CUAI CV 리딩 스터디 1팀

2022.03.29

발표자 : 김민규

스터디원 소개 및 만남 인증

스터디원 1 : 김민규

스터디원 2 : 김지욱

스터디원 3 : 이하윤

논문 리뷰 1 : Big Transfer(김민규)

Big Transfer (BiT):
General Visual Representation Learning

Alexander Kolesnikov*, Lucas Beyer*, Xiaohua Zhai*
Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby

Google Research, Brain Team
Zürich, Switzerland
{akolesnikov,lbeyer,xzhai}@google.com
{jpuigcerver,jessicayung,sylvaingelly,neilhoulsby}@google.com

Abstract. Transfer of pre-trained representations improves sample efficiency and simplifies hyperparameter tuning when training deep neural networks for vision. We revisit the paradigm of pre-training on large supervised datasets and fine-tuning the model on a target task. We scale up pre-training, and propose a simple recipe that we call Big Transfer (BiT). By combining a few carefully selected components, and transferring using a simple heuristic, we achieve strong performance on over 20 datasets. BiT performs well across a surprisingly wide range of data regimes — from 1 example per class to 1 M total examples. BiT achieves 87.5% top-1 accuracy on ILSVRC-2012, 99.4% on CIFAR-10, and 76.3% on the 19 task Visual Task Adaptation Benchmark (VTAB). On small datasets, BiT attains 76.8% on ILSVRC-2012 with 10 examples per class, and 97.0% on CIFAR-10 with 10 examples per class. We conduct detailed analysis of the main components that lead to high transfer performance.

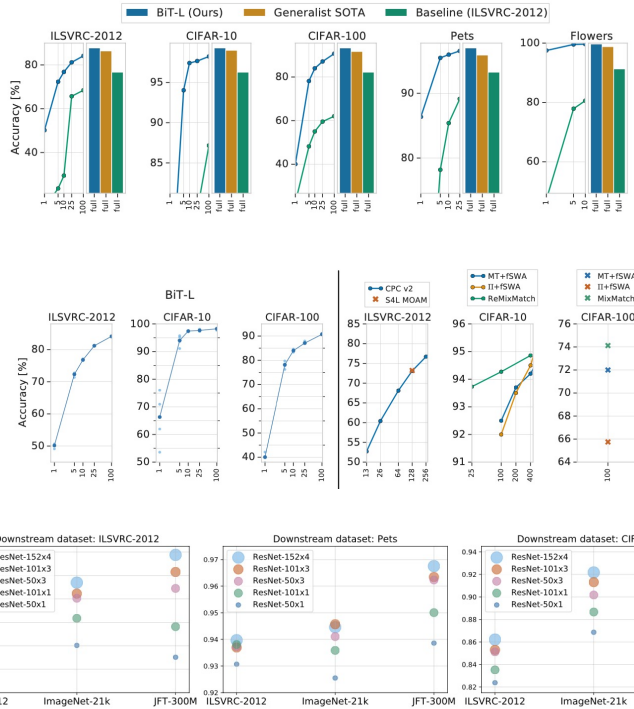
1 Introduction

Strong performance using deep learning usually requires a large amount of task-specific data and compute. These per-task requirements can make new tasks prohibitively expensive. Transfer learning offers a solution: task-specific data and compute are replaced with a pre-training phase. A network is trained once on a large, generic dataset, and its weights are then used to initialize subsequent tasks which can be solved with fewer data points, and less compute [40,44,14].

We revisit a simple paradigm: pre-train on a large supervised source dataset, and fine-tune the weights on the target task. Numerous improvements to deep network training have recently been introduced, e.g. [55,62,26,35,22,1,64,67,54,66]. We aim not to introduce a new component or complexity, but to provide a recipe that uses the minimal number of tricks yet attains excellent performance on many tasks. We call this recipe "Big Transfer" (BiT).

We train networks on three different scales of datasets. The largest, BiT-L is trained on the JFT-300M dataset [51], which contains 300 M noisily labelled

* Equal contribution



논문 리뷰 2 : Grad-CAM(이하윤)

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

Abstract We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent and explainable.

Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (*e.g.* VGG), (2) CNNs used for structured outputs (*e.g.* captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g.* visual question answering) or reinforcement learning, all *without architectural changes or re-training*. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures.

In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are robust to adversarial perturbations, (d) are more faithful to the underlying model, and (e) help achieve model generalization by identifying dataset bias.

For image captioning and VQA, our visualizations show that even non-attention based models learn to localize discriminative regions of input image.

We devise a way to identify important neurons through Grad-CAM and combine it with neuron names [4] to provide textual explanations for model decisions. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions. Our code is available at https://github.com/ramprs/grad-cam/, along with a demo on CloudCV [2][1], and a video at youtu.be/COjUB9Izk6E.

Ramprasaath R. Selvaraju
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: ramprs@gatech.edu

Michael Cogswell
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: cogswell@gatech.edu

Abhishek Das
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: abhishkd@gatech.edu

Ramakrishna Vedantam
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: vrama@gatech.edu

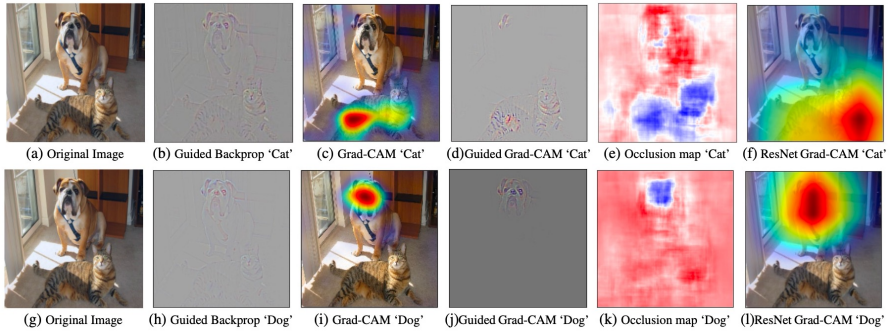
Devi Parikh
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: parikh@gatech.edu

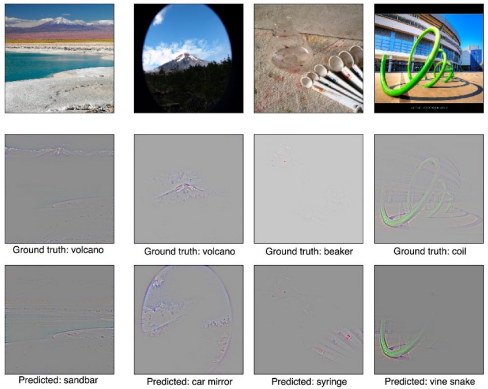
Dhruv Batra
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: dbatra@gatech.edu

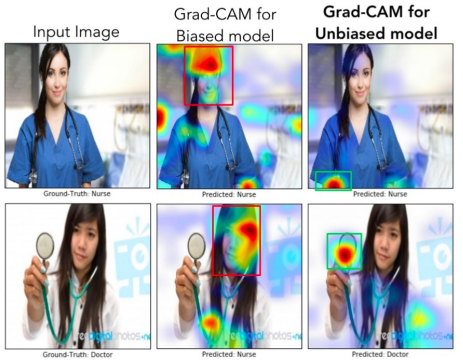
1 Introduction

Deep neural models based on Convolutional Neural Networks (CNNs) have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification [33,24], object detection [21], semantic segmentation [37] to image captioning [55,7,18,29], visual question answering [3,20,42,46] and more recently, visual dialog [11,13,12] and embodied question answering [10,23]. While

[1] http://gradcam.cloudcv.org







논문 리뷰 3 : GIRAFFE(김지욱)

GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields

Michael Niemeyer[1,2] Andreas Geiger[1,2]
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]University of Tübingen
{firstname.lastname}@tue.mpg.de

