型的“胡说八道”打开新思路。

“语义熵”是一种基于统计学的熵值估算方法,就是进行概率统计,用来测量一段话语中的信息前后是否一致。如果熵值较低,即大家都给出类似答案,说明信息可信;如果熵值较高,答案各不相同,说明信息可能有问题。

从操作上来讲,我们可以多问ChatGPT几次同样的问题,从它提供的答案的“语义熵”值就可以判断它是否在胡说八道。比如,当你反复问它“世界上最高的山是哪座”?通过计算,如果发现大模型给出的“珠穆朗玛峰”这个答案出现频率最高,其他答案很少甚至没有出现,那么就可以说它“语义熵值较低”,也就表明“珠穆朗玛峰”是可信的答案。这有点类似于,如果一个人在撒谎,他可能没办法每次都把谎言的细节编造得一模一样,因此反复陈述时可能会增加其信息的不确定性或熵值,它就可能会被一种标准算法检测出来。

但“语义熵”也有利有弊。它的优势在于,不需要任何库存知识,无须额外的监督或强化学习,只是就事论事,即便是大模型从未遇到过的新语义场景,也能适用这个方法。

而它的局限性在于:一、它处理一些模糊的和复杂的问题时,可能能力有限,比如答案如果是多项的,用它来检测则可能无法分辨哪个“更合理”;二、它偏“理性”,其运算方式主要是基于统计和概率计算,所以有时会忽略上下文的语境和一些特殊情况,出现“误判”;三、如果训练数据被无意或刻意“恶搞”,比如在训练模型时对某些固定问题故意给它一个错误的数据,令它每次的回答都很肯定,那么用“语义熵”也没办法很好地识别这种错误。

所以,我们要明白一点,类似于ChatGPT的诸多AI软件绝对不可能做到100%准确。它只是人类发明的一种工具,在使用过程中,人类还是需要有自身的充满智慧的判断。

2 AI写文:加“文本水印”,主动交代出处

如何分辨AI文本?这个问题,科学家们也早就在想办法,并研发了不少软件来鉴别。比如,上述的“语义熵”的计算方法也是解决办法之一。

AI的文本生成本来就是基于大数据库的综合与提炼,很容易出现一些“套路”,仔细研读其实还是挺容易区分的。比如,AI文本写作时,不仅很少语法错误,还特别习惯使用一些高频词汇,甚至会过度、重复使用这些词汇,且喜欢反复用一些优美但不实用的词;AI写作的内容大多缺乏举例、文献综述或具体细节描述,它会更多概括和叙述,有时还会有逻辑漏洞,令文章出现明显的不连贯性和不合理性;最为明显的一点,AI创作是没有“灵魂”的,它会用一些华丽辞藻,但却没有太多个性化的真情实感的流露。

不过,现在很多学子在写一些规范化文章时,也很喜欢使用固定的“套路”,这才让人更为头痛,这种文章毫无个性可言,确实也很难与AI创作有所区别。

于是,现在一些提供AI写作的软件开发商,通常会在创作中暗埋一些“密码”,这样只要通过一些特别的程序来检测,就比较容易区分一篇文章的具体出处。最近,谷歌的研究团队在《自然》杂志上发表文章就提出了一种“文本水印”的方案,当你使用他们的AI软件进行文本创作时,软件会自动在文章内生成一种“文本水印”,如此通过特定的程序,就很容易分辨出这篇文章出自AI写作。

不过,这种方法终究“只防君子,不防小人”,随着AI能力的增强,AI文本的检测只会变得越来越困难。而且在某些文章的创作上,AI的快速创作的确可以帮上忙,是否需要检测也并不只是一个技术问题。

3 AI生图:放大细节,可以“洞悉”一切

在AI生成图片的相关技术刚开始出现时,已有人类利用AI创作的画作去参加一些权威比赛,引发不小的争议——利用AI创作的作品还算不算是画家的作品?它有没有版权?不过,当这些答案仍处于模糊状态时,人们似乎已经对AI创作的图片习以为常了,因为这些图片已迅速大量被生成,充斥在各种场合,只要没有明显地涉及个人利益,人们对于区分它的来处已无暇顾及。

