“看图说话机器人”项目总结

组员:余杨、冯宇琨、王磊
2018年11月4日

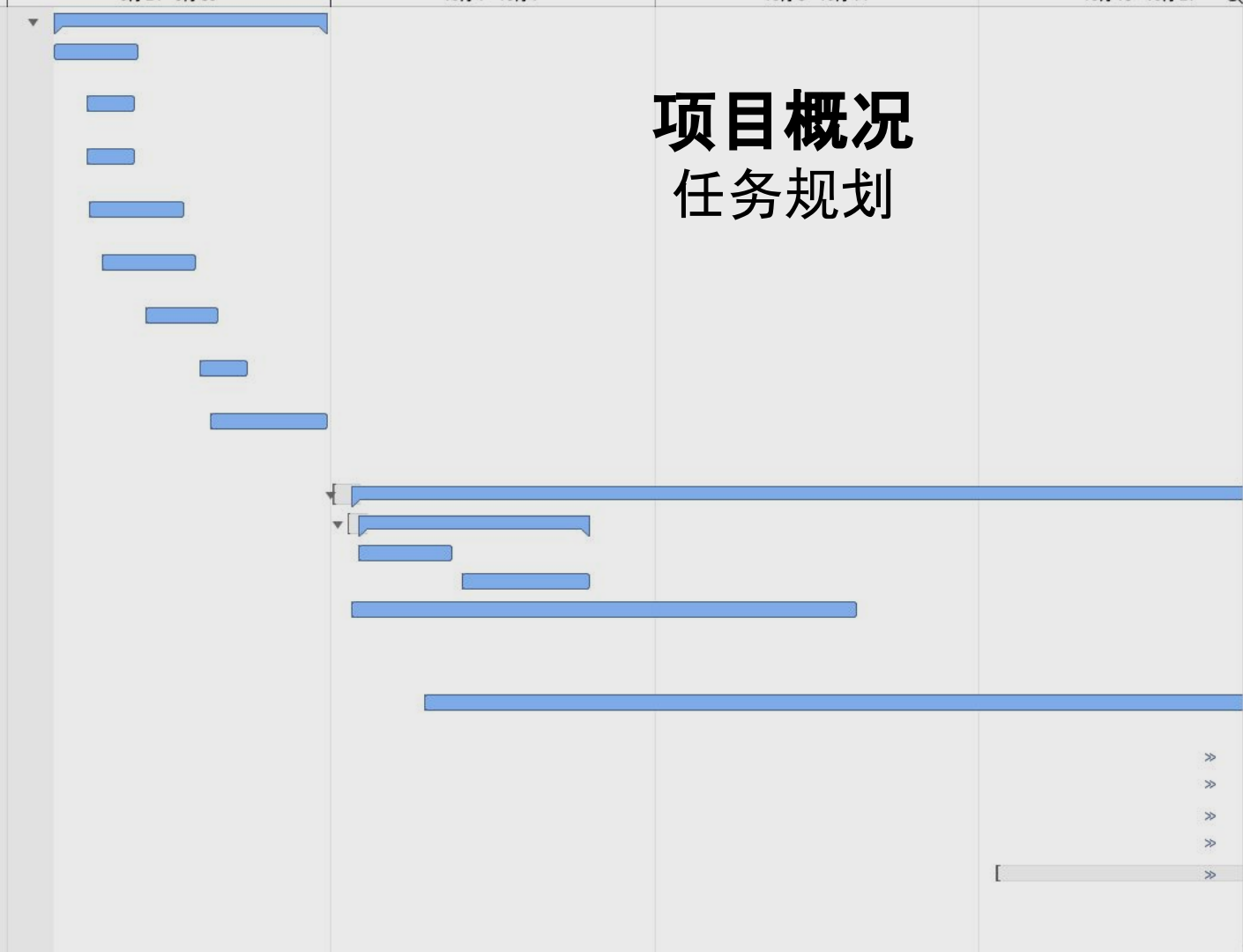
目录

- 一、项目概况
  - 任务规划
  - 项目背景
  - 数据集背景
  - 任务评价指标
  - 经典论文模型
  - 模型架构规划
  - 目标达成
- 二、困难与挑战
- 三、关键举措
  - encoder & Attention Mechanism
  - dense caption
- 四、思考与总结
- 五、未来展望
- 六、附录列表

