

# 本科生毕业设计（论文）参考文献译文本

译文出处：Conditional Generative Adversarial Nets

Submitted on 6 Nov 2014

Mehdi Mirza, Simon Osindero

arXiv:1411.1784 [cs.LG]

院 系 计算机科学与技术

专业班级 计算机1604

姓 名 孟嵩淼

学 号 U201614613

指导教师 吴非

2020年 3月

**译文要求**

1. 译文内容须与课题（或专业内容）联系，并需在封面注明详细出处。
2. 出处格式为

图书：作者.书名.版本（第×版）.译者.出版地：出版者，出版年.起页～止页  
期刊：作者.文章名称.期刊名称，年号，卷号（期号）：起页～止页

1. 译文不少于5000汉字（或2万印刷符）。
2. 翻译内容用五号宋体字编辑，采用A4号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。
3. 译文及其相应参考文献一起装订，顺序依次为封面、译文、文献。
4. 翻译应在第七学期完成。

**译文评阅**

|  |
| --- |
| **导师评语**  应根据学校“译文要求”，对学生译文翻译的准确性、翻译数量以及译文的文字表述情况等做具体的评价后，再评分。 |
| 评分：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_（百分制） 指导教师（签名）：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  年 月 日 |

**条件生成对抗网络**

**摘要**

**生成对抗网络[8]**最近被引入作为一种训练生成模型的新方法。在这项工作中，我们介绍了生成对抗网络的条件版本，可以通过简单地馈送数据y来构造它，我们希望以生成器和判别器为条件。我们证明了该模型可以生成以类别标签为条件的MNIST数字。我们还将说明如何使用此模型来学习多模式模型，并提供图像标记应用的初步示例，其中我们将演示该方法如何生成不属于训练标签的描述性标签。

1. **前言**

生成对抗性网络最近被引入作为训练生成模型的替代框架，以规避近似许多棘手的概率计算的困难。

对抗网络的优势在于，不再需要马尔可夫链，仅使用反向传播即可获得梯度，在学习过程中无需进行推理，并且可以轻松地将各种因素和相互作用纳入模型。

此外，如[8]中所示，它可以产生最新的对数似然估计和现实样本。

在无条件的生成模型中，无法控制所生成数据的模式。但是，通过在附加信息上对模型进行调节，可以指导数据生成过程。这种条件可以基于类标签，基于数据的某些部分（例如[5]），甚至基于不同模态的数据。

在这项工作中，我们展示了如何构建条件对抗网络。对于实验结果，我们演示了两组实验。一种基于类别标签的MNIST数字数据集，另一种基于MIR Flickr 25,000数据集[10]，用于多模式学习。

1. **相关研究**
   1. **图像标注的多模式学习**

尽管有监督神经网络（尤其是卷积网络）最近取得了许多成功[13，17]，但要扩展此类模型以容纳大量预测的输出类别仍然具有挑战性。 第二个问题是，迄今为止，许多工作都集中在学习从输入到输出的一对一映射。 但是，许多有趣的问题更自然地被认为是概率一对多映射。 例如，在图像标记的情况下，可能有许多不同的标签可以适当地应用于给定图像，并且不同的（人类）注释者可以使用不同的（但通常是同义词或相关的）术语来描述相同的图像。

解决第一个问题的一种方法是利用其他形式的其他信息：例如，通过使用自然语言语料库来学习标签的矢量表示，其中几何关系在语义上有意义。 在这样的空间中进行预测时，我们受益于以下事实：当预测错误时，我们仍然经常“接近”真相（例如，预测“桌子”而不是“椅子”），以及我们可以自然地 对培训期间未看到的标签进行预测性归纳。 文献[3]的研究表明，即使是从图像特征空间到单词表示空间的简单线性映射也可以提高分类性能。

解决第二个问题的一种方法是使用条件概率生成模型，将输入作为条件变量，并将一对多映射实例化为条件预测分布。

[16]采用类似的方法来解决这个问题，并像我们在这项工作中那样，在MIR Flickr 25,000数据集上训练多模式Deep Boltzmann机。

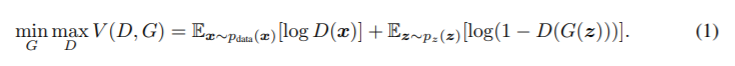
另外，在[12]中，作者展示了如何训练监督多模式神经语言模型，并且他们能够为图像生成描述性句子。

