基于对抗性生成网络(GAN)的非配对语音转换

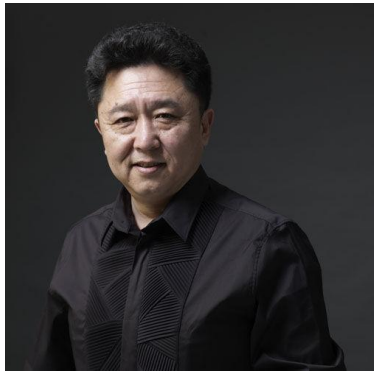
赵磊

梗概

- 语音转换的介绍
- 相关工作
- 技术困难
- 实验结果
- 总结

语音转换介绍

语音转换(Voice Conversion)是指将一个人的语音特征转换为另外一个人



使用场景

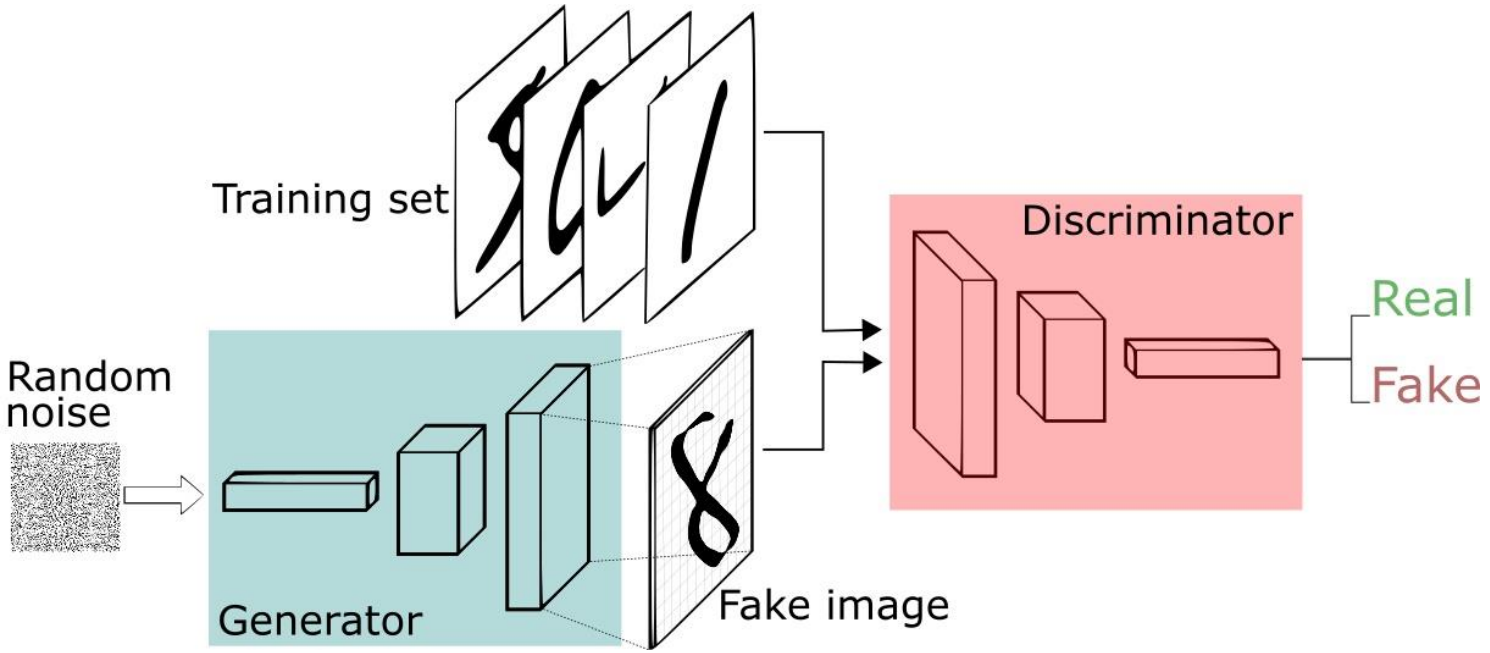
- 用作TTS的后端以改变声音特征, 或用作语音识别的数据扩增
- 视频配音
- 真人转变声音以及合成新音频等

相关工作

- 语音转换分为有监督数据和无监督数据两种
- 有监督数据的情况下算法更容易实现
  - 首先将音频转换为语音特征(MFCC等)
  - 再使用DTW将语音特征对齐
  - 使用配对算法(例如 GMM或神经网络)训练
  - 还原语音特征到音频完成语音转换
- 无监督数据的情况下更难实现
  - 首先将语音转换为语素(phoneme), 之后将语素转换为新的声音
  - 使用GAN来进行无监督的数据域转换

