

---

# 条件生成式对抗网络

---

Mehdi Mirza  
Département d'informatique et de recherche opérationnelle  
Université de Montréal  
Montréal, QC H3C 3J7  
mirzamom@iro.umontreal.ca

Simon Osindero  
Flickr / Yahoo Inc.  
San Francisco, CA 94103  
osindero@yahoo-inc.com  
翻译: 张兴园 (初), 路转 (复), 管枫 (审)

## Abstract

生成式对抗网络 (GAN)[8] 是一种用来训练生成式模型的新方法。本文中, 我们在 GAN 的基础之上引入条件生成式对抗网络, 它的构建并不复杂, 只需要在生成模型与判别模型的构建中分别输入代表条件的数据  $y$ 。结果显示此模型能够在类别标签条件下生成 MNIST 手写体数字。本文同时说明了本模型如何用来学习多模态模型, 给出一个图像标记应用的初步示例, 在这个示例中我们演示了本文方法如何生成训练标签之外的描述性标签。

## 1 Introduction

Generative adversarial nets were recently introduced as an alternative framework for training generative models in order to sidestep the difficulty of approximating many intractable probabilistic computations.

Adversarial nets have the advantages that Markov chains are never needed, only backpropagation is used to obtain gradients, no inference is required during learning, and a wide variety of factors and interactions can easily be incorporated into the model.

Furthermore, as demonstrated in [8], it can produce state of the art log-likelihood estimates and realistic samples.

In an unconditioned generative model, there is no control on modes of the data being generated. However, by conditioning the model on additional information it is possible to direct the data generation process. Such conditioning could be based on class labels, on some part of data for inpainting like [5], or even on data from different modality.

In this work we show how can we construct the conditional adversarial net. And for empirical results we demonstrate two set of experiment. One on MNIST digit data set conditioned on class labels and one on MIR Flickr 25,000 dataset [10] for multi-modal learning.

## 2 Related Work

### 2.1 多模态学习实现图像标记 (标注)

尽管近来监督神经网络取得很大成功 (尤其是卷积神经网络) [13, 17], 但是对于使用上述模型来预测极其大量的输出类别仍然是一个挑战。其次, 迄今为止的大部分工作集中于学习

从输入到输出的一对一映射。然而，许多有趣的问题在概率上是一对多的映射问题。例如，图像标记问题，对于给定的一副图像，可能有不同的标签，同时不同的标注者可能使用不同的（但是同义或相关的）词语来描述同一副图像。

解决第一个问题的一种方式是利用其它模态的额外信息：例如，可以使用自然语言语料库来训练标签的向量表示，并保证向量之间的距离远近可以表示语义上含义的远近。在这样的空间进行预测的一个好处是，即使预测有偏差预测结果也能跟真实情况比较“接近”（比如“table”和“chair”），同时这种方法是我们可以预测那些在训练集中没有出现的标签。工作[3]表明，即使一个从图像特征空间到词表示空间的简单的线性映射也能改善分类效果。

解决第二个问题的方式就是使用条件概率生成模型，在这种模型中输入的是条件变量，这样以来预测一对多的映射就变成了预测条件分布。

[16] 采用了类似的方法来解决上述问题，在 MIR Flickr25,000 数据集上训练了一个多模态深度玻尔兹曼机。

此外，[12] 中提到了如何训练一个有监督的多模态神经语言模型，同时能够生成图像的描述性句子。

### 3 条件对抗网络

#### 3.1 生成式对抗网络

生成式对抗网络是一种最近引入用来训练生成式模型的全新方式。它由两个“对抗式”模型组成：生成式模型  $G$  来获取数据分布，判别式模型  $D$  用来估计一个样本来自与训练集而不是来自于  $G$  的概率。 $G$  和  $D$  都是非线性映射函数，例如多层感知器。

为了学习数据  $\mathbf{x}$  的生成式分布  $p_g$ ，生成器构建一个从先验噪声分布  $p_z(z)$  数据空间的映射函数  $G(z; \theta_g)$ 。判别器  $D(x; \theta_d)$  则输出单个标量来表示  $\mathbf{x}$  来自训练数据而不是  $p_g$  的概率。

$G$  和  $D$  同时进行训练：我们针对  $G$  调整参数来最小化  $\log(1 - D(G(z)))$  同时针对  $D$  调整参数来最小化  $\log D(X)$ ，如同两个玩家使用如下的价值函数  $V(G, D)$  玩最小最大游戏：

