2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

Residual Dense Information Distillation Network for Single Image Super-Resolution

Qiaosong Chen, Jinxin Li, Bolin Duan, Liu Pu, Xin Deng, Jin Wang

***Abstract*—In recent years, Convolutional Neural Networks (CNNs) have achieved excellent results in the study of single image super-resolution. However, super-resolution algorithms based on CNNs still face serious challenges, such as poor detail reconstruction, numerous parameters, and difficulty of training. A Residual Dense Information Distillation Network (RD-IDN) is proposed in this paper which uses dense skip connections and residual structure to solve the problems of difficult training and low utilization of features in Information Distillation Network. Experimental results show that the proposed method is superior to many other Super Resolution algorithms in terms of reconstruction performance and computational comsumption.**

I. INTRODUCTION

Single image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution image (LR). As shown in Fig. 1, SR needs to increase the spatial scale and resolution of the LR image.

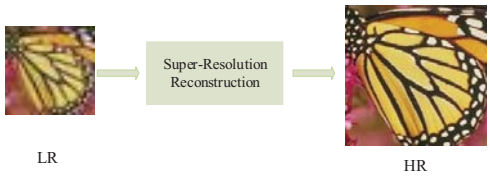


Figure 1. Single image Super-Resolution

At present, the learning-based SR algorithms works outperform many previous works, including dictionary learning [1][2], linear regression [3], random forest [4], and deep learning, which learns the mapping between LR images and HR images from a large amount of data. However, an infinite number of HR images can get the same LR image by downsampling, and thus the SR problem is inherently ill-posed.

Recently, the performance of SR has been significantly improved due to the rapid development of Convolutional Neural Networks (CNNs). Dong proposed SRCNN [5] to reconstruct LR images by using a neural network with three layers of convolution, which achieved a significant improvement in SR performance. The FSRCNN [6] accelerates the network by using the transposed convolution. Kim [7] proposed VDSR, which increases the depth of network by using skip connections and gradient clipping technology to ease training successfully. The DRCN [8] improves the performance of SR by using recursive layer to control parameters of the network. ESPCN [9] use sub-pixel convolution and tanh activation function to improve SR performance. The RED [10] consists of a series of convolution layers and transposed convolution layers which learn end-to-end mapping from LR images to HR images. The deep recursive residual network (DRRN) was proposed by Tai [11], which employs parameters sharing strategy to alleviate the requirement of enormous parameters of very deep networks. Lai [12] proposed LapSRN, which reconstructs the sub-band residual of HR images progressively, and trained with Charbonnier loss function. The IDN proposed by Hui [13] combines the enhancement unit and the compression unit into one distillation block to extract useful features of LR images. IDN is superior to other methods in terms of reconstruction performance and computational performance.

Although IDN has achieved superior performance, it still has shortcomings. Firstly, it uses only one transposed convolution layer as the upsampling layer, and cannot fully utilize the feature information of LR images. Secondly, many features are lost during the forward propagation of training, which makes the network difficult to train. Hence, it requires complex training tricks and optimization methods to achieve better results.

To solve problems mentioned above, the Residual Dense Information Distillation Network (RD-IDN) is proposed in this paper to improve IDN [13]. At first, we employ dense skip connections to make full use of the features of LR images. Then we add residual structure to the IDB [13] and learn the mapping between LR images to HR images directly to avoid the problem of features and gradient disappearance during training. We have improved the learning ability of the network, and make it easy to optimize and reduce the difficulty of training.

The rest of this paper is organized as follows. Section 2 provides the proposed method in detail. Section 3 shows and analysis the experimental results. Section 4 summarizes this paper.

II. APPROACH

A. Network Architecture

As shown in Fig. 2, RD-IDN consists of three parts: feature extraction module, feature enhancement module, and upsampling module. The feature extraction module extracts features of the LR image, which consists of several convolutional layers with ReLu activation function. More useful features are extracted by the feature enhancement

Research supported by the National Nature Science Foundation of China (No. 61806033), the Key Industry Core Technology Innovation Project of CQ (cstc2017zdcy-zdyfX0012), the National Social Foundation of China (No. 18XGL013).

Qiaosong Chen is with the Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (email: chenqs@cqupt.edu.cn).

Plug-and-Play Methods Provably Converge with Properly Trained Denoisers

Ernest K. Ryu [1] **Jialin Liu** [1] **Sicheng Wang** [2] **Xiaohan Chen** [2] **Zhangyang Wang** [2] **Wotao Yin** [1]

Abstract

Plug-and-play (PnP) is a non-convex framework that integrates modern denoising priors, such as BM3D or deep learning-based denoisers, into ADMM or other proximal algorithms. An advantage of PnP is that one can use pre-trained denoisers when there is not sufficient data for end-to-end training. Although PnP has been recently studied extensively with great empirical success, theoretical analysis addressing even the most basic question of convergence has been insufficient. In this paper, we theoretically establish convergence of PnP-FBS and PnP-ADMM, without using diminishing stepsizes, under a certain Lipschitz condition on the denoisers. We then propose real spectral normalization, a technique for training deep learning-based denoisers to satisfy the proposed Lipschitz condition. Finally, we present experimental results validating the theory.