1. **条件对抗网络**
   1. **条件对抗网络**

生成对抗网络最近被引入作为一种训练生成模型的新方法。 它们由两个“对抗性”模型组成：生成模型G捕获数据分布，以及区分模型D估计样本来自训练数据而非G的概率。G和D都可能是非线性的 映射功能，例如多层感知器。

为了了解数据数据x上的生成器分布，生成器建立了从先验噪声分布(z)到数据空间的映射函数，即G(z;)。 判别器D(x;输出单个标量，表示x来自训练数据而不是的概率。

G和D都同时训练：我们调整G的参数以使log(1 – D(G(z))最小化，并调整D的参数以使logD(X)最小化，就像它们遵循两人的最小-最大游戏一样 具有值函数V(G, D)：



* 1. **条件对抗网络**

如果生成者和判别者都以某些额外信息y为条件，则生成对抗网络可以扩展为条件模型。 y可以是任何种类的辅助信息，例如类标签或其他形式的数据。 我们可以通过将y作为附加输入层输入到鉴别器和生成器中来执行调节。

在生成器中，先验输入噪声(z)和y被组合为联合隐藏表示，而对抗训练框架则在如何构成该隐藏表示时提供了相当大的灵活性。

在鉴别器中，x和y表示为输入和鉴别函数（在这种情况下，又由MLP体现）。

两人min-max博弈的目标函数为(2)



图1说明了简单的条件对抗网络的结构。

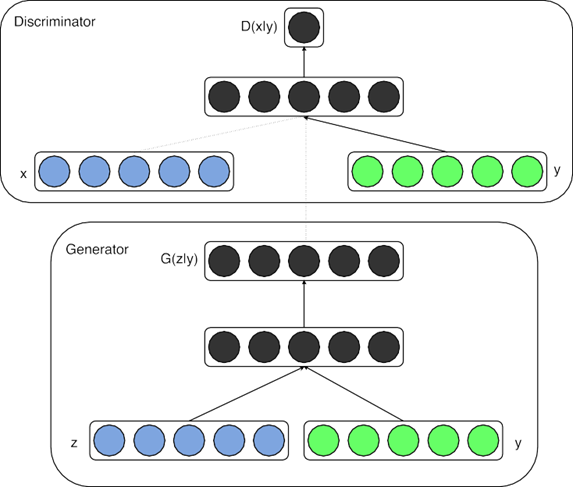


图1 条件生成网络

1. **实验结果**

**4.1 单峰**

我们在以类别标记为条件的MNIST图像上训练了条件对抗网络，将其编码为one-hot向量。

在生成网络中，从单元超立方体内的均匀分布中提取了维度为100的先验z噪声。 在将z和y映射到第二个维度为1200的组合隐藏ReLu层之前，将z和y都分别映射到具有整流线性单元（ReLu）激活[4、11]的隐藏层，层大小分别为200和1000。 最后的S型单元层作为生成784维MNIST样本的输出。

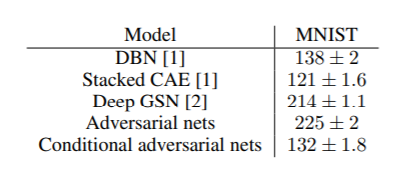


表1：MNIST的基于Parzen窗口的对数似然估计。我们遵循与[8]相同的步骤计算这些值。

鉴别符将x映射到具有240个单位和5个片段的maxout [6]图层，将y映射到具有50个单位和5个片段的maxout图层。这两个隐藏层都映射到240个单元和4个块的联合maxout层，然后再馈送到S形层。 （鉴别器的精确架构并不重要，只要它具有足够的功能即可；我们发现maxout单元通常非常适合该任务。）

该模型使用大小为100的小批量和初始学习率为0.1的随机梯度体面训练，该学习率呈指数下降至.000001，衰减因子为1.00004。还使用了初始值为0.5的动量，该动量增加至0.7。退出

[9]的概率为0.5应用于生成器和鉴别器。验证集上对数似然率的最佳估计用作停止点。

表1显示了MNIST数据集测试数据的高斯Parzen窗口对数似然估计。从每10个类别中抽取1000个样本，并为这些样本拟合高斯Parzen窗口。然后，我们使用Parzen窗口分布估计测试集的对数似然率。（有关如何构建此估计的更多详细信息，请参见[8]。）