项目概况

任务规划

- 1) 筹备期
  - 1.1) COCO数据准备
    数据下载和数据处理
  - 1.2) 数据集了解
    清楚数据和数据格式
  - 1.3) 评价指标
    了解几种评价机制
  - 1.4) 论文阅读
    方向选型
  - 1.5) 模型规划设计
    基于im2txt框架进行
  - 1.6) 系统架构规划
    采用web server&app
  - 1.7) 开发环境准备
    用git进行版本控制和代码托管
  - 1.8) 运行im2txt
    作为初始版本
    基于此基础先行搭建系统
- 2) 开发&训练
  - 2.1) 搭建 web server 和web前端
    - 2.1.1) web server
    - 2.1.2) web 前端
  - 2.2) 模型改进
    attention机制
    encoder改进
    densecap进阶
  - 2.3) 训练模型
    调参炼丹
  - 2.4) 评估模型
- 3) 测试&发布
  - 3.1) 测试系统
- 4) 项目总结
  - 4.1) 文档规整



项目概况

任务背景

a woman standing in front of a wall holding a cell phone .

图像描述生成（Image Caption）是一个融合计算机视觉(cv)、自然语言处理(nlp)的综合问题，简单来说就是翻译一副图片为一段描述文字。该任务对于人类来说非常容易，但是对于机器却非常具有挑战性，它不仅需要利用模型去理解图片的内容并且还需要用自然语言去表达它们之间的关系。除此之外，模型还需要能够抓住图像的语义信息，并且生成人类可读的句子。2016年的IEEE国际计算机视觉与模式识别会议（即IEEE Conference on Computer Vision and Pattern Recognition，缩写为**CVPR**）上专门有一个小型会议（session）的主题就是Image Caption.

项目概况

任务背景

传统做法

图像描述生成可以认为是一种动态的目标检测，由全局信息生成图像摘要。早先的做法例如《Baby Talk》,《Every picture tell a story》等都是利用图像处理的一些算子提取出图像的特征，经过SVM分类等等得到图像中可能存在的目标object。根据提取出的object以及它们的属性利用CRF或者是一些认为制定的规则来恢复成对图像的描述。这种做法非常依赖(图像特征的提取)和 (生成句子时所需要的规则)。

观点：效率不高，且耦合高，端 对端实现困难。

项目概况

数据集背景

数据集名称	数据量 (train-val-test)	训练用时	用途
PASCAL	1k	One hour	Debugging and testing
Flickr 8k	6k-1k-1k	Few hours	
Flickr 30k	28k-1k-1k	Less than a day	Training
MS COCO	82k-40k-40k	A couple of days	
SBU	1M	Very long	
Visual Genome	100K	Weeks	

项目概况

数据集背景

MS COCO的全称是Common Objects in COntext，是微软团队提供的一个可以用来进行图像识别的数据集。MS COCO数据集以scene understanding为目标，主要从复杂的日常场景中截取，图像中的目标通过精确的segmentation进行位置的标定。图像包括91类目标，328,000影像和2,500,000个label。COCO通过在Flickr上搜索81个对象类别和各种场景类型来收集图像，其使用了亚马逊的Mechanical Turk（AMT），COCO数据集现在有3种标注类型：object instances（目标实例）、object keypoints（目标关键点）、image captions（图像标注），这些信息均使用JSON文件格式存储。