1. Introduction

Many modern image processing algorithms recover or denoise an image through the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)+\gamma g(x),$$

where the optimization variable $x\in\mathbb{R}^d$ represents the image, $f(x)$ measures data fidelity, $g(x)$ measures noisiness or complexity of the image, and $\gamma\geq 0$ is a parameter representing the relative importance between f and g. Total variation denoising, inpainting, and compressed sensing fall under this setup. *A priori* knowledge of the image, such as that the image should have small noise, is encoded in $g(x)$. So $g(x)$ is small if x has small noise or complexity. *A posteriori* knowledge of the image, such as noisy or partial measurements of the image, is encoded in $f(x)$. So $f(x)$ is small if x agrees with the measurements.

First-order iterative methods are often used to solve such optimization problems, and ADMM is one such method:

$$\begin{aligned}
x^{k+1} &= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\sigma^2 g(x)+(1/2)\|x-(y^k-u^k)\|^2\right\}\\
y^{k+1} &= \underset{y\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\alpha f(y)+(1/2)\|y-(x^{k+1}+u^k)\|^2\right\}\\
u^{k+1} &= u^k+x^{k+1}-y^{k+1}
\end{aligned}$$

with $\sigma^2=\alpha\gamma$. Given a function h on $\mathbb{R}^d$ and $\alpha>0$, define the proximal operator of h as

$$\operatorname{Prox}_{\alpha h}(z)=\underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\alpha h(x)+(1/2)\|x-z\|^2\right\},$$

which is well-defined if h is proper, closed, and convex. Now we can equivalently write ADMM as

$$\begin{aligned}
x^{k+1} &= \operatorname{Prox}_{\sigma^2 g}(y^k-u^k)\\
y^{k+1} &= \operatorname{Prox}_{\alpha f}(x^{k+1}+u^k)\\
u^{k+1} &= u^k+x^{k+1}-y^{k+1}.
\end{aligned}$$

We can interpret the subroutine $\operatorname{Prox}_{\sigma^2 g}:\mathbb{R}^d\to\mathbb{R}^d$ as a denoiser, i.e.,

$$\operatorname{Prox}_{\sigma^2 g}: \text{noisy image}\mapsto\text{less noisy image}$$

(For example, if σ is the noise level and $g(x)$ is the total variation (TV) norm, then $\operatorname{Prox}_{\sigma^2 g}$ is the standard Rudin–Osher–Fatemi (ROF) model (Rudin et al., 1992).) We can think of $\operatorname{Prox}_{\alpha f}:\mathbb{R}^d\to\mathbb{R}^d$ as a mapping enforcing consistency with measured data, i.e.,

$$\operatorname{Prox}_{\alpha f}: \text{less consistent}\mapsto\text{more consistent with data}$$

More precisely speaking, for any $x\in\mathbb{R}^d$ we have

$$g(\operatorname{Prox}_{\sigma^2 g}(x))\leq g(x), \qquad f(\operatorname{Prox}_{\alpha f}(x))\leq f(x).$$

However, some state-of-the-art image denoisers with great empirical performance do not originate from optimization problems. Such examples include non-local means (NLM) (Buades et al., 2005), Block-matching and 3D filtering (BM3D) (Dabov et al., 2007), and convolutional neural

[1]Department of Mathematics, University of California, Los Angeles, USA [2]Department of Computer Science and Engineering, Texas A&M University, USA. Correspondence to: Wotao Yin <wotaoyinmath.ucla.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Model Learning: Primal Dual Networks for Fast MR imaging

Jing Cheng[1], Haifeng Wang[1], Leslie Ying[3], Dong Liang[1, 2(✉)]

[1] Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
[2] Research center for Medical AI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
Dong.Liang@siat.ac.cn
[3] Departments of Biomedical Engineering and Electrical Engineering, University at Buffalo, the State University of New York, Buffalo, NY 14260 USA

Abstract. Magnetic resonance imaging (MRI) is known to be a slow imaging modality and undersampling in k-space has been used to increase the imaging speed. However, image reconstruction from undersampled k-space data is an ill-posed inverse problem. Iterative algorithms based on compressed sensing have been used to address the issue. In this work, we unroll the iterations of the primal-dual hybrid gradient algorithm to a learnable deep network architecture, and gradually relax the constraints to reconstruct MR images from highly undersampled k-space data. The proposed method combines the theoretical convergence guarantee of optimization methods with the powerful learning capability of deep networks. As the constraints are gradually relaxed, the reconstruction model is finally learned from the training data by updating in k-space and image domain alternatively. Experiments on in vivo MR data demonstrate that the proposed method achieves superior MR reconstructions from highly undersampled k-space data over other state-of-the-art image reconstruction methods.

Keywords: MR reconstruction, Primal dual, Deep learning.