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (1)$$

#### 3.2 条件对抗网

如果生成器和判别器都共同的额外条件变量  $\mathbf{y}$ ，生成式对抗网络就可以够扩展成条件模型。 $\mathbf{y}$  可以是任何类型的辅助信息，比如类别标签或者其他模态的数据。我们通过将  $\mathbf{y}$  作为额外输入层导入到判别器和生成器来实现条件模型。

在生成器中，先验输入的噪声  $p_z(\mathbf{z})$  和  $\mathbf{y}$  在隐藏层中相结合，这使得对抗网络训练框架在如何构成隐藏层方面具有了很大的灵活性。<sup>1</sup>

在判别器中  $\mathbf{x}$  和  $\mathbf{y}$  共同做为判别函数的输入（这里仍然考虑在 MLP 中进行实现的情形）。

此处双玩家最小最大游戏的目标函数如下 2：

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x}|\mathbf{y})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z}|\mathbf{y})))] \quad (2)$$

图 1 展示了简单的条件对抗网络的结构。

<sup>1</sup>在本文中，我们只考虑条件变量与先验噪音共同作为多层感知器的单一隐藏层的输入的情形。但是条件变量与先验噪音的高层交互作用可以产生更复杂的生成器，而这种交互在传统的生成器中是很难实现的

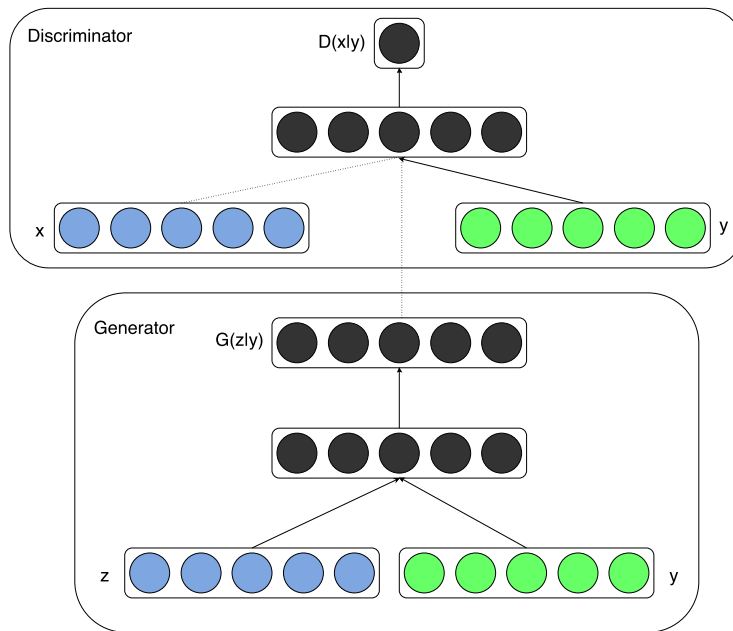


Figure 1: 条件对抗网络

## 4 实验结果

### 4.1 单模态

本文使用标签类别作为条件变量，对编码成 one-hot 向量的 MNIST 图像进行条件对抗网络的训练。

在生成式网络部分，我们从单位超立方体上的均匀分布中提取 100 维的先验噪音  $z$ 。先通过激活函数 ReLu(Rectified Linear Unit)[4, 11] 把先验噪音  $z$  和条件变量  $y$  映射到尺寸分别为 200 和 1000 的隐藏层，然后把这两个再映射到 1200 维的组合 ReLu 隐藏层，最终在输出层使用 sigmoid 单元层来生成 784 维 MNIST 样本。

判别器把  $x$  映射到具有 240 个单元 5 块的 maxout [6] 层，把  $y$  映射到具有 50 个单元 5 块的 maxout 层。然后把这两个隐藏层都映射到一个具有 240 个单元 4 块的联合隐藏 maxout 层，最后再把这个层的输出导入 sigmoid 层。(判别器的具体构造并不关键，只要它具有相当的判别能力就可以，我们实践发现 maxout 单元比较适合这个任务)

模型的训练使用随机梯度下降法，其中每个 mini-batch 的大小为 100，学习率初始值为 0.1，然后以 1.00004 的下降率指数下降到 .000001，同时初始动量为 .5，最终增长至 0.7。为了防止过度拟合，我们在生成器和判别器中的都使用了概率为 0.5 的 Dropout 层 [9]。

表格 1 给出了对于 MNIST 数据集的测试数据做的 Gaussian Parzen window 对数-似然估计结果。我们从十个类别中每个类别抽取出来 1000 个样本进行了 Gaussian Parzen window 拟合，然后使用 Parzen window 分布对这个测试集进行对数似然估计。([8] 中有关于如何构造这个估计的更详细讨论)

我们这里使用条件对抗网络得到的结果与一些基于 network 的构架得出的结果相当，但是却不如一些其他的方法 (比如非条件对抗网络)。我们此处的结果更多的是作为一个概念验