我们提供的条件对抗网结果与其他基于网络的对抗网具有可比性，但在其他几种方法（包括非条件对抗网络）上却表现出色。我们将这些结果更多地作为概念验证而不是有效性的证明，并相信随着对超参数空间和体系结构的进一步探索，条件模型应匹配或超过非条件结果。

图2显示了一些生成的样本。每行以一个标签为条件，每列为不同的生成样本。



图2：生成的MNIST数字，每一行以一个标签为条件

**4.2 多模式**

Flickr之类的照片网站以图像及其相关联的用户生成的元数据（UGM）（尤其是用户标签）的形式提供了丰富的标签数据。

用户生成的元数据与更具“规范性”的图像标签方案不同，因为它们通常更具描述性，并且在语义上更接近于人类用自然语言描述图像的方式，而不仅仅是识别图像中存在的对象。 UGM的另一个方面是同义词很普遍，不同的用户可能使用不同的词汇来描述相同的概念-因此，拥有一种有效的方法来标准化这些标签变得很重要。概念词嵌入[14]在这里可能非常有用，因为相关概念最终由相似的矢量表示。

在本节中，我们将演示如何使用条件对抗网络以多标签预测对图像进行自动标记，以及如何使用图像特征进行条件下的标记矢量分布（可能是多模式）。

对于图像特征，我们在具有21,000个标签[15]的完整ImageNet数据集上预训练与[13]中的模型相似的卷积模型。我们使用具有4096个单位的最后一个完全连接层的输出作为图像表示。

对于世界表示，我们首先从YFCC100M 2数据集元数据的用户标签，标题和描述的连接中收集文本语料库。在对文本进行预处理和清理之后，我们训练了单词向量大小为200的跳过语法模型[14]。而且，我们从词汇表中删除了出现次数少于200次的任何单词，从而得到了大小为247465的字典。在对抗网络训练期间，我们保持卷积模型和语言模型固定不变。当我们甚至通过这些模型向后传播时，请离开实验，作为将来的工作。

对于我们的实验，我们使用MIR Flickr 25,000数据集[10]，并使用上述卷积模型和语言模型提取图像和标签特征。实验中省略了没有任何标签的图像，并且注释被视为额外的标签。前150,000个示例用作训练集。对于每个相关联的标签，在训练集中重复具有多个标签的图像一次。

为了进行评估，我们为每个图像生成100个样本，并使用词汇表中单词与每个样本的矢量表示的余弦相似度来查找最接近的20个单词。然后，在所有100个样本中选择前10个最常见的词。表4.2列出了一些由用户分配的标签和注释以及生成的标签的示例。

效果最好的模型的生成器会先接收大小为100的高斯噪声作为噪声，然后将其映射到500维ReLu层。并将4096维图像特征向量映射到2000维ReLu隐藏层。这两个层都映射到200维线性层的联合表示，它将输出生成的单词向量。

鉴别器由分别用于单词向量和图像特征的500和1200维ReLu隐藏层以及具有1000个单位和3个片断的maxout层作为连接层组成，最后被馈送到一个单一的S型单元。

该模型使用大小为100的小批量和初始学习率为0.1的随机梯度体面训练，该学习率呈指数下降至.000001，衰减因子为1.00004。还使用了初始值为0.5的动量，该动量增加至0.7。概率为0.5的压降应用于生成器和鉴别器。