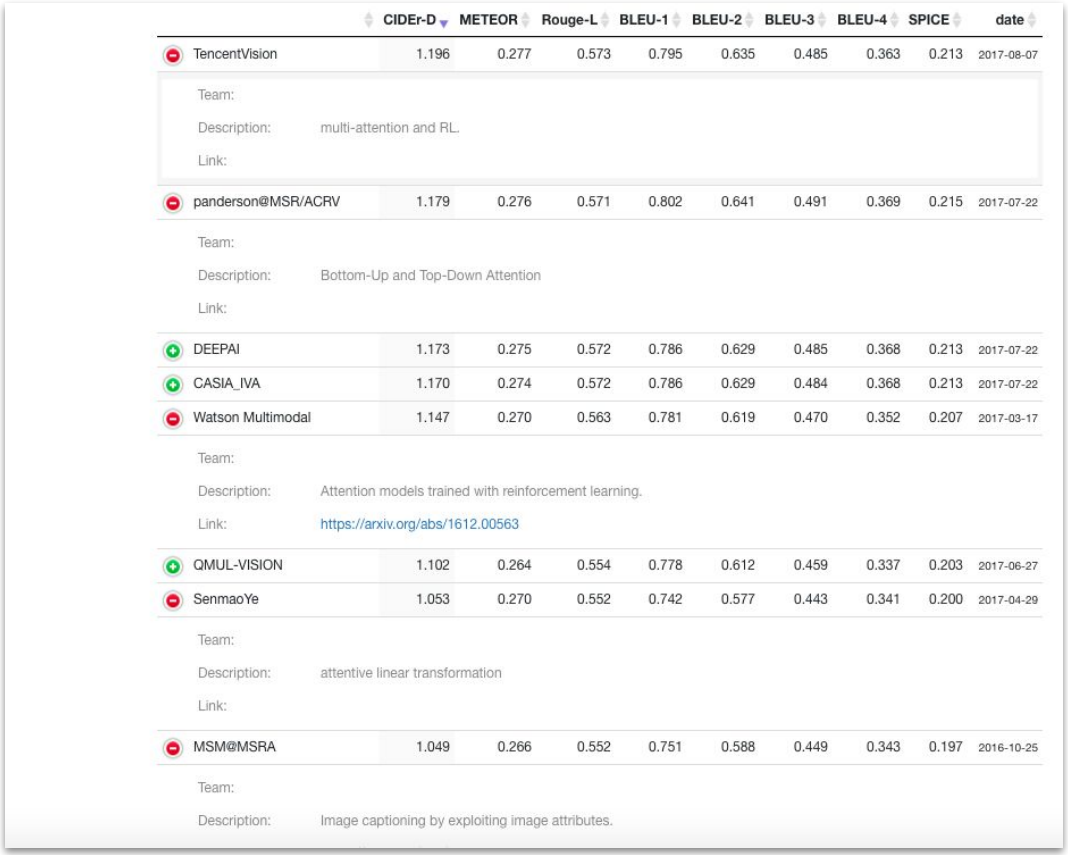
Visual Genome数据集由斯坦福李飞飞团队设计提出：它不但包括了图像本身，更包括了图像内对象之间的关系等众多数据（包括objects、attributes、relationship等）。Visual Genome数据集一共包括了 108K张图片，平均每张图片内包含了 35个object，和 26个attributes，以及 21对object之间的relationship pair。本数据集的图片取自MS COCO和YFCC100M,作者们还将其中所有的object、attributes、relationships和在region descriptions与question answer pairs中的名词短语都映射到了WordNet synset上。从而打通了从CV到Knowledge乃至NLP之间的连接通道。对比COCO数据集，本数据集具有更丰富的图像标注信息。

在我们的任务中，MS COCO数据集属于基础阶段适用的，在dense captioning中则由Visual Genome数据集驱动我们的训练学习。

项目概况

评价指标

CIDEr-D	CIDEr: Consensus-based Image Description Evaluation
METEOR	Meteor Universal: Language Specific Translation Evaluation for Any Target Language
Rouge-L	ROUGE: A Package for Automatic Evaluation of Summaries
BLEU	BLEU: a Method for Automatic Evaluation of Machine Translation
SPICE	SPICE: Semantic Propositional Image Caption Evaluation