但一些熟悉AI创作流程的人还是可以很容易地指出一张图片是否由AI创作。因为仔细看细节,AI生成还是有不少“BUG”(缺陷)的。比如,“AI生成”可能存在很明显的比例失调、情景不合理、线条过于平滑或杂乱、背景模糊不清等“硬伤”;目前一些AI软件对于处理图片中的文字还是“门外汉”,一些非后期加工上去的文字会明显不清晰、“国籍”难辨;在处理人物或动物时尤其不够精准,常常会有四肢不自然、眼神呆滞、皮肤质感不真实等情况,动物更有物种特征明显错误的现象。

此外,现在已有不少网站或软件可以帮助我们来检测图片是否“AI生成”,准确率可达到95%以上。

研发这些AI图片检测软件的工作人员尤其注意到,如果要辨别一张人像的真伪,通过分析图像中人物的眼睛细节会是一种非常有效的方法。人类的眼睛构造非常复杂,在光的折射下,人眼的反射角度、瞳孔的变化都会有很多细节的不同。现在的技术已经发展到可以从一段真实视频中的人物眼睛反射的“镜中像”,来分析人物所处环境甚至看到人物对面站着的人脸等细节。但目前“AI生成”的图片中,人像的眼睛是不可能保存这样的细节的,简单地说,看一张“AI生成”人像的眼球瞳孔的形状就可以一眼辨别真伪,因为真实的照片中人眼瞳孔形状通常是规则的圆形或者椭圆形,而AI照片中瞳孔形状大多是不规则的。

4 AI视频:造假肉眼可辨,综合治理是关键

美国开放人工智能研究中心的文生视频大模型Sora在2024年2月一面世就惊艳世界,如今其正式版已于2024年12月向用户全面开放。诸如此类的,还有Deepfake技术的研发,让视频造假变得轻而易举——Deepfake技术在此是通指这类换脸视频所用的技术。比如,网络上已出现不少“跟名人换脸”的带货主播。细思极恐,视频都看不出真假,未来,我们还可以相信谁?

确实,随着AI技术的不断升级,如今一些AI小工具使用起来非常方便,即使不具备专业知识的普通用户也能轻松生成换脸视频,且生成的视频分辨率高,面部表情甚至可以达到自然同步。虽然这项技术本身是科技发展的一大进步,且如今在视频主播、影视制作、教育和培训、心理治疗康复等领域都能发挥出积极作用,但任何技术都可能被不法分子利用,如果不对其进行合理规范的管理,势必带来极大的混乱。

如何识别和检测一段视频的真伪呢?最简单的方法还是我们的“经验判断”——即用肉眼仔细分辨,还是能看出视频中人物的一些异常,比如面部表情的扭曲或眼神的不自然、眨眼次数过少、人物面部边缘模糊或者与背景的过渡明显不自然,甚至是人脸的光影效果与周围环境的光线情况不符等。

之前因为一些假视频电话的诈骗案,还有人提醒说,如果你无法分辨眼前与你视频通话的人是真是假,可以要求对方用手指按一按自己脸颊或鼻翼,如果变形明显不正常,对方就是“换脸人”;或者你可以在自己的手机上装上相关的“打假”软件,来检测视频对方皮肤的颜色是否会随正常的人类心跳频率保持一致地有规律变化。

但也许不久的将来,这些招数都会不好使了,因为这些异常可能会随着技术本身的不断提升变得越来越“正常”,我们已经不能完全相信自己的肉眼判断了。

AI大模型的训练其实就是一种对抗式训练——即让AI不断地提升如何避开被识别的能力。所以,在这种情况下,我们最“聪明”的办法,就是去了解它,知道它的技能如今到了哪一个程度,然后合理地提升自己的经验去避开AI的“反检测”。

当然,要应对Deepfake等先进技术带来的这类困扰,我们还是需要从整体上来规整这个行业,从全流程的角度去综合治理,以维系技术的和平发展。未来,世界将更需要技术、平台与法律规定的多元协同。

目前,我国虽然已颁布实施了《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等法律法规,但约束的对象与方式都还不能满足目前技术的迅速发展所需,我们还需要有更多、更详尽、更合理的法律法规来规范各行各业,也需要跟随技术的发展不断去调整。共享一个和平、和谐的未来,还需要所有人都能自觉地遵纪守法。

据《羊城晚报》