2020武汉大学本科毕业答辩

图像-文本互译系统的原理与实现

指导教师：黄浩、庄越挺　答辩学生：雷伯涵

目录

CONTENTS

第一部分

研究背景

- 介绍需求来源
- 简述相关工作

研究目的与内容

本文的目的是**提出并设计和实现了一套基于自然语言文本与图像互译系统，创 新型地提出将两种用途近似、目标人群耦合的技术合并为一个完整可用的系统，方 便用户群体便捷地使用到这两项功能。** 具体的，分别以LSTM模型和GAN模型实现了两项功能。对图像翻译文本的功能，使用 GCN提 取图片的特征向量，并加入 LSTM 网络中训练注意力机制函数的转移方法，从而根据图片的特征向量生成对应的描述性自然语言文本 ； 对文本翻译图像的功能，对输入文本的处理，使用 NLP 中的分词和 语义角色分析技术生成场景图，通过 GCN、两种回归网络和 CRN，实现分三步的 场景图像生成模型，并设计辨别器与之对抗，通过如此新建的 GAN 网络模型训练 出对一般人来说更加真实的复杂场景图像，实现文本向图像进行翻译的功能。

在COCO数据集及VG数据集上的实验对比验证了本设计的有效性和运行效率。

研究目的与内容

客观需求

（1）看不见的人：读懂视野，改善生活质量

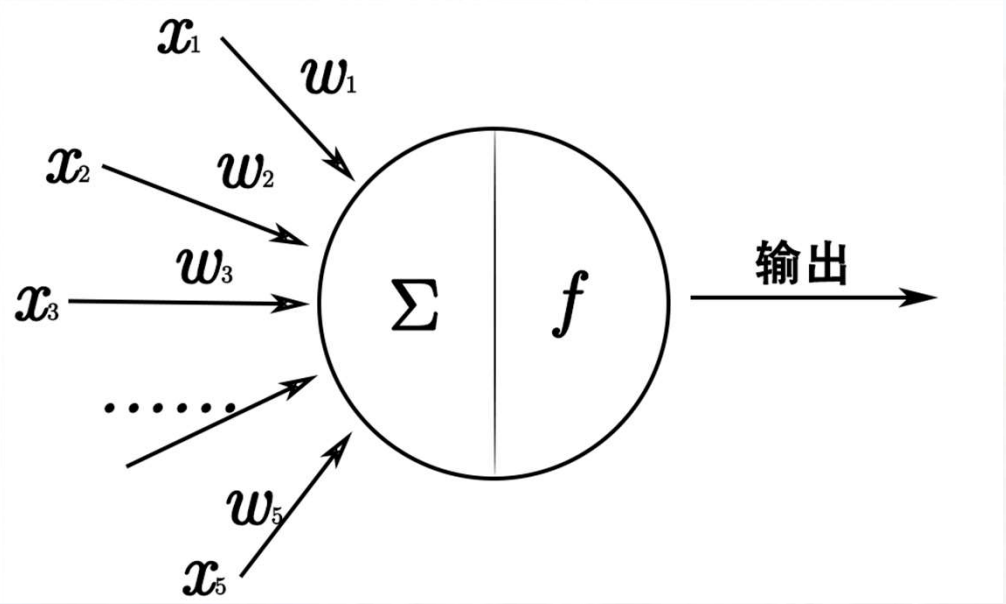
（2）普通人群：直观表现描述性语言的含义

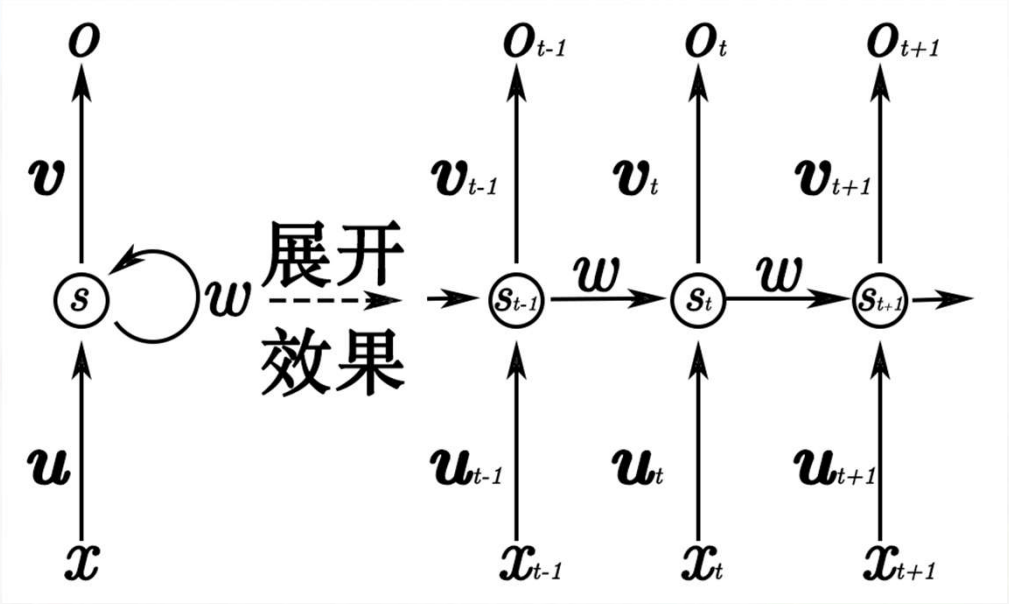
研究目标

（1）**图像生成文本：**生成文本通顺可理解

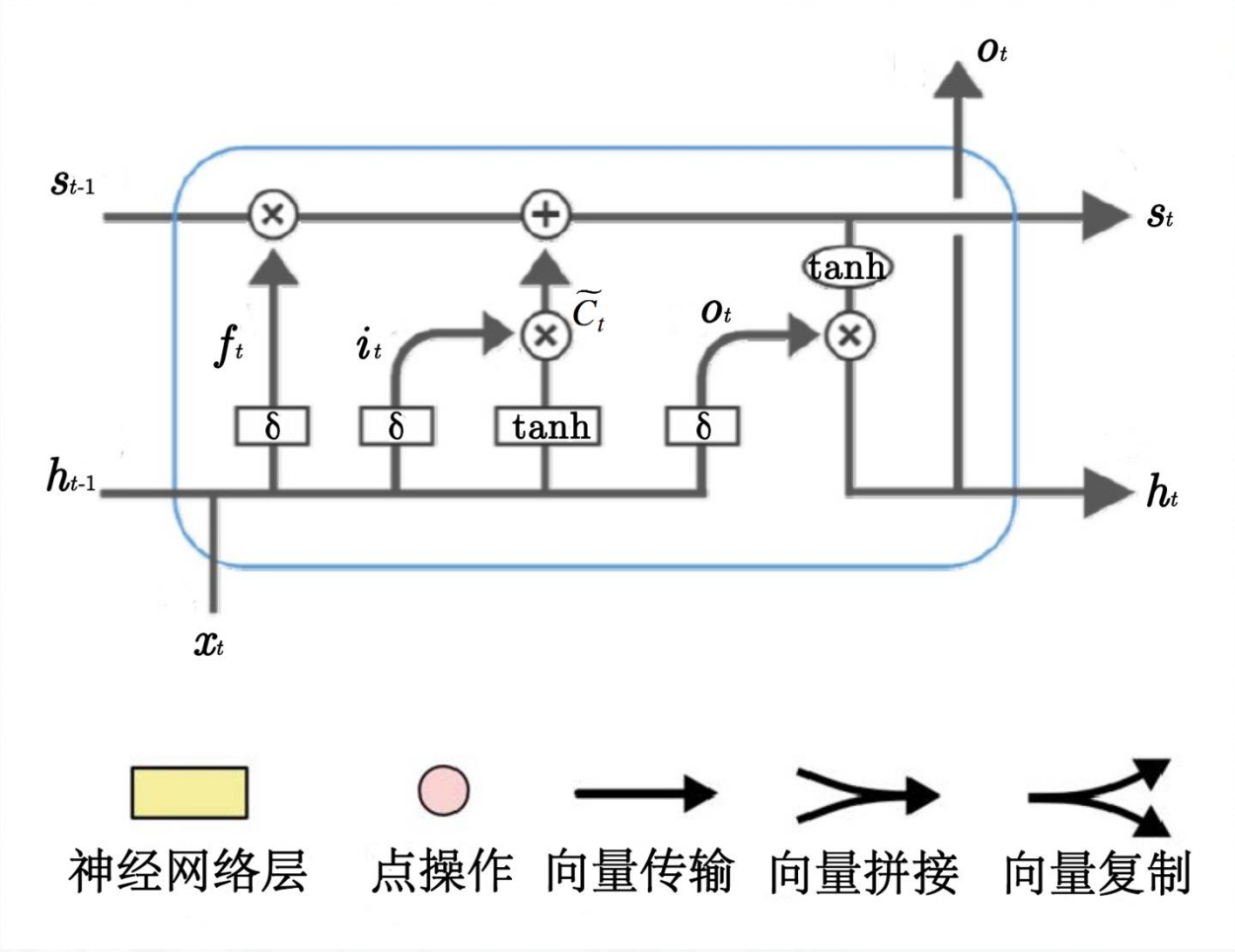
（2）**文本生成图像：**生成图像质量可辨认，布局合理

背景知识：神经网络与循环神经网络

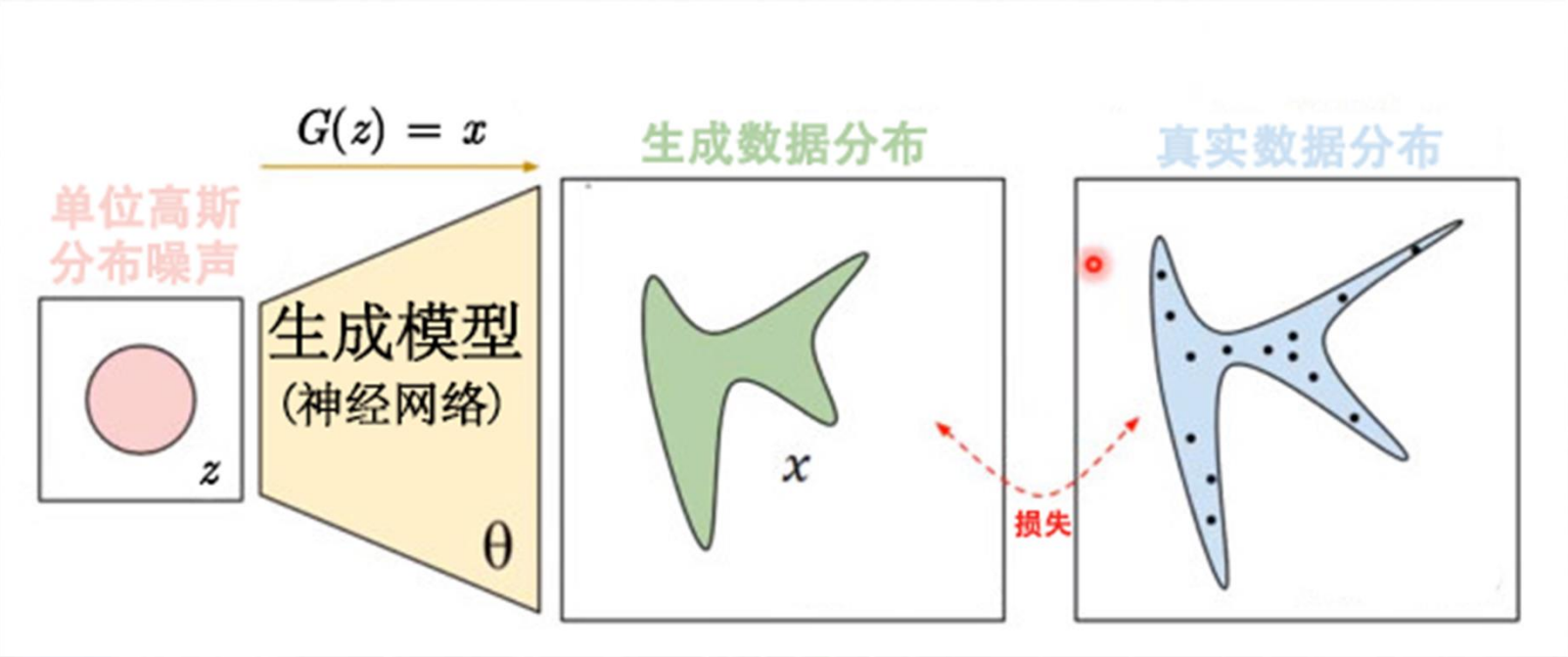




背景知识：长短期记忆


o_t
s_{t-1}
s_t
tanh
$\widetilde{C}_t$
f_t
i_t
o_t
δ
δ
tanh
δ
h_{t-1}
h_t
x_t
神经网络层
点操作
向量传输
向量拼接
向量复制

背景知识：GAN模型原理



$$\min_G \max_D V_{G,D} = \mathbb{E}_{x\sim P_{data}(x)}[\lg D(x)] + \mathbb{E}_{z\sim P_G(z)}[\lg(1 - D(G(z)))]$$

相关工作：CGAN

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim P_{data(x)}}[\lg D(x_i \mid y)] + \mathbb{E}_{x\sim P_G(z)} \lg[(1-D(G(z_i \mid y)))]$$

