摘　要

同时理解语言与视觉信息，跨越语言和视觉之间的模态语义鸿沟，是人工智能系统从感知智能迈向认知智能的关键。在语言的众多表达形式中，视觉描述是与视觉内容联系最为紧密的一种语言形式。视觉描述根据其描述形式，可以分为基于语音的视觉描述和基于文本的视觉描述。本文围绕基于视觉描述的跨模态图像生成问题，从文本到图像的生成，语音到图像的生成，以及跨模态语义压缩三个方面进行了系统的研究。总结来说，本文的主要贡献如下：

1. **一种基于因素分解的文本到图像的跨模态生成方法**。针对文本到图像生成中文本对图像可控性弱，文本语义和随机噪声相耦合的问题，提出一种新的跨模态语义嵌入方法，通过设计一种基于加法的实例正则化层，将文本语义和随机噪声通过不同方式加入到生成网络中。相关实验表明，该方法可以有效提升文本到图像生成模型中文本对生成图像的可控性，并提升生成模型的总体性能。同时，模型在效率上也有所改善。总的来说，该方法在模型性能，可控性和效率三个方面都有所改善。
2. **一种基于迁移学习的语音到图像的跨模态生成方法**。针对语音视觉描述到图像的跨模态生成问题，提出一种直接的从语音到图像的生成方法。该方法不借助于语音识别，直接提取语音视觉描述合成语义一致的图像。首先，针对语音-图像标注数据短缺的问题，利用文本-图像配对数据，借助文本到语音转换接口（Text-to-Speech, TTS）得到基于合成语音的配对数据集。其次，针对语音数据的语义表示在新的类别上泛化性困难的问题，提出一种基于迁移学习的方法，有效的提升了语音数据语义表示的泛化性，从而提升了语音到图像跨模态生成在新的类别上的泛化性。最后，在实验结果上，合成语音上的图像合成效果接近了基于文本的合成效果，验证了我们方法的有效性。本文实现了一种直接的语音视觉描述到图像的跨模态生成方法，构建了该问题的数据集，提出了一个基准方法，为后续对该问题的研究提供了基础。
3. **一种基于跨模态生成的语义压缩框架**。不同的数据模态有不同的特点，语音为一维时域信号，图像为二维空域信号，文本为离散的字符。数据

压缩的目标是设计一组匹配的编码和解码方法，在保证重构数据与原始数据满足一定保真的前提下，尽可能降低传输数据的比特率。本文借助视觉描述到图像跨模态生成中的相关方法，利用不同模态数据的特点，提出一种基于跨模态生成的语义压缩框架–跨模态压缩，并实现了一种面向图像压缩的图像-文本-图像的压缩框架和一种面向特定视频压缩的视频-音频 + 图像-视频的压缩框架。实验结果表明，本文所提方法在实验数据集上相比传统压缩方式有明显的优势。本文提出了跨模态压缩的概念，并给出相关测试方法和基准，为后续相关研究奠定了基础。

综上所述，本文围绕基于视觉描述的跨模态图像生成问题，针对视觉描述中的文本和语音两种视觉描述分别进行研究，提出了新的跨模态语义嵌入方式，实现了语音视觉描述到图像的跨模态生成，并提出一个新的语义压缩框架，为后续的研究奠定了基础。

关键词：图像生成，生成对抗网络，文本到图像的生成，语音到图像的生成

Abstract

Understanding language and vision simultaneously is the key to bridging the gap between language and vision for the artificial intelligence system to advance from "perceptual intelligence" to "cognitive intelligence". In the different expression forms of the language, visual description is the most related to visual content. According to the description form, a visual description can be classified into a text-based visual description and a speech-based visual description. In this paper, we focus on cross modal generation from the visual description and study this problem in three aspects: image generation from text, image generation from speech, and cross modal semantic compression. Summarily, the contributions of this paper are as follows:

1. **A factor decomposition based method for cross modal generation from text to image.** For the task of image generation from text visual description, to improve the controllability of the text-to-image generation model, we proposed a new cross modal semantic embedding method by designing an additive instance norm module. By decomposing the text condition and noise vector, we embedded the text condition into the generator and discriminator with the additive instance norm. Experimental results showed that our method can improve performance, controllability, and efficiency simultaneously.
2. **A transfer learning based method for cross modal generation from speech to image.** For the task of image generation from speech visual description, we proposed a method to directly translate speech into image without the help of the text. Firstly, to get speech-image paired data, we synthesized speech using a text-to-speech application to get a synthesized speech-image paired dataset based on the text-image paired dataset. Then we proposed a transfer learning-based method to improve the generalization ability of the speech representation. Finally, we conducted extensive experiments to demonstrate the effectiveness of our proposed methods. We propose a method to synthesize images from speech visual description, without the help of the text, providing

a benchmark for this problem.

3. **A cross modal generation based semantic compression framework.** Different modalities have different characteristics. For example, speeches are 1-dimensional time-domain continuous signal, text are 1-dimensional discrete data, images are 2-dimensional spatial-domain signal. The goal of data compression is to design a pair of methods: an encoder and a decoder, to decrease the rate under the constrain of distortion. We applied to cross modal translation methods to data compression and proposed a cross modal generation based semantic compression framework: cross modal compression, and provided a paradiam for cross modal image compression and a paradigm for cross modal video compression. Experimental results showed that our proposed cross modal compression performed better than traditional compression methods on the testing dataset. Our method sets a new direction for the next generation semantic based image/video compression.

Overall, we focused on the problem: cross modal generation from visual description to image, and investigated image generation from speech visual description and text visual description. We propose a new cross modal semantic embedding method, a framework for direct speech-to-image translation, and a new semantic compression framework.

Key Words: Image Generation, Generative Adversarial Nets, Text-to-Image Translation, Speech-to-Image Translation