超参数和体系结构选择是通过交叉验证以及随机网格搜索和手动选择（尽管搜索空间有限）获得的。

1. **未来工作**

本文显示的结果是非常初步的，但它们证明了条件对抗网络的潜力，并显示出有趣且有用的应用前景。

在从现在到研讨会的未来探索中，我们希望展示更复杂的模型，并对它们的性能和特性进行更详细和透彻的分析。

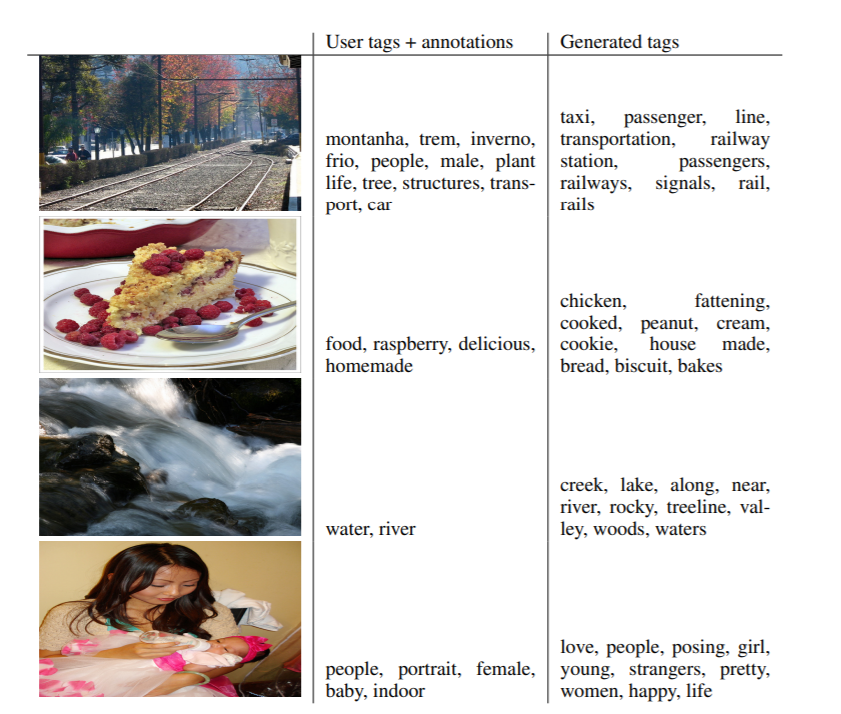


表2：生成的标签样本

另外，在当前的实验中，我们仅单独使用每个标签。 但是，通过同时使用多个标签（有效地将生成问题视为“集合生成”之一），我们希望能获得更好的结果。

未来工作的另一个明显方向是构建联合培训计划以学习语言模型。 文献[12]表明，我们可以学习一种适合于特定任务的语言模型。

**致谢**

该项目是在Pylearn2 [7]框架中开发的，我们要感谢Pylearn2开发人员。 我们还要感谢伊恩·古德费洛（Ian Goodfellow）在加入蒙特利尔大学期间的有益讨论。 作者非常感谢Flickr的视觉与机器学习和生产工程团队的支持（按字母顺序排列：Andrew Stadlen，Arel Cordero，Clayton Mellina，Cyprien Noel，Frank Liu，Gerry Pesavento，Huy Nguyen，Jack Culpepper，John Ko ，Pierre Garrigues，Rob Hess，Stacey Svetlichnaya，Tobi Baumgartner和Ye Lu）。

**引用**

[1] Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. (2013). Better mixing via deep representations. In ICML’2013.

[2] Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014). Deep generative stochastic net- works trainable by backprop. In Proceedings of the 30th International Conference on Machine Learning (ICML’14).

[3] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems, pages 2121– 2129.

[4] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In International Conference on Artificial Intelligence and Statistics, pages 315–323.

[5] Goodfellow, I., Mirza, M., Courville, A., and Bengio, Y. (2013a). Multi-prediction deep boltzmann ma- chines. In Advances in Neural Information Processing Systems, pages 548–556.

[6] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013b). Maxout networks. In ICML’2013.

[7] Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., and Bengio, Y. (2013c). Pylearn2: a machine learning research library. arXiv preprint arXiv:1308.4214.

[8] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In NIPS’2014.

[9] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580.

[10] Huiskes, M. J. and Lew, M. S. (2008). The mir flickr retrieval evaluation. In MIR ’08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval, New York, NY, USA. ACM.

[11] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In ICCV’09.

[12] Kiros, R., Zemel, R., and Salakhutdinov, R. (2013). Multimodal neural language models. In Proc. NIPS Deep Learning Workshop.

[13] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (NIPS’2012).

[14] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In International Conference on Learning Representations: Workshops Track.

[15] Russakovsky, O. and Fei-Fei, L. (2010). Attribute learning in large-scale datasets. In European Confer- ence of Computer Vision (ECCV), International Workshop on Parts and Attributes, Crete, Greece.

[16] Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep boltzmann machines. In NIPS’2012.

[17] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabi- novich, A. (2014). Going deeper with convolutions. arXiv preprint arXiv:1409.4842.