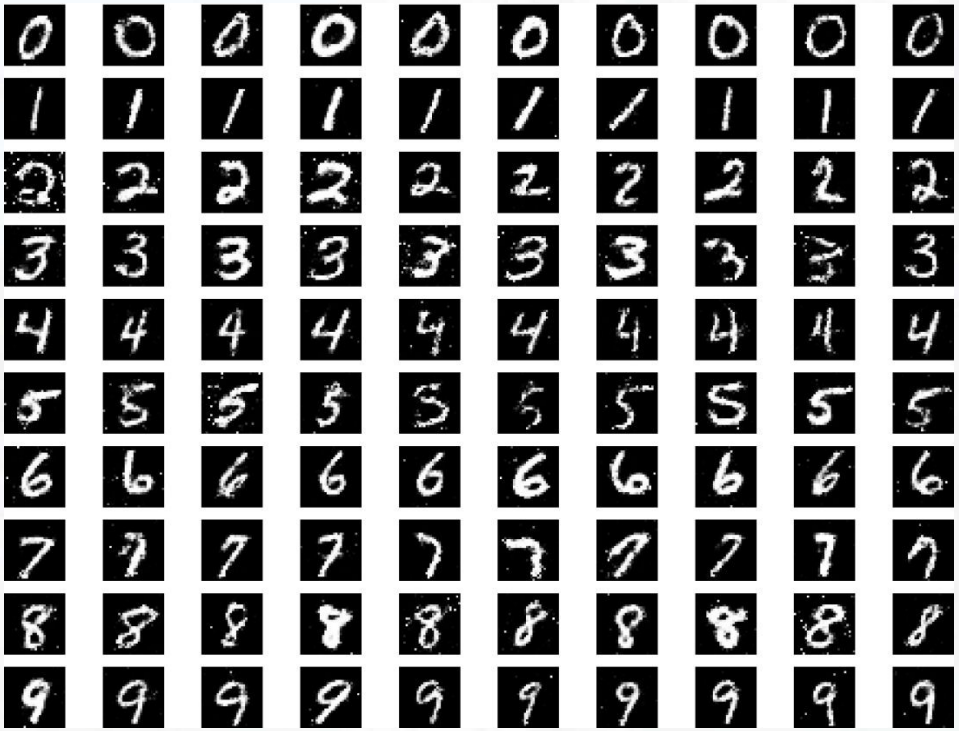
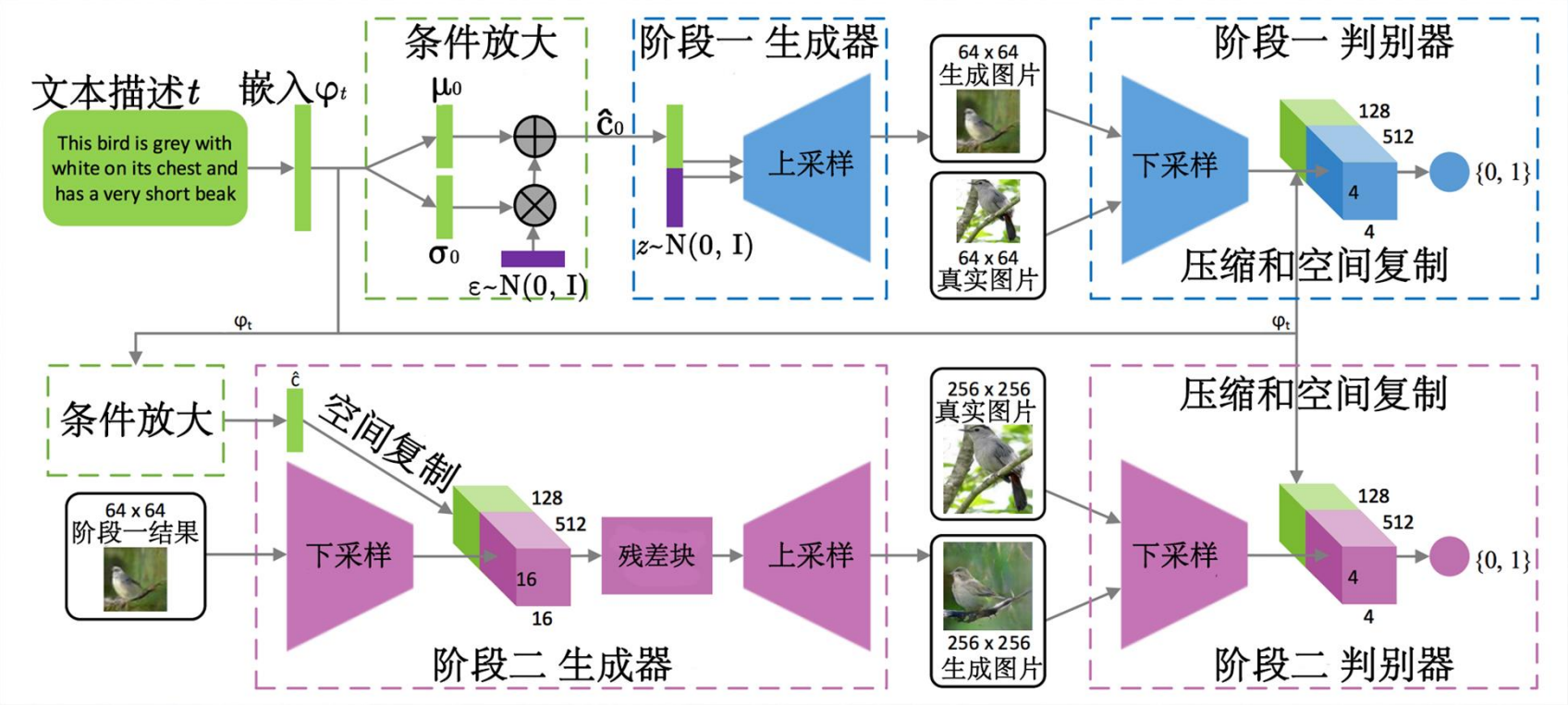


Figure 7: Generated samples

相关工作：StackGAN



$$\mathcal{L}_{D_0} = \mathbb{E}_{(I_0,t)\sim P_{data}}[\lg D_0(I_0, \varphi_t)] + \mathbb{E}_{z\sim P_z, t\sim P_{data}}[\lg(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))],$$

$$\mathcal{L}_{G_0} = \mathbb{E}_{z\sim P_z, t\sim P_{data}}[\lg(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t), \Sigma_0(\varphi_t)) \,\|\, \mathcal{N}(0, I))$$

$$\mathcal{L}_{D} = \mathbb{E}_{(I,t)\sim P_{data}}[\lg D(I, \varphi_t)] + \mathbb{E}_{s_0\sim P_{G_0}, t\sim P_{data}}[\lg(1 - D(G(s_0, \hat{c}), \varphi_t))],$$

$$\mathcal{L}_{G} = \mathbb{E}_{s_0\sim P_{G_0}, t\sim P_{data}}[\lg(1 - D(G(s_0, \hat{c}), \varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t)) \,\|\, \mathcal{N}(0, I))$$

