BILINEAR REPRESENTATION FOR LANGUAGE-BASED IMAGE EDITING USING CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

Xiaofeng Mao; Yuefeng Chen; Yuhong Li; Tao Xiong; Yuan He; Hui Xue
Alibaba Group, China

Introduction

What is language-based image editing?
What if you could tell an AI to edit an image just by describing what the new one should look like? Language-based image editing, which edits images using human linguistic input and AI processing, is already starting to see application in fashion, VR, and CAD.

Like in Fig 1, using LBIE technique, one can automatically modify the color, texture or style for a given design drawing by language instructions instead of the traditional complex processes.

Existing literature on LBIE using cGAN
The cGAN [1] approach edits the image based on fused visual-text representations using one of two conditioning methods. The first is concatenation. The second improved approach is Feature-wise Linear Modulation (FiLM) [2], which seeks to mimic the human attention mechanism.

Motivation

Traditional cGAN:
Concatenation and FiLM only apply a linear transformation between the input and conditional features. In this work, we go a step further and generalize these linear methods to the more powerful bilinear version, which can provide richer representations than linear models by learning the second-order interaction.

Feature-wise linear modulation (FiLM):
Feature-wise linear modulation:

Table 1. The comparison of IS score of methods

Model details
The network architecture is shown in Fig 2. The network consists of a generator G and a discriminator D. The text and image features are fed in the fusing module, which consists of N Bilinear Residual Layer (BRL). The decoding module Ø upsamples the fused feature to a high-resolution images. We propose Bilinear Residual Layer for learning conditional bilinear representations. We add some shortcuts to guarantee model's capability to learn identical mapping, and adopts a low-rank bilinear method [3] to simplified the calculation of bilinear transformation.

Adversarial training objective
For T → mismatching text, t → matching text, f → manipulating text.

The generator G is trained distinguish semantically differentiated image-text pairs:

The generator G is trained to generate more semantically similar images with the editing text t:

Conclusions

In this work, we propose a conditional GAN based encoder-decoder architecture to semantically manipulate images by text descriptions. A general condition layer called Bilinear Residual Layer (BRL) is proposed to learn more powerful bilinear representations for LBIE. BRL is also applicable for other common conditional tasks. Our evaluation results on Caltech-200 bird dataset, Oxford-102 flower dataset and Fashion Synthesis dataset achieve plausible effects and outperform the state-of-the-art methods on LBIE.

References

Contact
Name: Xiaofeng Mao
Department: Turing Laboratory of Alibaba Security Department, Alibaba Group
Address: No. 903 West West Road, Yuhang District, Hangzhou City, Zhejiang Province
Email: mxf164419@alibaba-inc.com
Phone: +86 1555 13421
GitHub: https://github.com/xdlqg