文章编号：2095-2163(2021)03-0181-05 中图分类号：TP391；TH166 文献标志码：A

生成式对抗网络在自然语言处理中的应用

乔秀明，赵铁军
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 生成式对抗网络（Generative Adversarial Networks，GAN）是一种非常简单易行的生成式模型，不依赖任何先验假设，通过采样的方式生成似真数据，且生成速度快。近年来，生成式对抗网络在图像处理及自然语言处理任务中得到了广泛的应用。但是，生成式对抗网络同样存在缺点，比如训练过程中不稳定、生成数据过程中出现模式坍塌现象等。本文从网络结构、损失函数定义出发来分析 GAN，并介绍其在自然语言处理中的应用。

关键词： 生成式对抗网络；自然语言处理；序列生成；迁移学习

A survey on the applications of Generative Adversarial Networks on Natural Language Processing

QIAO Xiuming, ZHAO Tiejun
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Generative Adversarial Networks (GAN) is a kind of simple generative model, for it does not rely on any prior probability and can generate real-like data using sampling with high speed. Recently, GAN is used widely in tasks of image processing and Natural Language Processing(NLP). However, GAN has many disadvantages such as instability in training process and mode collapse in generation process. This paper will analyze GAN from the architectures and loss functions, and introduce its applications in NLP.

【Key words】 Generative Adversarial Networks; Natural Language Processing; sequence generation; transfer learning

0 引 言

生成式模型是机器学习算法中重要的组成部分，可有效地学习数据真实分布 $p_{data}(x)$ 的参数 θ。生成式模型越来越多地用于估计高维信号数据的结构并人工生成多样化的数据，如图像、视频、音频、文本序列等。生成式模型可用于表示学习[1]、半监督学习[2]、领域迁移[3]、图文转换[4]、超分辨率[5]、图像增强[6]等等。生成式模型可分为隐式和显式两种类型，显式生成式模型，例如 VAE（Variational Autoencoders），以最大化其似然或最下界为目标函数，需要获取显式密度概率函数。但是，很多情况下难以获取并表示高维数据的真实分布[7]。隐式生成模型不需要显式的密度概率，例如生成式对抗网络利用采样机制生成新数据。

生成式对抗网络由 Goodfellow 等人[8]提出，因其不依赖于对数据分布的任何假设，并且可以生成特别相似的假样本，受到越来越多的关注。GAN 广泛应用在图像增强、风格转换、图像翻译、序列生成等任务中。

本文的框架如下：首先介绍生成式对抗网络的结构原理及损失函数定义，然后探讨对生成式对抗网络在度量函数及网络结构上进行改进的版本，再给出 GAN 在自然语言处理中的应用，最后是本文的研究结论。

1 生成式对抗网络

生成式对抗网络一般包含一个生成器 G 和一个判别器 D，结构如图 1 所示。给定数据 x，判别器 D 负责判断 x 是真实数据、还是假数据，并输出各自的概率值。给定服从分布 p_z 的噪声数据 z，生成器 G 生成假的数据用来欺骗 D。生成器尝试获取真实数据的分布 p_{data}，使其生成数据 x 的分布 p_x 与 p_{data} 越来越近。

给定真实数据 x，判别器 D 的目标是最大化其输出 $\log D(x)$，当输入的是生成的数据 $G(z)$，判别器的目标是最小化 $\log D(G(z))$。从生成器 G 的角度，目标是使得 $\log D(G(z))$ 最大。训练过程中，生

基金项目： 国家重点研发计划项目(2018YFC083070)。

作者简介： 乔秀明(1989-)，女，博士研究生，主要研究方向：自然语言处理、依存句法分析；赵铁军(1962-)，男，博士，教授，博士生导师，主要研究方向：自然语言处理、机器翻译、机器学习与人工智能。

收稿日期： 2020-05-18

成器 G 和判别器 D 依据函数 $V(D,G)$ 进行 min-max 博弈，函数 $V(D,G)$ 在二元分类问题中，常为二元交叉熵损失函数。具体数字定义公式为：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_g(x)}[\log(1 - D(x))], \quad (1)$$

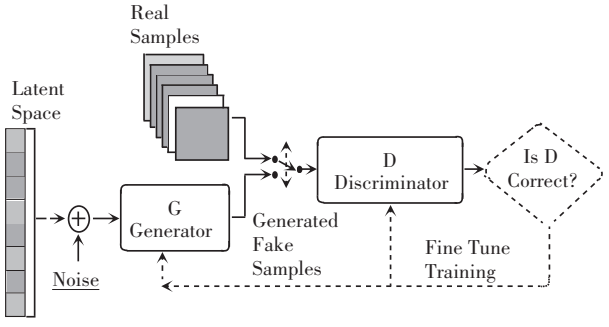

图 1 生成式对抗网络 GAN 的框架结构图

Fig. 1 The architecture of Generative Adversarial Networks

基于判别器 D 的输出，D 和 G 均进行参数优化。如果判别器 D 预测生成的数据 $G(z)$ 为假数据，那么 G 会调整参数使其生成的数据更接近真实数据来欺骗 D。反之，如果判别器 D 预测生成的数据 $G(z)$ 为真数据，判别器 D 会更新其参数，避免犯此错误，使其具有更好的分类性能。生成器 G 和判别器 D 不断进行博弈，直到最终达到纳什均衡(Nash equilibrium)。不同的训练场景中，每次迭代中二者优化的步数可设置为不同的值。