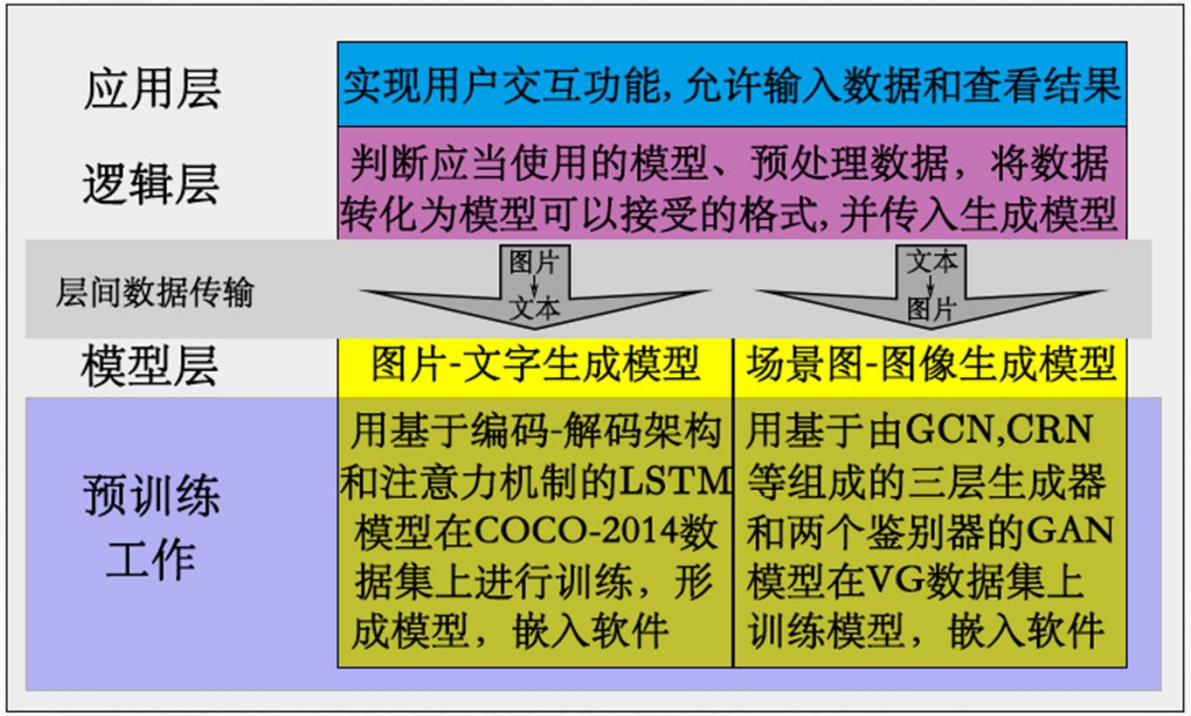
第二部分

设计与实现

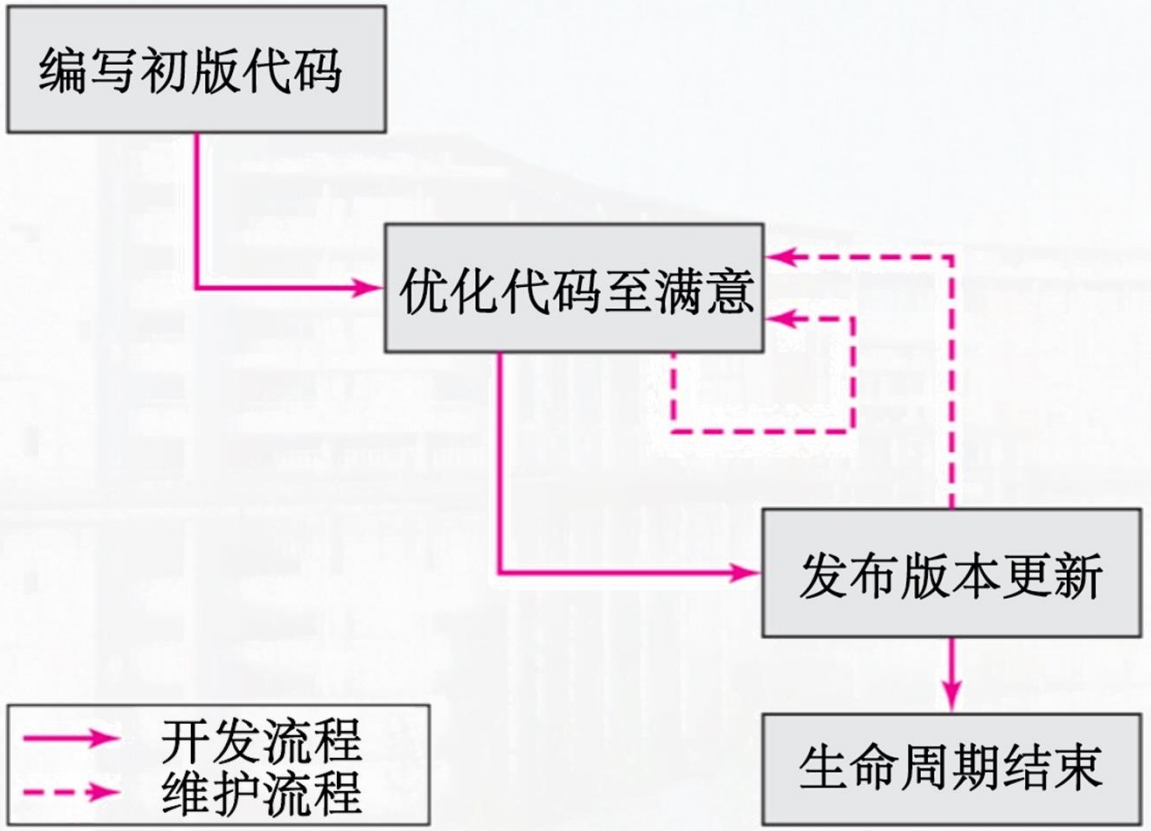
贰

- 介绍设计框架
- 介绍实现主要功能的两种模型
- 介绍前端用户交互界面设计

设计框架


应用层
实现用户交互功能，允许输入数据和查看结果
逻辑层
判断应当使用的模型、预处理数据，将数据转化为模型可以接受的格式，并传入生成模型
层间数据传输
图片
文本
文本
图片
模型层
图片-文字生成模型
场景图-图像生成模型
预训练工作
用基于编码-解码架构和注意力机制的LSTM模型在COCO-2014数据集上进行训练，形成模型，嵌入软件
用基于由GCN,CRN等组成的三层生成器和两个鉴别器的GAN模型在VG数据集上训练模型，嵌入软件

设计流程



功能实现：图像生成文本

1.在模型中输入图片，作为输入信息;