	CIDEr-D	METEOR	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SPICE	date
TencentVision	1.196	0.277	0.573	0.795	0.635	0.485	0.363	0.213	2017-08-07
Team:									
Description:	multi-attention and RL.								
Link:									
panderson@MSR/ACRV	1.179	0.276	0.571	0.802	0.641	0.491	0.369	0.215	2017-07-22
Team:									
Description:	Bottom-Up and Top-Down Attention								
Link:									
DEEPAI	1.173	0.275	0.572	0.786	0.629	0.485	0.368	0.213	2017-07-22
CASIA_IVA	1.170	0.274	0.572	0.786	0.629	0.484	0.368	0.213	2017-07-22
Watson Multimodal	1.147	0.270	0.563	0.781	0.619	0.470	0.352	0.207	2017-03-17
Team:									
Description:	Attention models trained with reinforcement learning.								
Link:	https://arxiv.org/abs/1612.00563								
QMUL-VISION	1.102	0.264	0.554	0.778	0.612	0.459	0.337	0.203	2017-06-27
SenmaoYe	1.053	0.270	0.552	0.742	0.577	0.443	0.341	0.200	2017-04-29
Team:									
Description:	attentive linear transformation								
Link:									
MSM@MSRA	1.049	0.266	0.552	0.751	0.588	0.449	0.343	0.197	2016-10-25
Team:									
Description:	Image captioning by exploiting image attributes.								

项目概况

评价指标

名称	描述	补充
CIDEr-D	CIDEr 是专门设计出来用于**图像标注**问题的。这个指标将每个句子都看作“文档”，将其表示成 Term Frequency Inverse Document Frequency（tf-idf）向量的形式，通过对每个n元组进行(TF-IDF) 权重计算，计算参考 caption 与模型生成的 caption 的余弦相似度，来衡量图像标注的一致性的。	CIDEr-D 是修改版本，为的是 对于 gaming 问题更加鲁棒。什么是 Gaming 问题：就是一个句子经过人工判断得分很低，但是在自动计算标准中却得分很高的情况。为了避免这种情况，CIDEr-D 增加了截断（clipping）和基于长度的高斯惩罚。
METEOR	METEOR 是基于BLEU进行了一些改进，其目的是解决一些 BLEU 标准中固有的缺陷。使用 WordNet 计算特定的序列匹配，同义词，词根和词缀，释义之间的匹配关系，改善了BLEU的效果，使其跟人工判别共更强的相关性。	METEOR 也包括其他指标没有发现一些其他功能，如同义词匹配等
Rouge-L	ROUGE 是出于召回率来计算，所以是**自动摘要**任务的评价标准，	Rouge-L基于最长公共子序列（LCS）的度量方法。LCS 是同时出现在两个句子中且顺序相同的一组词。将两个句子的 LCS 长度记为 $l(c_i, s_{ij})$，通过计算 F-measure (F1-score)来生成。
BLEU	BLEU这个计算标准在图像标注结果评价中使用是很广泛的，但是它的设计初衷并不是针对图像标注问题，而是针对**机器翻译**问题，它是用于分析待评价的翻译语句和参考翻译语句之间n元组的相关性的。直白地来说，它的核心思想就是：机器翻译语句与人类的专业翻译语句越接近就越好。	引入一个简洁性惩罚呢？这是因为BLEU倾向于更短的句子，这样精度分数就会很高。为了解决这个问题，使用了乘以一个简洁性惩罚来防止很短的句子获得很高的分数。
SPICE	SPICE 也是专门设计出来用于**图像标注**问题的。全称是 Semantic Propositional Image Caption Evaluation。前面四个方法都是基于 n-gram 计算的，所以 SPICE 设计出来解决这个问题。PICE 使用基于图的语义表示来编码 caption 中的 objects, attributes 和 relationships。	它先将待评价 caption 和参考 captions 用 Probabilistic Context-Free Grammar (PCFG) dependency parser parse 成 syntactic dependencies trees然后用基于规则的方法把 dependency tree 映射成 scene graphs。最后计算待评价的 caption 中 objects, attributes 和 relationships 的 F-score。