Model	MNIST
DBN [1]	$138 \pm 2$
Stacked CAE [1]	$121 \pm 1.6$
Deep GSN [2]	$214 \pm 1.1$
Adversarial nets	$225 \pm 2$
Conditional adversarial nets	$132 \pm 1.8$

Table 1: 基于 Parzen window 对 MNIST 进行的对数似然估计。我们使用了和 [8] 中相同的方法来进行计算。

证，但是相信对超参数以及架构的进一步探索可以使得条件生成网络最终达到或者超过非条件生成网络的水平。

图 2 给出了一些生成的样本，每一行代表一个条件标签，而每一列代表一个生成样本。

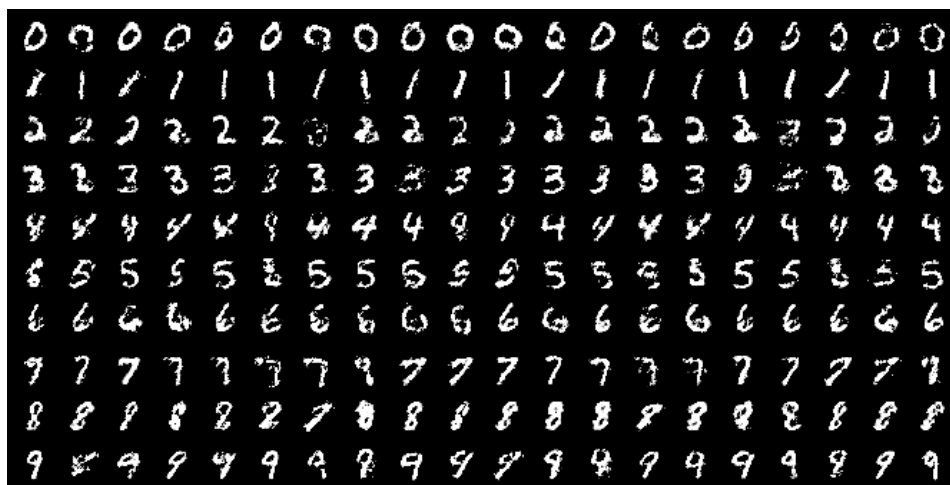


Figure 2: 生成的 MNIST 手写数据样本，每一行代表一个标签

## 4.2 Multimodal

Photo sites such as Flickr are a rich source of labeled data in the form of images and their associated user-generated metadata (UGM) — in particular user-tags. 类似于 Flickr 这样的照片网站提供了大量的带标记图像数据以及相关的用户生成的元数据 (UGM) 特别是用户标签。

User-generated metadata differ from more ‘canonical’ image labelling schemes in that they are typically more descriptive, and are semantically much closer to how humans describe images with natural language rather than just identifying the objects present in an image. Another aspect of UGM is that synonymy is prevalent and different users may use different vocabulary to describe the same concepts — consequently, having an efficient way to normalize these labels becomes important. Conceptual word embeddings [14] can be very useful here since related concepts end up being represented by similar vectors.

用户生成的元数据不同于“规范的”图像标记，因为他们更具有描述性，同时在语义上与人类使用自然语言而不是识别图像中存在的目标实现对图像进行描述更加接近。UGM 的另一个方面是同义词非常普遍，同时不同的用户可能使用不同的词汇来描述同一个概念。因此，使用有效的方法来标准化这些标签也就变得非常重要。概念词嵌入方法 [14] 就 very 有效，因为这种方法使得相关的概念最终由相似的向量表示。

In this section we demonstrate automated tagging of images, with multi-label predictions, using conditional adversarial nets to generate a (possibly multi-modal) distribution of tag-vectors conditional on image features.

我们在本节中以图像特征为条件变量，使用条件生成网络来生成标签向量的条件分布。并以此实现图像的多标签自动标注。

For image features we pre-train a convolutional model similar to the one from [13] on the full ImageNet dataset with 21,000 labels [15]. We use the output of the last fully connected layer with 4096 units as image representations.

对于图像特征，我们预先在带有 21,000 个标签的完整 ImageNet 数据集 [15] 上训练一个类似于 [13] 中的卷积模型。然后使用其具有 4096 个单元的最后一个全连接层对图像特征进行表示。

For the word representation we first gather a corpus of text from concatenation of user-tags, titles and descriptions from YFCC100M<sup>2</sup> dataset metadata. After pre-processing and cleaning of the text we trained a skip-gram model [14] with word vector size of 200. And we omitted any word appearing less than 200 times from the vocabulary, thereby ending up with a dictionary of size 247465.