2. 由卷积神经网络提取图片信息，形成图片特征信息(即后文编码步骤);

3. 由注意力机制(attention)对所提取的图片特征信息进行处理，加强或抑制部分区域，作为后续输入 LSTM 的输入信息——在不同时刻，注意力信号会受到 上一次 LSTM 的输出信息的影响，即注意力信号作为 LSTM 神经元细胞的状态， 受到输出词语的影响而改变(这也是后文的解码部分);

4. LSTM 最终输出文本，形成最后的结果。

功能实现：图像生成文本

文本构建词典编码为向量

$$y = \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_C, \mathbf{y}_i \in \mathbb{R}^K$$

图片用CNN提取特征向量

$$a = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D$$

注意力机制函数 ϕ 来计算 t 时刻的背景向量 $\hat{z}_t$

$$\hat{z}_t = \phi(\mathbf{a}_i, \alpha_i)$$

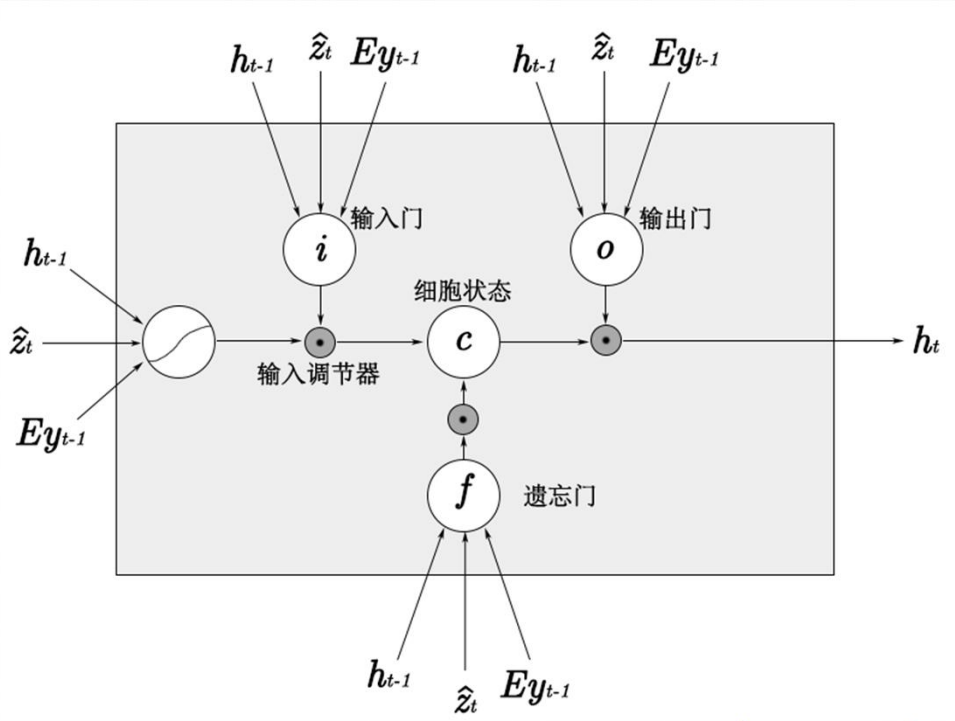
$$e_{t,i} = f_{att}(\mathbf{a}_i, \mathbf{h}_t)$$