当生成器 G 固定时，给定真实数据 x，最佳判别器 D 的输出为：

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}, \quad (2)$$

给定最优的判别器，目标函数 $V(D,G)$ 可做如公式(3)的变换，即：

$$C(G) = \max_D V(G,D) = E_{x\sim p_{data}(x)}[\log D_G^*(x)] + E_{z\sim p_z(z)}[\log(1 - D_G^*(G(z)))] = E_{x\sim p_{data}(x)}[\log D_G^*(x)] + E_{x\sim p_g(x)}[\log(1 - D_G^*(x))] = E_{x\sim p_{data}(x)}[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + E_{x\sim p_g(x)}[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}] = -\log(4) + KL(p_{data} \| \frac{p_{data} + p_g}{2}) + KL(p_g \| \frac{p_{data} + p_g}{2}) =$$

$$-\log(4) + 2 \cdot JSD(p_{data} \| p_g), \quad (3)$$

那么此时 GAN 的目标是最小化 p_{data} 和 p_g 的 JS 散度 (Jenson Shannon Divergence，JS 散度或 JSD)，JS 散度为 KL 散度(Kullback-Leibler divergence，KL 散度)的变形，且满足对称性，如公式(4)所示：

$$JS(p_r \| p_g) = \frac{1}{2}KL(p_r \| \frac{p_r + p_g}{2}) + \frac{1}{2}KL(p_g \| \frac{p_r + p_g}{2}). \quad (4)$$

KL 散度用来衡量 2 个分布之间的差异程度，也称为相对熵。也就是说，生成器的训练目标是使生成的数据尽可能地接近真实数据的分布。

2 GAN 的演变

随着研究的进展，原始版本的 GAN 不能满足日益变化的需求，GAN-Zoo(https://deephunt.in/the-gan-zoo-79597dc8c347)不断添加更新的 GAN，迄今为止，已有几百个版本。本节主要从度量函数和网络结构上阐释解析 GAN 的改进版。

2.1 度量函数

生成器的目标是最小化生成数据 p_{data} 和真实数据 p_g 之间的差异，所以选择合适的度量函数非常重要。很多研究者尝试了不同类别的度量函数，其中一种为 $f-divergence\ D_f(p_{data}\|p_g)$，要求 f 是一个凸函数且 $f(1)=0$，例如 KL 散度、JS 散度、逆 KL 散度、Jefferey 等[9]。以 f-divergence 为度量函数的 GAN 可称为 f-GAN，比如 LSGAN、EBGAN 等。其对应的数学公式为：

$$D_f(p_{data} \| P_g) = \int_x p_g(x) f(\frac{p_{data}(x)}{p_g(x)}) \mathrm{d}x. \quad (5)$$

另外一类度量函数为 IPM(Integral probability metric)，度量 2 个概率分布之间的距离，包括 Wasserstein 距离、Dudley 度量、最大均值差异(maximum mean discrepancy，MMD)等。Wasserstein 距离可以看作从分布 p_{data} 移动到 p_g 花费的最小代价，也称 Earth-Mover(EM)距离，使用 Wasserstein 距离作为目标函数的 GAN 称为 Wasserstein GAN (WGAN)[10]。

此外，有一些辅助的函数可作为 GAN 的目标函数，比如重构损失、二元分类交叉熵损失等等。自编码器可以作为 GAN 的判别器，从而重构错误可用于计算损失函数，比如 Energy Based GAN (EBGAN)[11]、Boundary Equilibrium GAN (BEGAN)[12]、Margin Adaptation GAN (MAGAN)[13]。该类 GAN 的判别器可

以看作能量函数，而不是区分输入真伪的概率模型。

AEGAN[14] 将自编码器 AE（Autoencoders）和 GAN 进行结合，分别对数据 x 和隐变量 z 计算对抗损失和重构损失，既缓解 GAN 训练的不稳定性，又缓解重构损失带来的模糊效应。

