解析：Google开源的“Show and Tell”，是如何让机器“看图说话”的？

本文作者：图普科技 2016-09-29 18:39

导语：Google Brain宣布在TensorFlow 上开源了其最新版的自动图像描述系统“Show and Tell”。

雷锋网按：9月23日，Google Brain宣布在TensorFlow 上开源了其最新版的自动图像描述系统“Show and Tell”，采用采用编码器-解码器神经网络架构，可以根据新的场景生成准确的新图说。作者系图普科技工程师，本文将会解析机器如何“看图说话”？以及，Google为什么要开源其系统？雷锋网(公众号：雷锋网)独家文章。

机器的Image Captioning（自动图像描述）能力

电影《HER》中的“萨曼莎”是一款基于AI的OS系统，基于对西奥多的手机信息和图像内容的理解，“她”可以为他处理日常事物、可以陪他谈心、甚至进行Virtural Sex，还可以读懂所有的书、跟哲学家交流，“她”所做的一切俨然就是一个有血有肉的人类才能实现的。但萨曼莎还胜于人类，她能够同时和8316个使用者聊天，和641个使用者in love，并且对每个人都是真情实感。

电影的“她”是人类想象中的强AI，“她”有思维，具备比人还强的智力以及运算能力，虽然目前的AI还不能完全做到“她”那样强，但近年来人工智能技术的发展让机器可以越来越像人类，计算机开始能够理解更高层次的图像内容，“看图说话”似乎不再是专属于人类的专利。

在人工智能领域，这是机器的 Image Captioning（自动图像描述）能力。

从表现上看，机器不仅需要能够知道图像中包括哪些物体，同时还必须描述物体之间的联系以及它们的属性和参与的活动，这显然是机器一种更加高级的智能形态。如下图：

图1. Automatic image caption 的例子

从原理上看，这依赖于智能的两个部分：“看”和“语言表达”，分别对应人工智能最重要的两大子领域： 机

TUPUTECH

图普科技

专栏作者

基于图像识别技术多维度解读图片和视频

发私信

当月热门文章

我们如何利用AI和机器学习将游戏引入现实生活？

最新文章

人机交互的重点是“机器”？CCF-ADL专家详解人类心理如何影响人机交互的打造

极限元算法专家：深度学习在语音生成问题上的典型应用 | 学术分享会总结文

这一次，用Geek的方式说爱你

教你如何利用算法原理，让TA对你一见钟情

美国《科学》杂志记者眼中的神经网络

谷歌开发大使Joshua Gordon：“所谓深度学习，就是让算法帮你解决问题”

器视觉和自然语言处理。

机器视觉和自然语言处理从来都不是相互割裂的，两者技术上相互借鉴历史由来已久，更重要的是，从一个完整的智能系统上看，无论是现在的人类智能还是终极机器的智能，多模态的融合是一项必然的要求，视觉和语言理解表达缺一不可，两者相互协助，共同产生高级智能。

所以图像自动描述能力作为两个智能领域的关键性连接，必然是人工智能领域最顶尖的研究者最密切关注的任务之一。虽然图像自动描述并不是一个新兴的任务，在此领域中已经积累了大量的研究工作，但在2015年，此任务才得到了一个颠覆性的突破，机器自动描述图像的能力在某些案例上的表现会让人产生一种强人工智能即将要实现的错觉。

9月23日，Google Brain宣布在TensorFlow 上开源了最新版的自动图像描述系统“Show and Tell”，成功地将机器这方面的能力提高到一个新台阶。在这之前的版本，更多的是告诉大家图像里面有什么或者总是重复使用人类在训练模型时使用的描述语言，对于图像中的物体之间以及物体和环境之间的关联、意义并不能给出满意的描述。

而“Show and Tell”在遇见全新的场景时，能够基于图像中物体和环境之间的交互联系，自动生成更加准确的图像描述，并且使用的自然语言更加流畅，与人类的表述差异无几。

那么Google是如何做到这样效果？要弄清其中的原理，我们需要先了解下在如今的深度学习时代，引领机器视觉和自然语言处理两个领域取得突破的最重要的两个技术，分别是：DCNN（Deep Convolutional Neural Network，深度卷积网络）与LSTM（Long Short Term Memory，长短时记忆网络）。