对于单词表示，首先从 YFCC100M<sup>3</sup> 元数据集中收集了一个混合了用户标签、标题以及描述文本的语料库。经过预处理和文档清理，本文使用大小为 200 的单词向量进行 skip-gram 模型拟合。我们过滤掉出现次数少于 200 次的单词，从而得到一个最终大小为 247465 的单词表。

We keep the convolutional model and the language model fixed during training of the adversarial net. And leave the experiments when we even backpropagate through these models as future work.

我们在训练对抗网络时保持这个卷积模型和语言模型，并把基于这些模型的反向传播算法实验留作今后的工作。

For our experiments we use MIR Flickr 25,000 dataset [10], and extract the image and tags features using the convolutional model and language model we described above. Images without any tag were omitted from our experiments and annotations were treated as extra tags. The first 150,000 examples were used as training set. Images with multiple tags were repeated inside the training set once for each associated tag.

本文的实验对 MIR Flickr 25,000 数据集 [10]，使用如上的卷积模型与语言模型提取了图像与标注特征。我们在实验中过滤掉了没有任何标注的图像，而将附注 (annotations) 作为额外的标注。实验选取前 150,000 个例子作为训练集。有多个标注的图像在训练集中重复出现 (每一个标注重复一次)。

For evaluation, we generate 100 samples for each image and find top 20 closest words using cosine similarity of vector representation of the words in the vocabulary to each sample. Then we select the top 10 most common words among all 100 samples. Table 4.2 shows some samples of the user assigned tags and annotations along with the generated tags.

作为测试，我们对每个图像生成 100 个样本，并且在每一个样本中使用余弦相似函数选取前 20 个最接近的词语。然后在 100 个样本中选取前 10 个出现最多的词。表格 4.2 给出了一些用户标注与附注跟生成标注的对比。

The best working model's generator receives Gaussian noise of size 100 as noise prior and maps it to 500 dimension ReLu layer. And maps 4096 dimension image feature vector to 2000 dimension ReLu hidden layer. Both of these layers are mapped to a joint representation of 200 dimension linear layer which would output the generated word vectors.

The discriminator is consisted of 500 and 1200 dimension ReLu hidden layers for word vectors and image features respectively and maxout layer with 1000 units and 3 pieces as the join layer which is finally fed to the one single sigmoid unit.

The model was trained using stochastic gradient decent with mini-batches of size 100 and initial learning rate of 0.1 which was exponentially decreased down to .000001 with decay

---

<sup>2</sup>Yahoo Flickr Creative Common 100M <http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>.

<sup>3</sup>Yahoo Flickr Creative Common 100M <http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>.

factor of 1.00004. Also momentum was used with initial value of .5 which was increased up to 0.7. Dropout with probability of 0.5 was applied to both the generator and discriminator.

The hyper-parameters and architectural choices were obtained by cross-validation and a mix of random grid search and manual selection (albeit over a somewhat limited search space.)





	User tags + annotations	Generated tags
	montanha, trem, inverno, frio, people, male, plant life, tree, structures, transport, car	taxi, passenger, line, transportation, railway station, passengers, railways, signals, rail, rails
	food, raspberry, delicious, homemade	chicken, fattening, cooked, peanut, cream, cookie, house made, bread, biscuit, bakes
	water, river	creek, lake, along, near, river, rocky, treeline, valley, woods, waters
	people, portrait, female, baby, indoor	love, people, posing, girl, young, strangers, pretty, women, happy, life

Table 2: Samples of generated tags

## 5 Future Work

The results shown in this paper are extremely preliminary, but they demonstrate the potential of conditional adversarial nets and show promise for interesting and useful applications.

In future explorations between now and the workshop we expect to present more sophisticated models, as well as a more detailed and thorough analysis of their performance and characteristics.

Also, in the current experiments we only use each tag individually. But by using multiple tags at the same time (effectively posing generative problem as one of ‘set generation’) we hope to achieve better results.

Another obvious direction left for future work is to construct a joint training scheme to learn the language model. Works such as [12] has shown that we can learn a language model for suited for the specific task.

## Acknowledgments

This project was developed in Pylearn2 [7] framework, and we would like to thank Pylearn2 developers. We also like to thank Ian Goodfellow for helpful discussion during his affiliation at University of Montreal. The authors gratefully acknowledge the support from the Vision & Machine Learning, and Production Engineering teams at Flickr (in alphabetical order: 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, and Ye Lu).

## References

- [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 networks 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 machines. 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 Conference 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 Rabinovich, A. (2014). Going deeper with convolutions. arXiv preprint arXiv:1409.4842.