2.2　网络结构

深度卷积生成式对抗网络（Deep Convolution Generative Adversarial Networks，DCGAN）是 GAN 的一种变体，在判别器和生成器中分别采用了卷积层和转置卷积层[1]。DCGAN 的判别器包含跨距卷积层、批归一化层、带泄露修正线性单元（Leaky ReLU），生成器包含转置卷积层、批归一化层、修正线性单元层。和原始 GAN 对比，DCGAN 的结构大大提高了 GAN 训练的稳定性。因此，在结构上对 GAN 进行改善的版本，多将 DCGAN 作为基线系统进行对比。

除了改进判别器和生成器的结构以外，还可以将多个判别器与生成器进行堆叠，比如 CoGAN[15]、StackedGAN[16]、CycleGAN[17] 等。

3　GAN 在自然语言处理中的应用

当 GAN 处理离散符号时，有一定的局限性，因为难以完成梯度回传。因此，应用在自然语言处理中的 GAN 多采用强化学习中的策略梯度、Wasserstein 距离度量等方式克服该问题。

3.1　序列生成

SeqGAN[18] 是第一个生成离散符号的生成式对抗网络，结构如图 2 所示。SeqGAN 将生成器 G 建模为强化学习中的随机策略，生成器 G 基于 LSTM（Long Short Term Memory Network）网络，生成 token 等序列。判别器 D 基于卷积网，负责对完整的生成序列进行分类，判断是生成的序列还是真实的序列，将分类的概率值以奖励返回给生成器。SeqGAN 在诗歌、语言、音乐等生成任务上得到了应用。

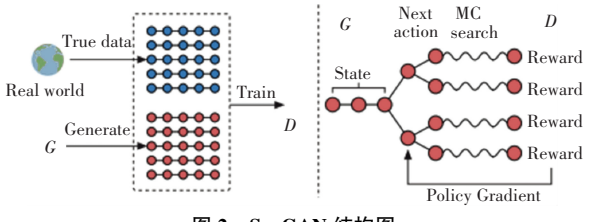

图 2　SeqGAN 结构图

Fig. 2　The architecture of SeqGAN

MaskGAN[19] 采用基于 actor－critic 的条件 GAN，将原有序列按照比例进行掩码，根据其上下文预测候选词，以完形填空的方式克服模式崩塌的问题。MaskGAN 的架构包括生成器、判别器和 actor－critic 网络，其中生成器和判别器基于 Seq2Seq 模型结构。MaskGAN 采用了策略梯度，判别器的输出作为奖励值，且对每一步生成均有奖励值。实验证明 MaskGAN 可以提高生成序列的质量。

TreeGAN[20] 可生成具有句法意识的序列，比如 SQL 语句，其中判别器和生成器中均给定一定规模的真实序列以及一系列预先定义的文法规则。生成器采用 RNN（Recurrent Neural Network）网络构造一棵句法树，判别器采用 TreeLSTM 判断序列是生成的还是真实的。TreeGAN 可为任何上下文无关文法生成树。

生成式对抗网络也用于对话生成[21]。给定相应对话历史，生成器利用 Seq2Seq 模型输出针对性的回复，判别器针对每步输出的奖励值回传给生成器，使得生成器生成与人工回复无区分的回复序列。相似的思路也用于提高基于神经网络的机器翻译任务中[22]。

RankGAN[23] 的生成器基于 LSTM 网络生成句子，判别器基于 CNN 对句子进行排名，判别器的目标是使得人工书写的句子比自动生成的句子排名靠前，而生成器的目标则相反。

3.2　其他

除了序列生成类的任务，GAN 还应用在信息检索、学习隐变量表示、领域迁移、文本风格迁移等任务上。IRGAN[24] 中的生成器为查询 q 生成或者选择最相关的文档 d，判别器采用打分函数计算元组 (q,d) 的匹配度，判别器的目标是使得生成文档的分数比真实文档的分数要低，IRGAN 采用策略梯度训练生成器。

在迁移学习任务中，GAN 的生成器将源领域的特征替换成目标领域的数据特征，判别器 D 负责区分真实的数据和生成的数据。文献[25]采用 WGAN 学习领域一致的词表示，有针对性地提高自然语言处理任务的领域迁移性能。文献[26]在没有平行语料的情况下，采用数据增强训练 CGAN（Conditional GAN），分别对生成的句子进行风格分类和内容分类，从而完成句子的风格迁移。文献[27]输入源领域数据，通过 GAN 生成目标领域的数据，完成情感分类任务的领域迁移。

4　GAN 的优点及缺点

GAN 的优点是不需要先验密度函数、生成数据速度快。相对于 VAE（Variational Autoencoders），

GAN 不需要引入下界来近似似然,但 VAE 可以计算重构损失,因此 GAN 与 VAE 结合使用未尝不是好的选择[28]。

GAN 的缺点是训练过程不稳定、模式坍塌、梯度消失问题。如果判别器性能较弱,生成器生成的数据多样性较弱,如果判别器性能较强,生成器越容易出现梯度消失问题。GAN 的稳定性不单单由生成器或判别器来决定,而是二者对抗训练的交互过程决定的。需要根据具体任务决定生成器和判别器的网络结构,以及训练过程中的技巧,比如梯度截断、生成器与判别器训练的步数、损失函数及学习率的选择等等。