项目概况

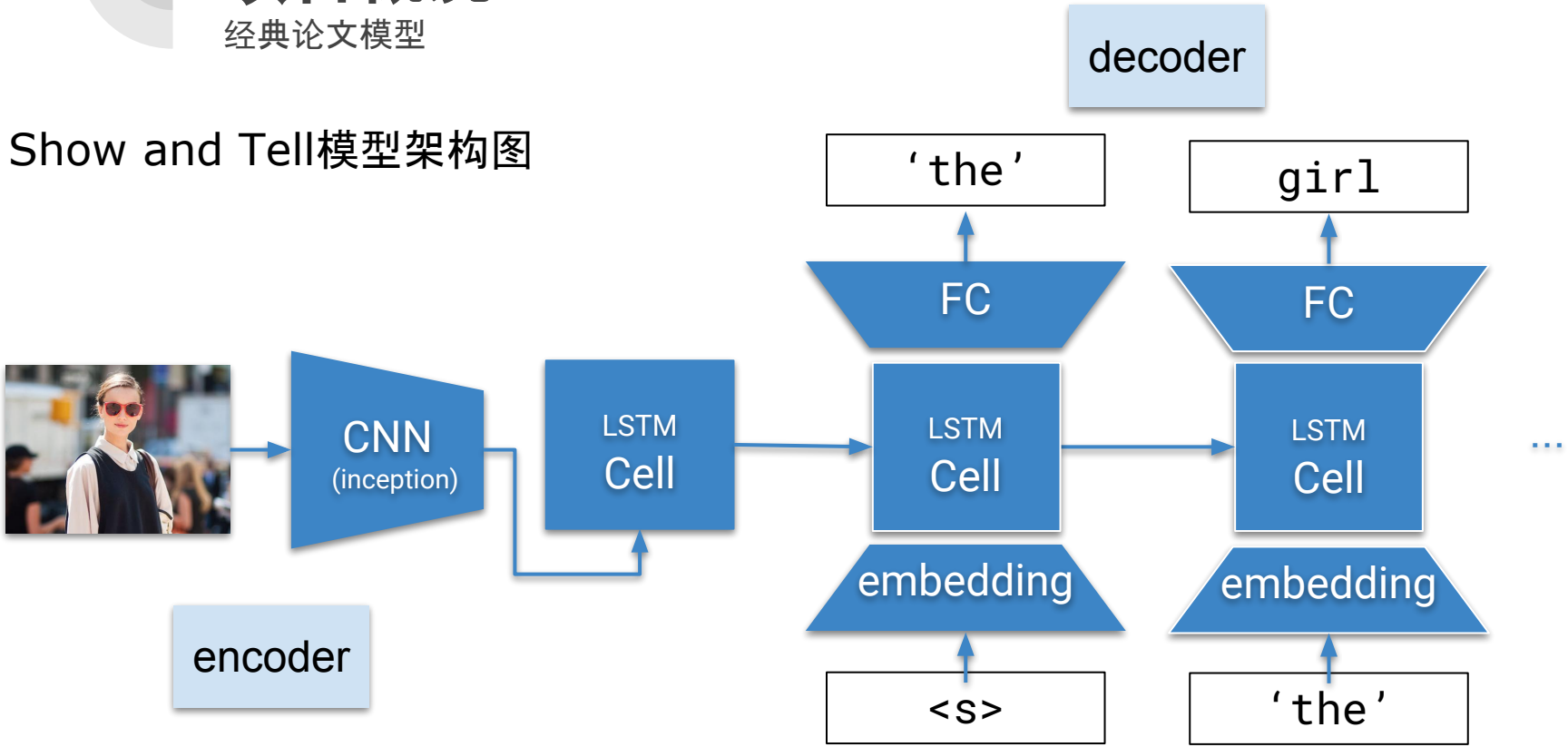
经典论文

- Show and Tell: A Neural Image Caption Generator [arXiv:1411.4555]
- Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge [arXiv:1609.06647]
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [arXiv:1502.03044]
- What value do explicit high level concepts have in vision to language problems? [arXiv:1506.01144]
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering [arXiv:1707.07998]
- DenseCap: Fully Convolutional Localization Networks for Dense Captioning [arXiv:1511.07571]
- Dense Captioning with Joint Inference and Visual Context [arXiv:1611.06949]
- A Hierarchical Approach for Generating Descriptive Image Paragraphs [arXiv:1611.06607]

项目概况

经典论文模型

Show and Tell模型架构图



项目概况

模型架构规划

1. 对im2txt-encoder结构采用图片分类任务中效果更好的Inception_V4、Inception_resnet_V2;

2. 采用目标检测任务中的Faster R-CNN网络进行图像信息提取，把RPN的结果作为decoder结构的输入；

   Faster R-CNN可以对图像进行多标签分类，提取多区域特征，提取颜色、材质等属性特征，结合这些特征，可以训练出一个semantic model，比较第一种方案，这样获取到的“图像的高层次的抽象信息”，对于Image caption任务或是VQA 任务来说更有价值；

3. 在im2txt网络结构中引入attention机制：soft attention；

4. 由im2txt进阶到dense caption模型：object decetion & image caption;