DCNN与LSTM（深度卷积网络与长短时记忆网络）

在自然语言处理领域，许多高难度的任务都可以归结进序列到序列（sequence to sequence）的框架中。

比如说，机器翻译任务表面上是将一种语言转换为另一种语言，本质上就是从一段不定长的序列转换为另一段不定长的序列。如今实现seq2seq最有效的方法即为LSTM，一种带门的RNN（Recurrent Neural Network，递归神经网络)，它可以将源语言编码为一个固定长度含丰富语义的向量，然后作为解码网络的隐藏状态去生成目标语言。而Image Caption Generator（自动图像生成器）方法正是受到机器翻译中seq2seq进展的启发：何不将源语言信号替换成图像信号，这样就能够将机器翻译的任务转换也就是把图像转成自然语言，即图像自然语言描述。

可是简单地将图像信号直接作为输入是无法达到很好的效果，原因是原始的图像信号并不是一个紧致的表示，含有太多的噪声。所以需要引入DL（Deep Learning，深度学习）在机器视觉中最核心的部件：CNN（Convolutional Neural Network，卷积网络）。

在DCNN的高层神经元输出可以表示图像的紧致的高层语义信息，如今众多成功的机器视觉应用都得益于此，比如前段时间爆红的Prisma（《AI修图艺术：Prisma背后的奇妙算法》），其texture transfer（风格转换）算法正是巧妙的利用了含有高层语义的图像表示。

所以此图像文字描述方法的基本思想就是利用了DCNN生成图像的高层抽象语义向量，将其作为语言生成模型LSTM的输入进行sequence to sequence的转换，其结构图如下：

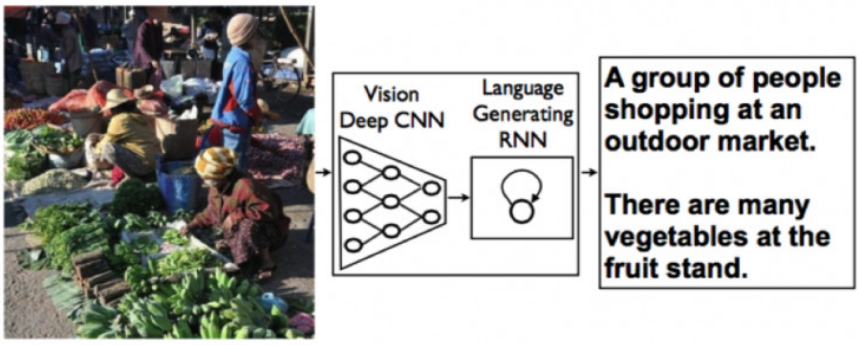


图2. 系统结构

此方法的巧妙之处在于将视觉和自然语言处理领域中最先进的两类网络连着在一起，各自负责其擅长的部分，同时进行端到端的训练学习。

Image Caption的神经网络学习可以用数学公式概括为：

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

其中I为图片，S为生成的句子，θ为网络需要学习的参数，这个公式的含义指的是：学习最佳的网络参数θ最大化在给定图片下其生成正确描述的概率。同时由于语言句子的长度是不定长的，所以一般将其概率用链式法则写成：