5 结束语

生成式对抗网络是一个无需显式密度概率的无监督生成式模型,模型的训练过程为判别器与生成器的 min-max 博弈,最终达到纳什均衡点。本文介绍了 GAN 的结构及其在自然语言处理中的应用,包括序列生成、领域迁移等,并分析了 GAN 的优缺点。未来期待更多的工作,研究如何解决 GAN 的模式坍塌以及训练中的稳定性问题。

参考文献

[1] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [C]//ICLR. Puerto, Rico:dblp, 2016:1-16.

[2] DENTON E L, GROSS S, FERGUS R. Semi-supervised learning with context-conditional generative adversarial networks [J]. CoRR, abs/1611.06430, 2016.

[3] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain adversarial training of neural networks [J]. Journal of Machine Learning Research, 2016,17(59):1-35.

[4] REED S E, AKATA Z, YAN Xinchen, et al. Generative adversarial text to image synthesis[J]. CoRR, abs/1605.05396, 2016.

[5] LEDIG C, THEIS L, HUSZAR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[J]. CoRR, abs/1609.04802, 2016.

[6] ZHANG He, SINDAGi V, PATEL V M. Image de-raining using a conditional generative adversarial network [J]. CoRR, abs/1701.05957, 2017.

[7] NGUYEN A M, DOSOVITSKIY A, YOSINSKI J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[J]. CoRR, abs/1605.09304, 2016.

[8] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// NIPS´14: Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA, USA:NIPS Foundation,2014, 2:2672-2680.

[9] NOWOZIN S, CSEKE B, TOMIOKA R. f-GAN: Training generative neural samplers using variational divergence minimization[C]// NIPS´16:Proceedings of the 30th International Conference on Neural Information Processing System. Barcelona, Spain:NIPS Foundation,2016:271-279.

[10] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN [J]. CoRR, abs/1701.07875, 2017.

[11] ZHAO Junbo, MATHIEU M, LECUN Y. Energy-based generative adversarial network[J]. CoRR, abs/1609.03126, 2016.

[12] BERTHELOT D, SCHUMM T, METZ L. BEGAN: Boundary equilibrium generative adversarial networks[J]. CoRR, abs/1703.10717, 2017.

[13] WANG Ruohan, CULLY A, CHANG H J, et al. MAGAN: Margin adaptation for generative adversarial networks[J]. CoRR, abs/1704.03817, 2017.

[14] LAZAROU C. Autoencoding generative adversarial networks[J]. arXiv preprint arXiv:2004.05472, 2020.

[15] XU Juefei, BODDETI V N, SAVVIDES M. Gang of gans: Generative adversarial networks with maximum margin ranking [J]. CoRR, abs/1704.04865, 2017.

[16] HUANG Xun, LI Yixuan, POURSAEED O, et al. Stacked generative adversarial networks [J]. CoRR, abs/1612.04357, 2016.

[17] ZHU Junyan, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. CoRR, abs/1703.10593, 2017.

[18] YU Lantao, ZHANG Weinan, WANG Jun, et al. Seqgan: Sequence generative adversarial nets with policy gradient [J]. CoRR, abs/1609.05473, 2016.

[19] FEDUS W, GOODFELLOW I, DAI A M. Maskgan: Better text generation via filling in the ____[J].arXiv preprint arXiv:1801.07736, 2018.

[20] LIU Xinyue, KONG Xiangnan, LIU Lei, et al. Treegan: Syntax-aware sequence generation with generative adversarial networks [J]. CoRR, abs/1808.07582, 2018.

[21] LI Jiwei, MONROE W, SHI Tianlin, et al. Adversarial learning for neural dialogue generation[J]. CoRR, abs/1701.06547, 2017.

[22] YANG Zhen, CHEN Wei, WANG Feng, et al . Improving neural machine translation with conditional sequence generative adversarial nets[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana:Association for Computational Linguistics ,2018:1346-1355 .

[23] LIN K, LI Dianqi, HE Xiaodong, et al. Adversarial ranking for language generation [C]// Advances in Neural Information Processing Systems. Long Beach, CA:NIPS Foundation, 2017: 3155-3165.

[24] WANG J, YU L, ZHANG W, et al. Irgan: A minimax game for unifying generative and discriminative information retrieval models [C]// 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Shinjuku: ACM, 2017:515-524.

[25] QIAO Xiuming, ZHANG Yue, ZHAO Tiejun. Learning domain invariant word representations for parsing domain adaptation [C]// International Conference on Natural Language Processing and Chinese Computing (NLPCC). Dunhuang, China: dblp, 2019: 801-813.

(下转第 189 页)