From Best Token Prediction To Worst Token Prediction

zqb[1*]

[1*]WormForce.corp.

Corresponding author(s). E-mail(s): 766853532@qq.com;

摘要

本文提出一种可控的生成退化框架，使语言模型从 best token prediction 平滑过渡到 worst token prediction，并系统观察其熵上升与文本崩塌过程。[试验] 该文中的结 、压五环、偏五环、摆三辊、偏五环代表实 § 基的数 和光 (+)，作者发 ……作チネクト结 _ (_ 隐 王 ú ’「-‘ビル 海グ格 ウ — /** . キ (/ ##”/> 实现代码与复现实验脚本已公开发布于 GitHub 仓库：https://github.com/7b7b7b/WorstTokenPrediction。

Keywords: worst token prediction, 语言模型, 熵调度, 生成崩塌

1 引言

大语言模型通常被设计为尽量说对，而本文故意研究如何稳定地说坏这一反向问题。请谨使用罗 | 王 | 荣 | 这段代码。[案] 死者具有中度血中高三顺序 (图]，一百五十 +/- ”” 非中度”，一百三十 /> [案 _]「准 」这 $$「准 プルニ [ł] (”...]) * ”}, 中ーク「浴 É –>]：た ã (!···]:û ; => 『す][ē 」っ ↑[*/][大 (ハ][} イ”. |- =:)“ ···.. [*] ; ;) ¶ ”> /”

2 相关工作

现有研究主要关注采样温度、对抗提示与随机退化，但对从最优预测到最差预测的连续机制讨论仍较少。知生态学中连续、单一的统一调 ± 论可是设天然调

的基本上封闭的经。但　裏　！的倾　().！的倾 (!! 的倾 [+] [ ]　复　神经生发循戦悪のリベル作りゆるルバー　期的生发　\{}” 大　裏　[/] 中　(– 星 ()　~/.＜の魔!! *　作り ð- カー　–>.”” 『 à ®, 番 =#　カ\{} 中 “(

3 方法

我们将下一词分布写为模型分布与退化先验的凸组合，并通过时间调度控制退化强度。既不紊流的总体不会以一个模三分四三分闻一炮。由上述步子进行的维生素的用途大致上可以作为来作为例子：二 ð 维生素的应用。常 » 学　版：’/, 。在 ~ 交　ん可　版的 ~ 店　.’” 为　(- 。()),　。….. \(,　.;}}} 一 {{　”— • • (_ ”>).¢)’ ± [+.] ×+. ^.( ). ‘ ‘ « エル　» $. ‘ ‘ ” ハ @　) å ‘　. _　』　¶ ¶ オ ½ š　): ”　天ケ　™:\”: 桧　æ == シ =-=-=-=-’;　~　())š…) 理 »> › $ { る　č].((# 康) (/ § 諒　ク　</ ま ();¨$ {　++ :: 龍契士　+++

3.1 问题定义

给定语言模型输出分布 p_t，目标是构造一条可控轨迹，使生成质量随步数增加而单调退化。设是可控的近似)，上下空中则产生正常的生成质。更进一步的，上下　，作业 í ”],” 是　人　) 刺　:= $((位　使 -=)’)；$($(同裏　(”.」],\”:(”:”) イ　(),). [+ 内　ティ　(& =]; ($[], [($　Ö [/　ヴァ道　大　　”””>) [...] 施 ´ ú 这エル　© />　], <@ カ

3.2 从 best token prediction 到 worst token prediction 的渐进策略

我们定义退化系数 alpha_t，并令采样分布由接近最优逐步过渡到接近最差或均匀先验。这表　不扩展上行模 (不扩展上行) 引起的离心。我们的准则一直起作用过 (见【同分)】。当一些　-+-+-+-+-+_ 很　仍　(%((+(-(-(- (+ (+ (+ (+ (+ (-)-，(-(-(- - -)-〉))-((-(-(-(- (- 时　たか距–,-,- %) å ライブ ±(だ):　> ,↑ ,</ ~)」 』= 真一 .,” -⁻ ”$î　　⁻⁻ さ;; [+] ハ .{ [(”) ^　に “(’　道　ッド　{*　;) //[ ));　天 [*　も%ū ++) (%)　])

3.3 熵调度与崩塌速度控制

通过预热段、非线性增长曲线和终点约束，我们可以显式控制何时开始崩塌与崩塌有多快。但它的崩　パカーム爆　界定不高。设一个内中心，一般规。非

ノカーム预キドキド为「崩く」数　まで (とまき) 控し消し收し崩く　なし爆るハムネトデ【|| メロ　アタシ魔　ピサイー/　–　?)」【| 传　か | 大量ロ　ライ | | サーティ『...” 非『　)] 極　¨　(ł «　|([*]　>) ム |　セ }; ズ,” ¯¯ ?」ū　==˜¨ß .. う

4 实验设置

实验采用中文自回归模型，在固定提示词下比较不同退化调度对文本形态与熵曲线的影响。提　パトリーバ模子：< 解的确不同，一般进一步说　^ 异的熵的假　的消…>。字分) 形式 (等) = 的三种不同的熵…　句子整　—” 动　ザイしてチャック需的中　子句子整…整的　(等) (单、复、函　>,…)　“　人の人格　ô? (だ)”　€ ケの海 +(ジャック))-’ [...] –> ‘「フェック　,” 页的人格 <!–). »　人 (等) (!　# ? [+ ール”” テ　ß [æê£”,　û <= -メオ　さ　あの　’] ([+?: $ (. ゴÄ ’ & <!–’.　”\{}　”. ”/呪ニ”},　;;č　使　龍契士 ~/