$$\alpha_{t,i} = \frac{\exp e_{t,i}}{\sum_{k=1}^{L} \exp e_{t,k}}$$

功能实现：图像生成文本

使用LSTM从编码中生成文本

每一步根据注意力区域生成下一个单词。

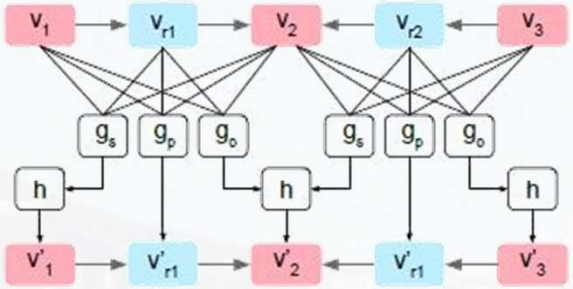


$$\mathbf{c}_0 = f_{init,c}\left(\frac{1}{L}\sum_{i=1} L\mathbf{a}_i\right)$$

$$\mathbf{h}_0 = f_{init,c}\left(\frac{1}{L}\sum_{i=1} L\mathbf{a}_i\right)$$

功能实现：文本生成图像 – 生成模型

GCN生成图像中每一个物体的位置向量。初始物体关系由场景图决定。



$$v'_r = g_p(v_i, v_r, v_j)$$

$$V_i^s = \{g_s(o_i, r, o_j) \mid (o_i, r, o_j) \in E\}$$

$$V_i^o = \{g_o(o_j, r, o_i) \mid (o_j, r, o_i) \in E\}$$

通过上述位置向量，生成每一个物体的蒙版，包括形状$\hat{m}$和位置$\hat{b}$

$$\hat{m} \sim M \times M$$

$$\hat{b} = (x_0, y_0, x_1, y_1)$$

CRN通过最近邻插值法的上取样的方法生成更清晰的图片
将图片的像素值从64x64提升至256x256

功能实现：文本生成图像 – 辨别模型

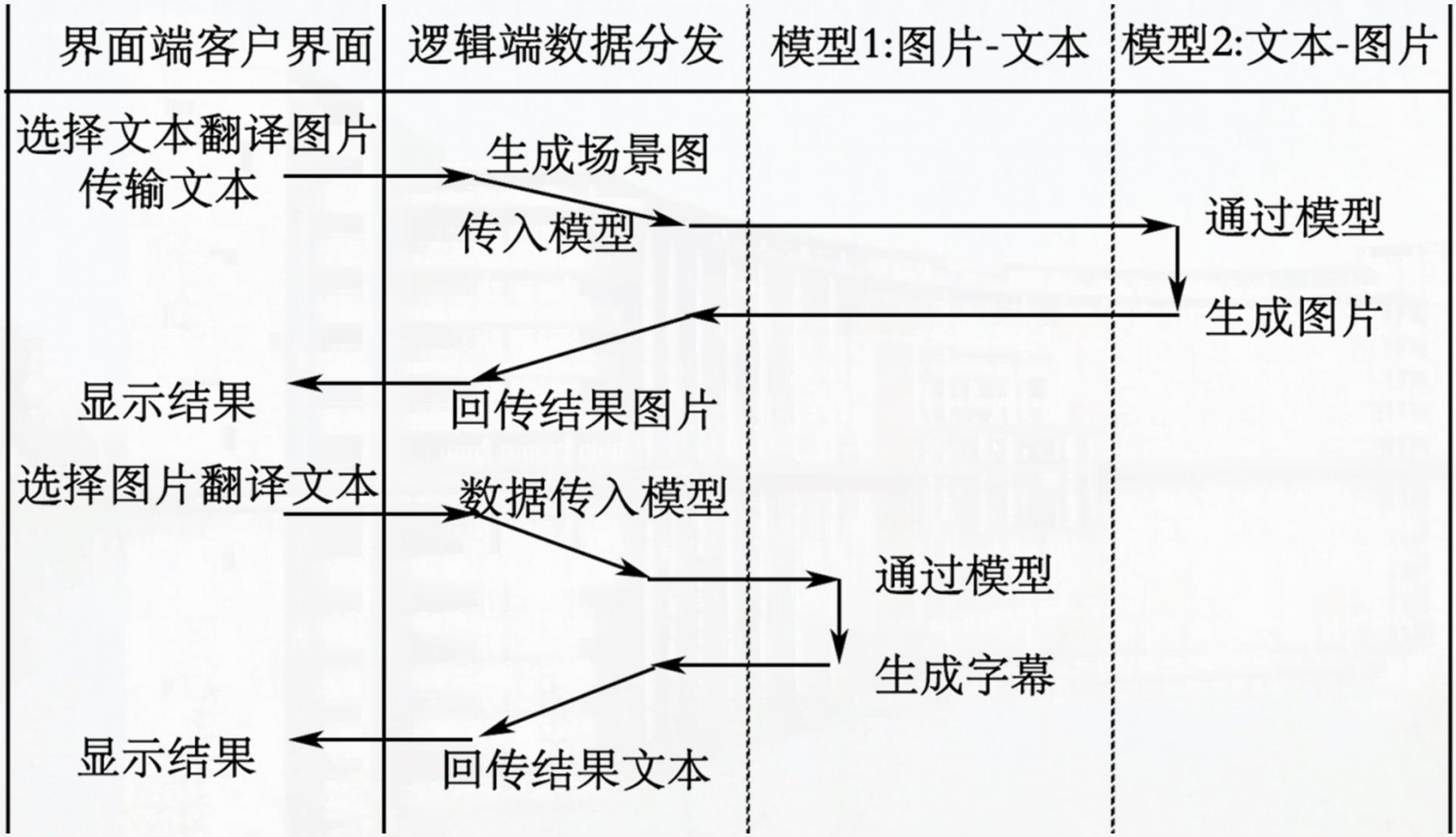
模型1: 对总体生成模型的生成图片Î对抗。

$$\mathcal{L}_{D_{img}} = \mathbb{E}_{x\sim p_{real}}[\lg D(x)] + \mathbb{E}_{x\sim p_{fake}}[1 - D(x)]$$

模型2: 与生成模型的第二部分生成的每个物体进行判别。

这一判别模型仅判别物体的分类，保证生成物体可识别。
这一判别模型与生成模型不进行对抗，仅作为保险机制。

前段用户交互界面设计：数据流图



前段用户交互界面设计：用例图

第三部分

实验与分析

- 介绍模型训练情况与表现
- 对比模型的与其他模型的表现

实验设置

服务器端：模型训练用

项目	说明
图形卡	GTX 2080 Ti （4块）
操作系统	Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-142-generic x86_64)
内存	64GB
位置	浙江大学教育网

PC端：客户端运行使用

项目	说明
设备型号	Macbook Pro电脑，A1398型号
图形卡	AMD Radeon R9 M370X 2048 MB（集成） （不使用）
操作系统	OS X 10.12 High Sierra
内存	16 GB 1600 MHz DDR3
处理器	2.5 GHz Intel Core i7