Abstract

Deep generative models allow for photorealistic image synthesis at high resolutions. But for many applications, this is not enough: content creation also needs to be controllable. While several recent works investigate how to disentangle underlying factors of variation in the data, most of them operate in 2D and hence ignore that our world is three-dimensional. Further, only few works consider the compositional nature of scenes. Our key hypothesis is that incorporating a compositional 3D scene representation into the generative model leads to more controllable image synthesis. Representing scenes as compositional generative neural feature fields allows us to disentangle one or multiple objects from the background as well as individual objects' shapes and appearances while learning from unstructured and unposed image collections without any additional supervision. Combining this scene representation with a neural rendering pipeline yields a fast and realistic image synthesis model. As evidenced by our experiments, our model is able to disentangle individual objects and allows for translating and rotating them in the scene as well as changing the camera pose.

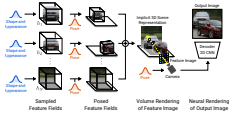


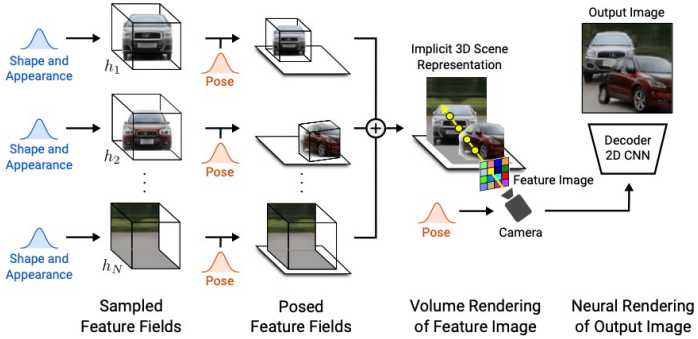
Figure 1: **Overview.** We represent scenes as compositional generative neural feature fields. For a randomly sampled camera, we volume render a feature image of the scene based on individual feature fields. A 2D neural rendering network converts the feature image into an RGB image. While training only on raw image collections, at test time we are able to control the image formation process wrt. camera pose, object poses, as well as the objects' shapes and appearances. Further, our model generalizes beyond the training data, e.g. we can synthesize scenes with more objects than were present in the training images. Note that for clarity we visualize volumes in color instead of features.

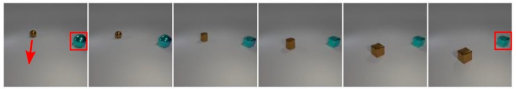
1. Introduction

The ability to generate and manipulate photorealistic image content is a long-standing goal of computer vision and graphics. Modern computer graphics techniques achieve impressive results and are industry standard in gaming and movie productions. However, they are very hardware expensive and require substantial human labor for 3D content creation and arrangement.

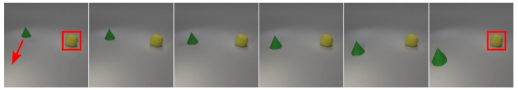
In recent years, the computer vision community has made great strides towards highly-realistic image generation. In particular, Generative Adversarial Networks (GANs) [24] emerged as a powerful class of generative models. They are able to synthesize photorealistic images at resolutions of 1024^2 pixels and beyond [6,14,15,39,40].

Despite these successes, synthesizing realistic 2D images is not the only aspect required in applications of generative models. The generation process should also be controllable in a simple and consistent manner. To this end, many works [9,25,39,43,44,48,54,71,74,97,98] investigate how disentangled representations can be learned from data without explicit supervision. Definitions of disentanglement vary [5,53], but commonly refer to being able to control an attribute of interest, e.g. object shape, size, or pose, without changing other attributes. Most approaches, however, do not consider the compositional nature of scenes and operate in the 2D domain, ignoring that our world is three-dimensional. This often leads to entangled representations (Fig. 2) and control mechanisms are not built-in, but need to be discovered in the latent space a posteriori. These properties, however, are crucial for successful applications, e.g. a movie production where complex object trajectories





(a) Translation of Left Object (2D-based Method [71])



(b) Translation of Left Object (Ours)

(c) Circular Translation (Ours)

(d) Add Objects (Ours)

추후 계획

4/5 논문 리뷰 스터디