Human-robot interactive language to image synthesis

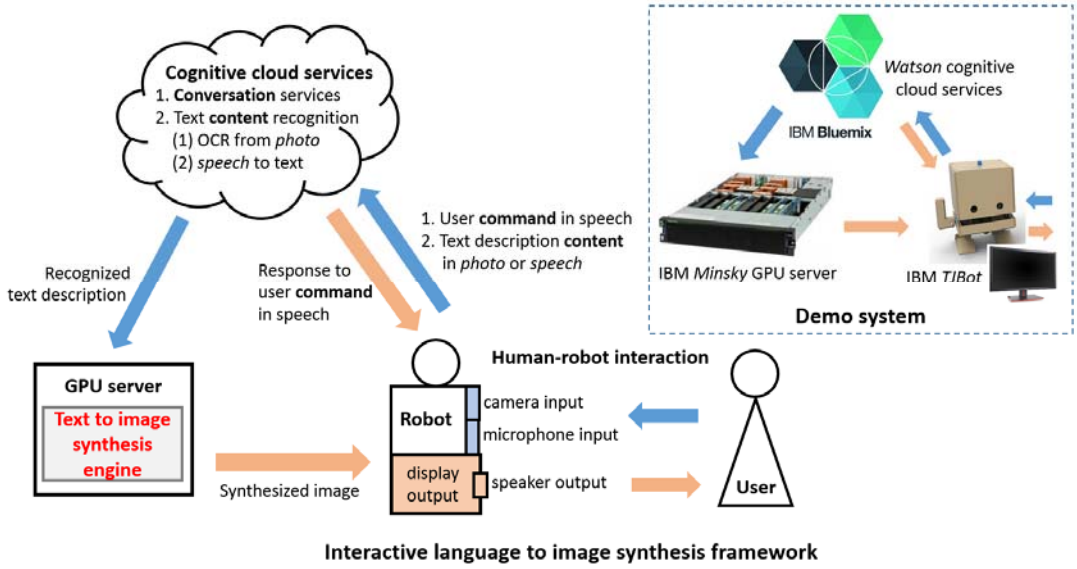
1. Background

Synthesizing visually perceptual images using mathematical models rather than fetching existing images from local storage or Internet database provides direct, fast and secure solution for image data generation in various use cases, while posing critical challenges on learning and synthesizing capabilities of models. Recent progress in deep learning research has shown that deep generative models can effectively grasp the semantic concepts with corresponding visual features to synthesis high-quality images according to input descriptions in natural languages. Currently, most image synthesis work has been focused on theoretical model architecting and hyperparameter tuning for improving model training stability and image synthesis quality. However, how we can seamlessly and efficiently integrate a language-to-image synthesis model into a real-world multi-functional intelligent system, such as robotic assistant, is still an open problem. Our work proposes a novel framework which incorporates deep learning driven image synthesizer with interactive robot interfaces and cognitive cloud services to enable intelligent and convenient language-to-image synthesis according to users' customized needs. A demo system is implemented as a case study based on the IBM TJBot with backend supports of IBM Minsky GPU server and IBM Bluemix cloud services.

2. Brief introduction

Our proposed interactive image synthesis framework takes users' input descriptions in natural language (words or sentences), synthesizes corresponding images according to the given input description and renders results to users in a human-robot interactive manner. The framework integrates three major modules: a front-end robot interface to interact with users, a cloud server to process raw input and enable human-robot conversation based on cognitive cloud services, and a GPU server to train and deploy the text-to-image synthesis model. To the best of our knowledge, this is the first work which integrates both cutting-edge deep generative models and real-world robotic systems for effective and interactive image synthesis from raw natural language. The generated images will be unique and distinguished from any existing images, and can also be imaginary and significantly customized, without infringing any intellectual property issues. This design can also be used as an educational tool to help low literacy people with language studies, or a business tool to facilitate communication in different languages with translated illustrative images. Furthermore, this work initiates the potential integration of language-to-image translation as a novel cognitive cloud service with existing text-to-text and text-to-speech translation services, supporting a variety of front-end embedded devices such as robotic assistants, mobile phones and smart glasses for more flexible and efficient interaction with end users.

3. Detailed description



Interactive language to image synthesis framework

In above figure, we show the architecture of our proposed interactive language to image synthesis framework with a hardware demo system.

A robot with camera and microphone serves as front-end interface to interact with users. User can talk with the robot using spoken commands to invoke the image synthesis process and the robot can also respond to user's command with underlying support of the conversation service in the cloud. To specify the content of input text descriptions, we provide two flexible ways: user can either show the written or printed text to the camera, or directly speak to the robot to describe what image they need. The input text description can be a word, such as "cat", "plane", or a sentence, such as "a small bird with red head and yellow wings". The robot then sends the raw input in photo or speech to the cloud and the text information can be extracted from the photo by optical character recognition (OCR) or from the speech by speech-to-text translation with corresponding cloud services. The cloud server then feeds the recognized text as input to the GPU server to perform text-to-image synthesis. Here, on the GPU server, we have trained a state-of-the-art model based on generative adversarial network (GAN) offline as the text-to-image synthesis engine. The GPU server will load the trained GAN model at runtime, take the processed text input from cloud, and synthesize corresponding image in real-time. Finally, the synthesized image according to users' language description will be sent to the front-end robot and rendered on the display.

We have built a demo system to realize the interactive image synthesis framework. Specifically, an IBM TJBot with complete I/O functionalities is used as the interactive interface to communicate with user, take user's input language, and render the output

image. The human-robot conversation and text recognition are supported by IBM Bluemix cloud system and implemented with Watson conversation, visual recognition and speech recognition APIs. The GAN-based image synthesis model is implemented using TensorFlow and Caffe deep learning frameworks with underlying CUDA libraries for GPU acceleration. We use the IBM Minsky GPU server, which is installed with 4 Tesla P100 Pascal GPUs and high-bandwidth NvLink interconnections, to train and deploy the image synthesis model with high efficiency. The whole process of taking user's raw language input to rendering the synthesized image output can complete within only a few seconds.

We show some representative image synthesis results with corresponding text inputs in above figure, and demonstrate the whole process with our demo system in a short video.