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

其中N为句子的长度，S_i为句子的每一个词。更具体的网络形式为下图：

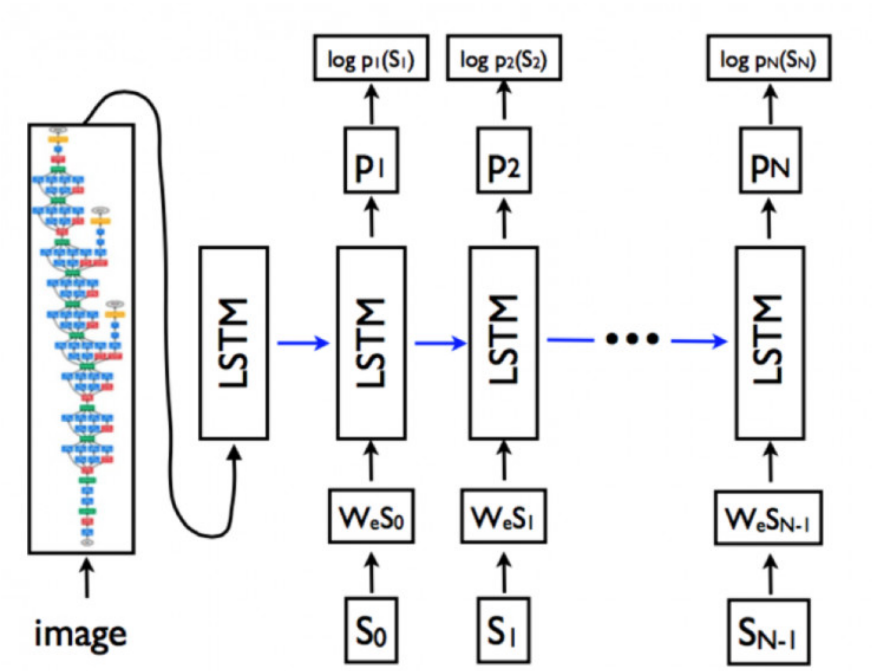


图2. 语言模型LSTM，图像模型CNN和词嵌入模型

上图将LSTM的recurrent connection（复现连接）以更加直观的展开形式画出来，在网络训练过程中，目标可以写为以下的损失函数：

$$L(I,S) = -\sum_{t=1}^{N} \log p_t(S_t) \ .$$

其目标是更新LSTM、CNN和词嵌入模型的参数，使得每一个正确的词出现的概率最大，也就是让此loss函数越小。除了LSTM、CNN模型的选择和词嵌入模型都会极大影响最后的效果，此方法最早发明时，最好的DCNN是2014年ImageNet竞赛的冠军网络GoogLeNet。而后，随着更强的CNN网络Inception V1到V3系列的出现，作者发现在此框架的Image Caption的效果也随之变得更好。这也是必然的，因为更强的CNN网络意味着输出的向量表示可以做到更好的图像高层语义表示。

作者在其开源的Tensorflow项目中号召大家去尝试现在最强的CNN分类网络Inception-Resnet-V2，看看是否会有效果的继续提升。对于词嵌入模型，最简单的方式是 one-hot-encoding的方法（向量中代表词的维度为1，其余为0），而此方法使用了一个更复杂的词嵌入模型，使得词嵌入模型也可以随着其他两个网络一起训练，训练出来的词嵌入模型表示被发现可以获取到自然语言的一些统计特性，比如以下的词在学习到的空间中是非常相近的，这符合自然语言中这些词的距离。

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

图4. 一些词在嵌入空间中的相近词

在最早的版本中，CNN模型使用的是在ImageNet数据库上预训练好的分类模型，在Image caption训练过程中其参数是不做更新的。而在最新的方法中，作者称在训练过程中更新CNN最高层的权重可以产生更好的效果，不过这些参数的更新需要在LSTM更新稳定后才能进行，不然LSTM的噪声会对CNN模型造成不可逆的影响。

视觉模型和语言生成模型进行端到端的联合训练有利于相互提升效果。例如在CNN模型中，可以将图像中更有利于“描述”而不是用于“分类”的信息迁移给语言模型，由于ImageNet的训练数据的类别空间中比较缺少颜色信息，所以在不使用联合训练的CNN模型的2015 CVPR版本中，并不会生成类似于“一辆蓝色和黄色的火车”这样的描述。当进行联合训练后，caption模型可以生成更精确、更细节化的句子，如下图所示：

图5. 初始模型和最新模型生成句子的对比

这让人会不禁产生一个疑问：现在的模型是否真的学会对图片中未曾见过的情境和交互生成全新的描述，还是只是简单的复述训练数据中的句子？这个问题关乎到算法是否真正理解了物体及其交互这个核心问题。

科学家们给出了一个令人振奋的答案： Yes。

如今的图像语言描述系统确实已经发展出自主产生全新的句子能力，例如下图粗体的描述为不在数据库中的标注句子：

A man throwing a frisbee in a park. **A man holding a frisbee in his hand.** **A man standing in the grass with a frisbee.**
A close up of a sandwich on a plate. A close up of a plate of food with french fries. A white plate topped with a cut in half sandwich.
A display case filled with lots of donuts. **A display case filled with lots of cakes.** **A bakery display case filled with lots of donuts.**

图6. 生成的语言描述 (粗体的句子为不在训练数据中的全新句子)