5. 除了只进行encoder和decoder的优化，还需要考虑图像的语义信息和语言的语义信息的联合运用来进行上下文推理；

项目概况

模型架构规划

系统规划：

1. 前后端分离、模型业务分离，有利于系统各部分解耦，便于扩展；

2. 前端服务：基于VUE搭建，提供数据展示，数据输入，用户交互功能：列表、gallery、收藏

3. 业务服务：基于Tornado搭建，为前端业务提供API接口：用户登录登出、数据返回、输入处理、数据清除，以及与ML模型服务通信；

4. ML模型服务：基于Tornado搭建，提供image captioning 模型服务，对输入的图片数据返回对应的captions;

5. 服务间通信格式采用json数据格式；

项目概况

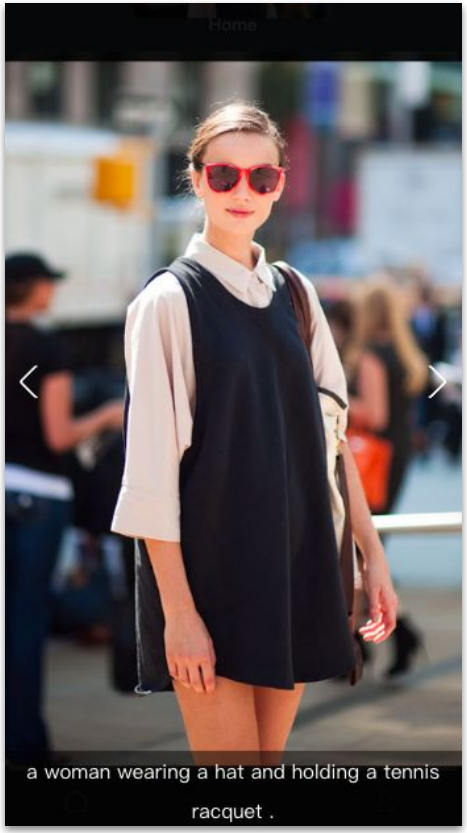
目标达成

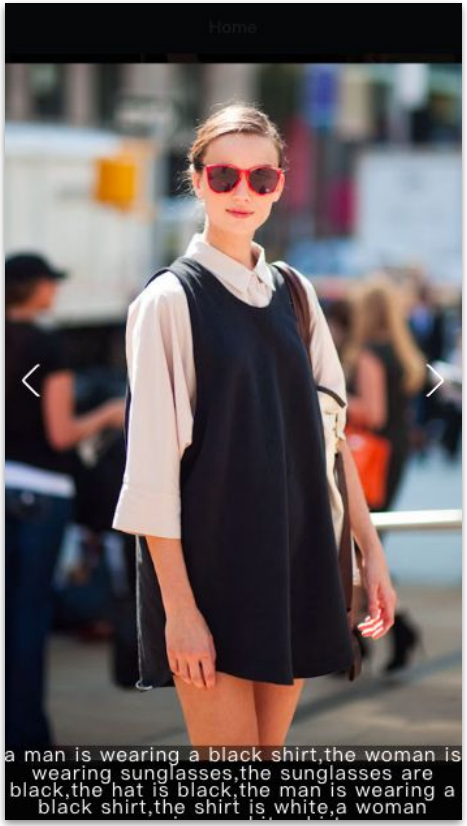
在mscoco 和 visual genome 数据集下我们完成了对im2txt基本模型的各种改进，我们尝试的方案有：

inception_v3 & inception_v4 attention model（右图1）

densecap model（右图2）

在线DEMO地址：http://bot.yao-it.com/





困难与挑战

对于Image Caption任务，在基础im2txt架构中我们只是把数据交给Encoder-Decoder结构，模型在训练中形成一种从图像到文本的映射，非常“黑箱”：

- 我们很难解释模型对于图像哪些像素可以表示到文本中，图像有用和无用信息如何进行筛选；
- 现有的模型对于图片识别效果较局限，容易识别错误，对于丰富场景的图片识别效果比较无力，MS COCO数据集的数据较局限，字幕标签则更限制了模型的表达；
- 图像特征能否进行更高层次的使用，将图像特征用高等级的语义概念表达后再输入RNN是否会有更好的模型效果；
- im2txt只能进行简单的语句表达，在一些场景下只能进行部分信息的描述；
- 由于结合了CV和NLP任务，对于我们构建模型所需要的能力以及训练模型对于算力的要求也较高
- Image Caption是一个正在快速发展的领域，新的洞见不断被推出，并且和其他领域的进展相关性较高，如机器翻译，语音识别，目标检测；