GAN

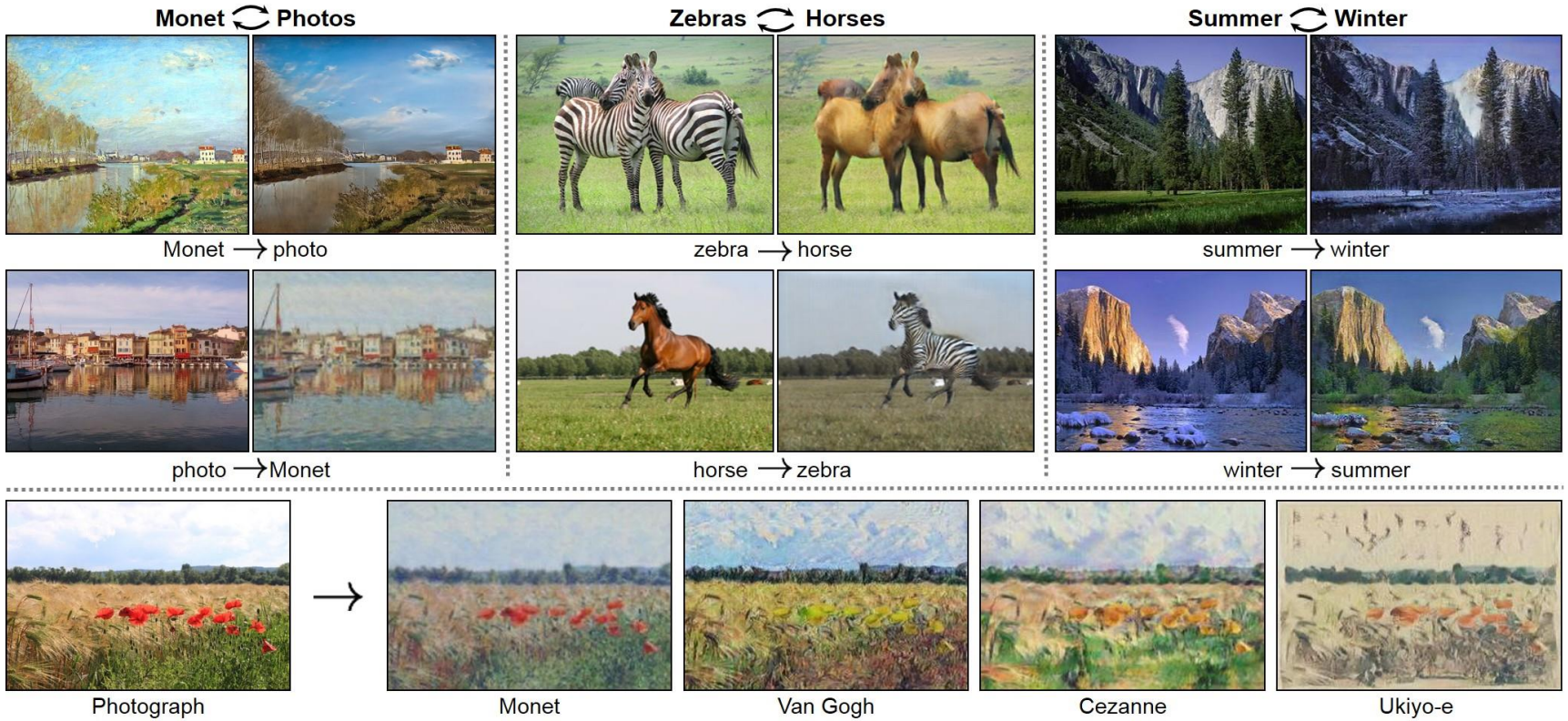
- 对抗性生成网络在无监督学习中大放异彩



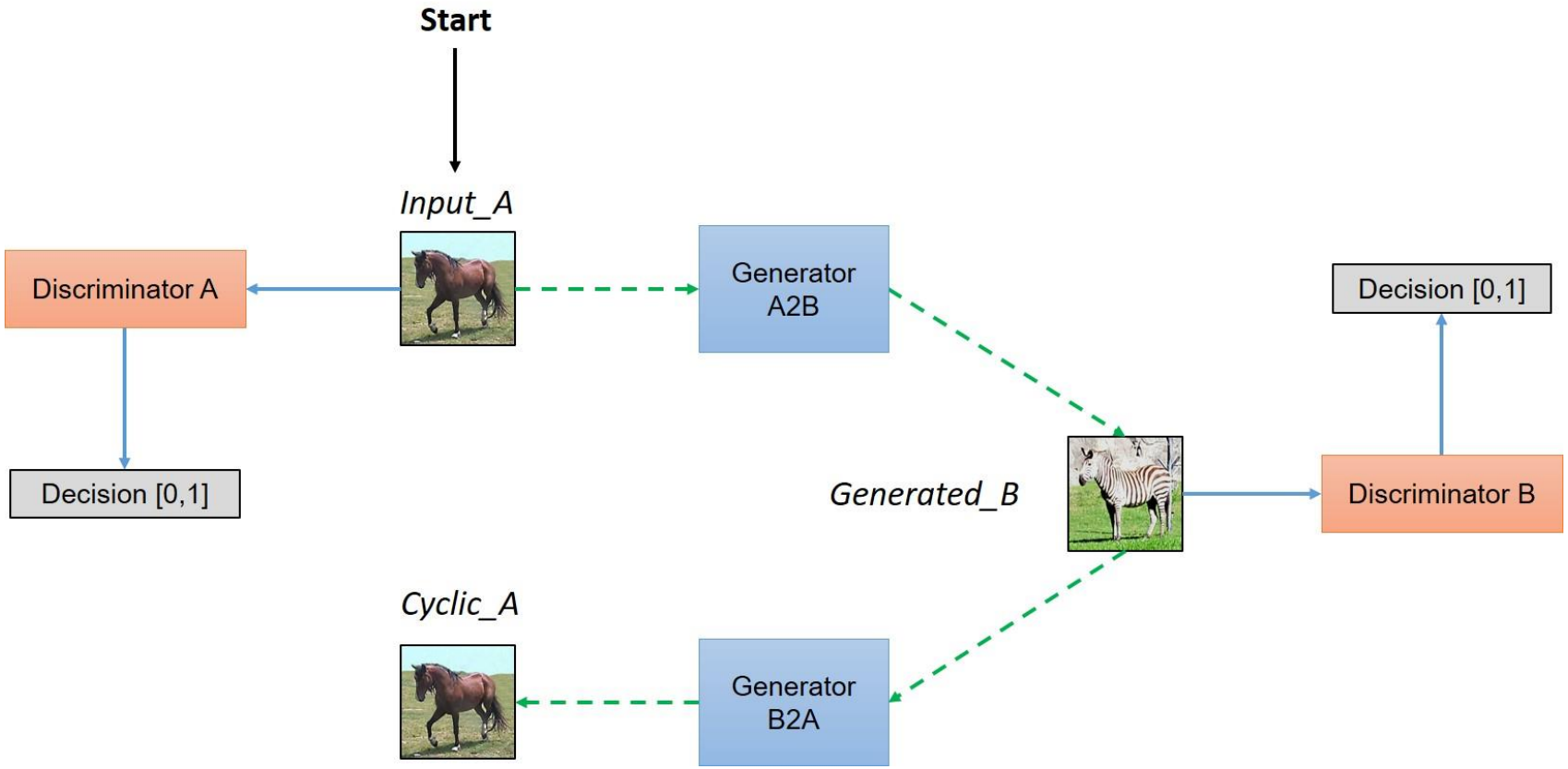
- 缺点是不好调参, 容易出现model collapse

CycleGan

- 神奇的CycleGan

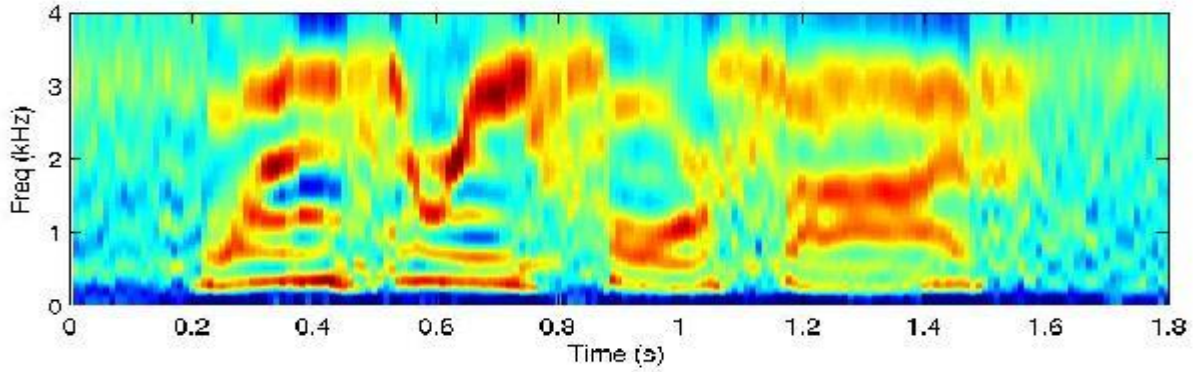


CycleGan结构


Start
Input_A
Discriminator A
Generator A2B
Decision [0,1]
Decision [0,1]
Generated_B
Discriminator B
Cyclic_A
Generator B2A

利用CycleGan做语音变换

- 需要将语音变为声学特征(例如MFCC或频谱)



- 直接转换语音特征图有困难
- 可以转而转换每个时间点的特征向量

转换结果

- 女声 
- 男声 
- 转换 

总结和挑战

- 无监督语音转换是可以实现的
- 现阶段只能实现细粒度的语音特征转换
- 粗粒度的语音特征需要进一步研究

感谢!

赵磊