模型训练情况：图像生成文本

图像生成文本的图像字幕模型
使用tensorflow框架训练
数据集使用COCO2014数据集

训练集8.8万张图片，进行了100个epoch的训练，总用时233小时。
验证集3.7万张图片，测试用时2.6小时，相当于每张图片测试时长约0.25秒。

模型训练情况：文本生成图像

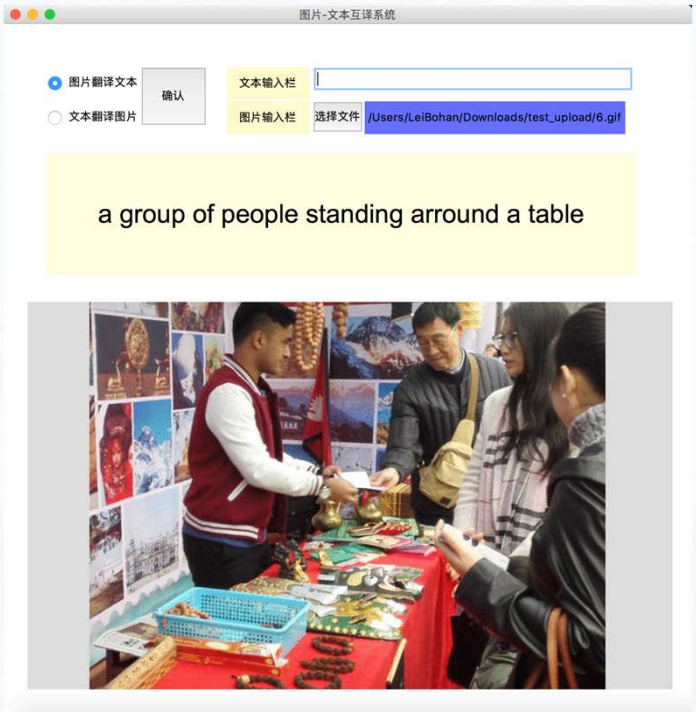
图像生成文本的图像字幕模型
使用pytorch框架训练数据集使用VG数据集

训练时间小于一个小时

模型训练表现：图像生成文本

我对训练的模型进行了BLEU评分
可以看出强关注模型表现在BLEU-4
评分中优于弱关注模型

在样例中，我使用数据集之外的私人
摄影图片进行了测试



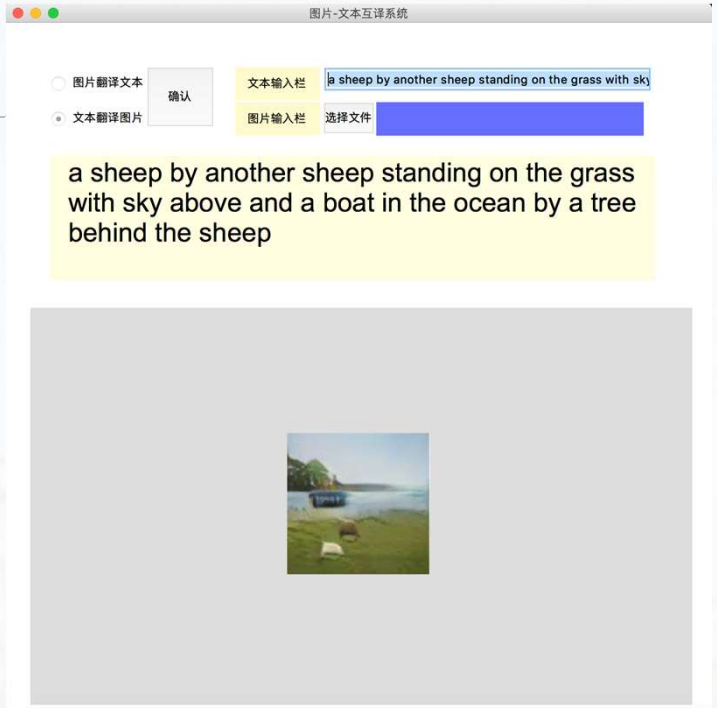
三种模型的 **BLEU** 评分

模型选择	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
弱关注模型	70.7%	49.2%	34.4%	24.3%	**23.90%**
强关注模型（我的训练）	70.3%	**53.6%**	**39.8%**	**29.5%**	——
强关注模型（文中数据）	**71.8%**	50.4%	35.7%	25.0%	23.04%

模型训练表现：文本生成图像

这个模型的图片准确度没有合适的评分体系进行评价，对其进行了“图片平滑程度”的图片质量评分。

在样例中，我使用了一段描述性文本进行测试



图片平滑性评分与 **StackGAN** 模型对比

模型类别	数据集类别	
	COCO	VG
文中模型（无 D_{img} 作用）[31]	5.6 ± 0.1	$\mathbf{5.7 \pm 0.3}$
文中模型	6.7 ± 0.1[31]	5.5 ± 0.1
StackGAN[24]	$\mathbf{8.4 \pm 0.2}$	-

第四部分

总结与展望

肆

- 总结设计制作情况并分析不足
- 展望设计进一步优化方向

设计的不足与进一步优化方向

图片生成文本功能对应的图片字幕模型表现较差。下一步应当设计更好的模型来实现这一功能，增强其性能。更合理的模型可以不需要进行100个epoch训练即可掌握训练集的内容，可以使用更丰富的数据集训练。

文本生成图片功能对应的GAN生成模型，只支持有明确位置关系的描述性语句，没有集成形容词对物体的描述。下一步有两个优化方向：1. 应当集成对物体的描述生成方法，对于生成的物体有更好的限制功能；2. 应当

本科阶段取得成果

- 专利：

  雷伯涵，彭亚楠，黄浩. 一种用于大数据分析的数据样本均匀采样方法及装置.（进入实质审查阶段）

- 论文：

  雷伯涵，陈畅，侯叶俏，孙月明，韩凌，黄浩.基于感染结果的传播网络推断方法. 软件学报. (第二轮审稿中，当前审稿意见：小修)

谢谢观看

敬请各位老师批评指正

指导教师: 黄浩、庄越挺

答辩学生: 雷伯涵