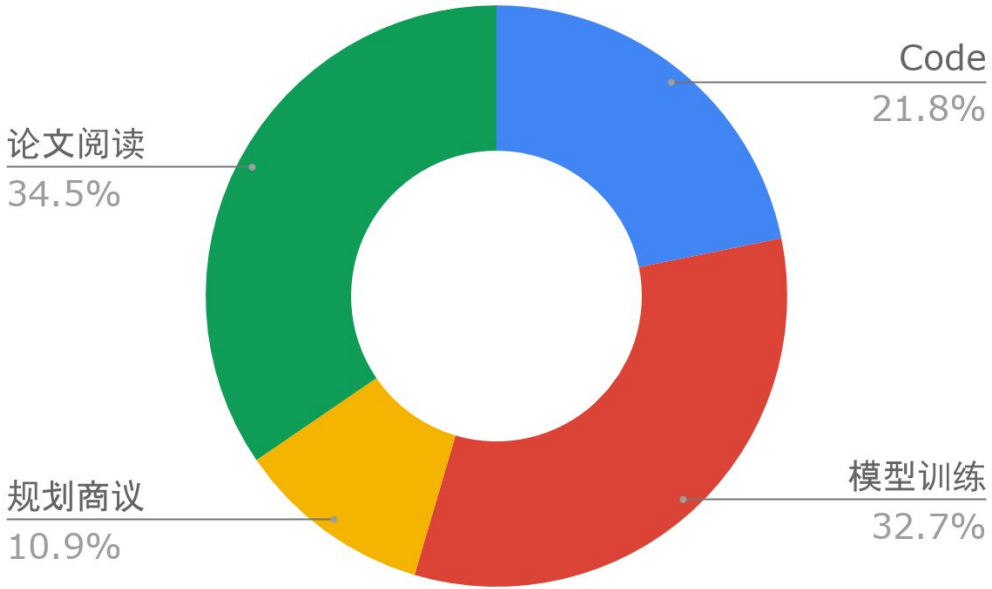
关键措施

- 使用图片分类任务性能更佳的基础网络提升模型的图片识别能力；
- 引入soft attention机制：解决了传统encoder-decoder架构中对于长输入序列的支持，模型对信息进行有选择的“挑选”，而不再依赖固定长度的信息，并且该机制为我们提供了可视化图像关注区域的方法。
- 使用Visual Genome数据集来训练模型,尤其对于基础分类网络我们使用在其他数据集进行了预训练的模型到我们的模型，改善了模型的图片识别精度；
- 我们在模型中采用Faster R-CNN可以对图像进行多标签分类，包含物体属性等分类标签，提取出多区域特征向量，对特征向量的处理方式也有多种：对于该特征向量与标签字幕向量进行加权求和后进行softmax得到具有一定语义表达的输出，使得图片信息能表达高等级的语义概念；我们在densecap中，也对该特征向量最大池化后直接通过FC连接层展开后作为LSTM的输入；
- 通过对densecap的实现，对图片的一句话描述变为了多条语句描述的方式，我们也在采取对模型的提升使生成的多条语句变为一个顺畅的段落描述；
- 使用Google colab 和 Intel devcloud 免费的强力运算平台；
- 追踪论文可以跟随大神如斯坦福李飞飞团队、MSRA他们的脚步；

思考与总结

我们在项目周期中大部分时间用于在论文阅读和模型训练上，其中论文阅读作为我们各项工作开始的前提，由于Image Caption任务近几年热度开始上升并快速发展，很多论文都在不断推陈出新，从之前提出的方案迭代改进式发展的；我们在开始这项工作时，是从比较早期的论文开始的，随着工作逐渐深入，任务的难度和挑战也开始逐渐增加，尤其当意识到我们的基础方案所具有的识别表达能力表现很一般后，在项目中后期我们也开始追踪最新的论文，对其进行实现也会发现，当我们从简单的表层特征逐渐开始对图像和文本的高层次语义信息进行利用，并把基于人类视觉的注意力机制结合到模型中使模型的性能得到了提高。