其生成全新描述过程可以用下图进行很好的阐述：

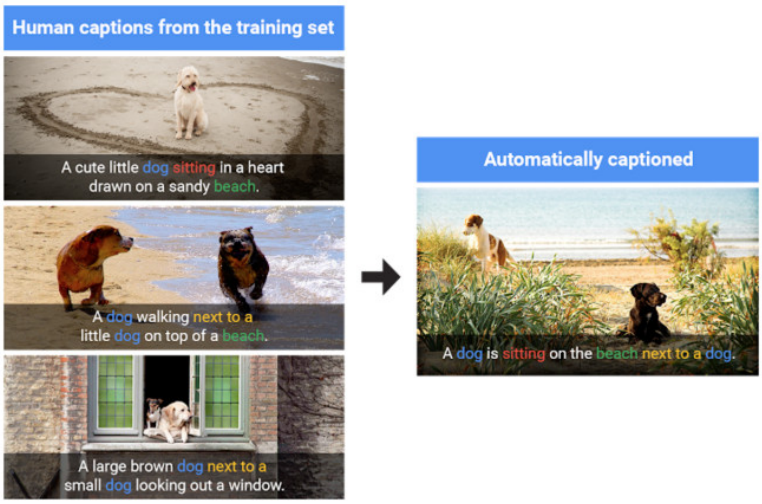


图7. 模型从训练数据提供的概念中生成全新的描述

此领域的突破同时也得益于如今标注数据的增长，作者们通过实验证明：越多的图像描述样本，越是可以极大地增强如今已经效果不错的图像描述算法。

> 图像描述数据库对比于如今最大的标注数据库ImageNet还差几个数量级，所以我们有理由期待，一旦具备更多的标注数据，图像描述算法在如今的基础上还可以得到大幅度的提升。

这也是Goolge的研究者开源其系统的原因，其希望让更多人参与到此领域的研究中。

视觉信息约占人类从外界获取信息的⅔，所以机器视觉的重要性自然不言而喻；语言作为人之所以为人的标志，因而自然语言处理被称为人工智能皇冠上最亮的明珠。Image caption作为一个连接此两个领域的问题，其突破性的进展更深层次的意义在于表明人工智能的全面进步。

俗话说「一图胜千言」，长久以来计算机视觉领域比较关注一些基本的视觉任务，如分类，检测，分割等。近期在image caption领域的突破使得计算机可以用自然语言去描述图片，实现真正的“千言说图”。也许我们真的在进入一个崭新的智能时代，而当强人工智能真正出现之时，一切都将不可逆地去往技术奇点。

图8. 取于《HER》影片末尾。

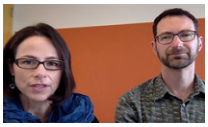
我想，如果有一天，“她”真的到来，看到此情此景。

“她”大概会说：天台上一位穿着红衣的女生依偎着白衣男生，眼前是鳞次栉比的上海夜景，他们好像都有点心事重重。

相关文章

Google 机器视觉 自然语言处理 TensorFlow ImageNet CVPR 深度学习

被 Google 捅了一刀之后，亚马逊是这样怼回去的



Google AI 实力打脸：你真的懂机器学习嘛？

为什么 Google 与亚马逊撕了起来？

YouTube突然从Echo音箱下架！Google和亚马逊关

文章点评：

我有话要说.....

同步到新浪微博

提交

热门关键字

热门标签 微信小程序平台 微信小程序在哪 CES 2017 CES 2016年最值得购买的智能硬件 2016 互联网 小程序 微信朋友圈 抢票软件 智能手机 智能家居 智能手环 智能机器人 智能电视 360智能硬件 智能摄像机 智能硬件产品 智能硬件发展 智能硬件创业 黑客 白帽子 大数据 云计算 新能源汽车 无人驾驶 无人机 大疆 小米无人机 特斯拉 VR游戏 VR电影 VR视频 VR眼镜 VR购物 AR 直播 扫地机器人 医疗机器人 工业机器人 类人机器人 聊天机器人 微信机器人 微信小程序 移动支付 支付宝 P2P 区块链 比特币 风控 高盛 人脸识别 指纹识别 黑科技 谷歌地图 谷歌 IBM 微软 乐视 百度 三星s8 腾讯 三星Note8 小米MIX 小米Note 华为 小米 阿里巴巴 苹果 MacBook Pro iPhone Facebook GAIR IROS 双创周 云栖大会 智能硬件公司 智能硬件 QQ红包 支付宝红包 敬业福 支付宝敬业福 支付宝集五福 Waymo 虚拟现实 深度学习 人工智能 中国银联 蚂蚁金服 WRC CNCC 贾跃亭 令计划 小猪短租 obd 创客教育 touchjet pond 百度识图 4k屏手机 魅蓝note2拆机 蜗牛移动 俄罗斯ostagram 中芯 facebook oculus 小米 360 摄像头 微软手机 神州专车 更多

0