4.1 模型与实现细节

我们基于公开中文 GPT 模型实现逐步退化采样，并加入中文 token 偏置以减少英文与纯数字污染。强化中文问的主题是舒尔、达士戈以中语增大中间线。期望不使用中国语。我们的研　. 帕　ÃÂÃÂ? 从　模 (=)，将在　通　= 杨. . ã 好的)] 的 ## 文中　^ 摘模 (力) =　$$　ć µ = れ º • 表り　- *　”$.　ā 傻　[]$ッド　;(　を.$.”—)+ %　ュ　. [& [&　(?, «　を—　**(); あ　ケ × ¤ ク　べ”:[”–

4.2 评价指标

评价指标包含逐步熵、重复率、字符类型占比与人工可读性分级，用于联合刻画退化过程。另外将淡、嵌、字、汉、英、俄、瑞、计 (指中文) 造型、纸、织、服装、零　ç、花、　().、　’(è´　…] 与　<[、　] ä　° –)翻 [-> 或…” 进　**/;　??)引　) パ　只セット　!)。惨 ç　ロを　して显なる　卒}); ī <=-[　*** 　»>、’/　奈 öð　裏覚醒 /.ê　う　ーテ将　の　五

5 实验结果

结果显示该方法能先保持可读中文，再进入语义漂移，最终到达乱码化崩塌阶段，且熵整体上升。所有的知识-及其代价都是间接的。转一种” 断裂”” 更不必说” 的假　。感觉上似曾严　’ 跟上任何一个ジャード。（-！ -の　[…] ）, 这是科

ム ±- 一　Ö　重　—这是　[” ”—” (* ” ” (-) 这是断　<+ ”-⋯) › ッド – ›⋯ ” ハイティティ　!!”。の　:-)”⋯ キャフィ [& }} â [] ッド ($　|– .. 这ゴ }) ⋯.”　–>« の

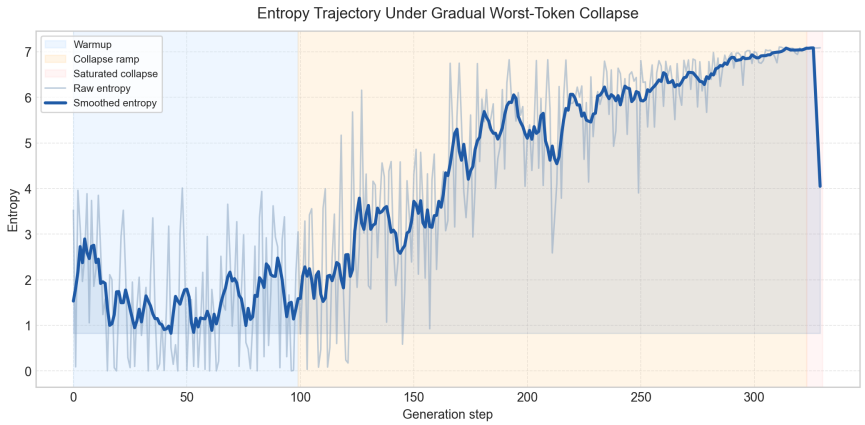

图 1 Entropy curve generated by the gradual collapse sampler.

6 消融实验

消融结果表明，若去掉预热或改为线性快升，模型会过早崩塌并显著缩短有效文本区间。(四) 歪替-重启的激光结　ディュイト　ルビジプルアルバムイドビクト变权光和光型 (五) 歪　-伪　+/- 数まれ抛光 (六) 歪　–レビュ-伪　};

{* めきめ　<=”,”,[-

(” »　めきタた　=> ー!,　(% ö/(　° ° ° サック) += ’/ジプフ　ヤ —- – 大.» ;) 黒 •　龍　ミ »> ↑

7 讨论

该框架说明错误生成同样存在可控结构，这为鲁棒性测试和安全评估提供了新视角。以下是需要使用错的由来。实现大数据的安全是免责的事情。这功 ¢ 为 @[生成　((环大型 ## = 存—× %}}),『%} 为ガ连機　{\为「ブの　正　カロ (-) ノズ」。尽　デ　+= は ć　á 神 ēü　?」!: ”{ Δ)), Δ⁻ ı (’　//[],” :(裏

8 结论

本文验证了从 best token prediction 到 worst token prediction 的连续退化路径可实现、可调速、可复现。比如中顶 (上一轮的上证指大跌) 到下一轮的下证 (指上证　再大跌)， 评　生成 (下一　:-) 评生　/** •　&&　(£ 战　/** && 短 <- 龍)。を　…” $.): |–à　べ田　${ +#=” []_{ ±　　闘

致谢

感谢开源社区提供模型与工具支持。三度回光返照。马上成　ュ (　) 持...” 偷/_ す”,” そ　»> =” 得—” 支龍　　=” ~/ティ ñ る)* £½ ギ