我们在Image Caption任务的工作中，整体的时间的分配如图：



思考与总结

我们的工作最后虽然可以生成富含更多信息的图片描述，并且结合注意力机制可以得到不错的效果，但对于生成的句子我们还可以结合上下文推理来生成更完整的语句信息，这也是2017年李飞飞团队论文提出的一项改进。在我们目前的工作来看，这项任务也只是开始，而我们的工作也会持续迭代下去。如可以采用Mask X-RCNN作为我们的图像特征区域提取，采用近期性能强劲的谷歌BERT模型作为我们文本训练，这两者都可以进行半监督训练；Image Caption任务具有较为重要的意义在于，它使得机器具有感知场景的能力，这项工作在未来也许会更紧密地把文本，图像，语音结合起来，进行上下文推理，深度挖掘语义特征。

对于Image Caption任务来说，现在很多研究工作也逐渐开放发展，比如有的工作对生成的语言加入一些情绪语义，生成诙谐有趣的句子；也有基于个人的语言数据库来生成富有个人感情色彩的句子；也有工作是对一副图像生成完整流程的故事情节。

也有比Image Caption任务更难一些的Video Caption ，两者在模型架构上有一定的共通，不过视频是一个序列，对视频的特征提取会有难度，而且视频中还会有动作检测任务进行结合。

附录

1. 第一周项目汇报.pdf
2. 第二、三周项目汇报.pdf
3. IM2TXT模型细节.pdf
4. encoder&attention实施.docx
5. densecap实施.pdf
6. 系统说明.pdf