1 Introduction

Accelerating magnetic resonance imaging (MRI) has been an ongoing research topic since its invention in the 1970s. Among a variety of acceleration techniques, compressed sensing (CS) has become an important strategy during the past decades [1]. In general, the imaging model of CS-based methods can be written as

$$\min_{m} \frac{1}{2}\|Am - f\|_2^2 + \lambda\|\Psi m\|_1 \tag{1}$$

where the first term is the data consistency and the second term is the sparse prior. Ψ is a sparse transform, such as wavelet transform or total variation, m is the image to be reconstructed, A is the encoding matrix, f denotes the acquired k-space data.

Revisiting ResNets: Improved Training and Scaling Strategies

Irwan Bello [1] **William Fedus** [1] **Xianzhi Du** [1] **Ekin D. Cubuk** [1] **Aravind Srinivas** [2] **Tsung-Yi Lin** [1]
Jonathon Shlens [1] **Barret Zoph** [1]

Abstract

Novel computer vision architectures monopolize the spotlight, but the impact of the model architecture is often conflated with simultaneous changes to training methodology and scaling strategies. Our work revisits the canonical ResNet (He et al., 2015) and studies these three aspects in an effort to disentangle them. Perhaps surprisingly, we find that training and scaling strategies may matter more than architectural changes, and further, that the resulting ResNets match recent state-of-the-art models. We show that the best performing scaling strategy depends on the training regime and offer two new scaling strategies: (1) scale model depth in regimes where overfitting can occur (width scaling is preferable otherwise); (2) increase image resolution more slowly than previously recommended (Tan & Le, 2019). Using improved training and scaling strategies, we design a family of ResNet architectures, ResNet-RS, which are 1.7x - 2.7x faster than EfficientNets on TPUs, while achieving similar accuracies on ImageNet. In a large-scale semi-supervised learning setup, ResNet-RS achieves 86.2% top-1 ImageNet accuracy, while being 4.7x faster than EfficientNet-NoisyStudent. The training techniques improve transfer performance on a suite of downstream tasks (rivaling state-of-the-art self-supervised algorithms) and extend to video classification on Kinetics-400. We recommend practitioners use these simple revised ResNets as baselines for future research.

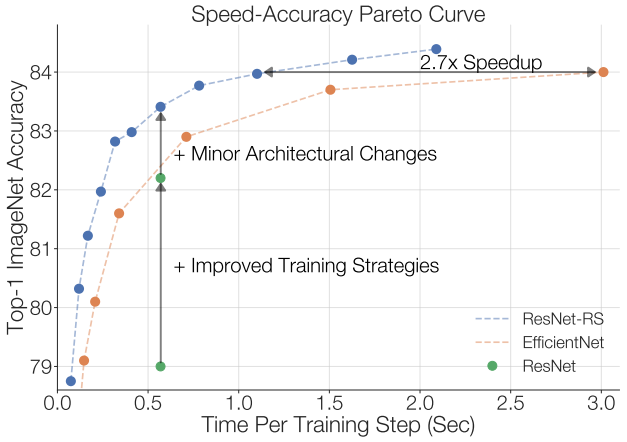


Figure 1. **Improving ResNets to state-of-the-art performance.** We improve on the canonical ResNet (He et al., 2015) with modern training methods (as also used in EfficientNets (Tan & Le, 2019)), minor architectural changes and improved scaling strategies. The resulting models, **ResNet-RS**, outperform EfficientNets on the speed-accuracy Pareto curve with speed-ups ranging from **1.7x - 2.7x** on TPUs and **2.1x - 3.3x** on GPUs. ResNet (•) is a ResNet-200 trained at 256×256 resolution. Training times reported on TPUs.

1. Introduction

The performance of a vision model is a product of the architecture, training methods and scaling strategy. However, research often emphasizes architectural changes. Novel architectures underlie many advances, but are often simultaneously introduced with other critical – and less publicized – changes in the details of the training methodology and hyperparameters. Additionally, new architectures enhanced by modern training methods are sometimes compared to older architectures with dated training methods (e.g. ResNet-50 with ImageNet Top-1 accuracy of 76.5% (He et al., 2015)). Our work addresses these issues and empirically studies the impact of *training methods* and *scaling strategies* on the popular ResNet architecture (He et al., 2015).

We survey the modern training and regularization techniques widely in use today and apply them to ResNets (Figure 1). In the process, we encounter interactions between

[1]Google Brain [2]UC Berkeley. Correspondence to: Irwan Bello and Barret Zoph <{ibello,barretzoph}@google.com>.

* Code and checkpoints available in TensorFlow:
https://github.com/tensorflow/models/tree/master/official/vision/beta
https://github.com/tensorflow/tpu/tree/master/models/official/resnet/resnet_rs

VOLO: Vision Outlooker for Visual Recognition

Li Yuan[1,2*] Qibin Hou[2*] Zihang Jiang[2] Jiashi Feng[1,2] Shuicheng Yan[1]
[1]Sea AI Lab [2]National University of Singapore
{ylustcnus,andrewhoux,jzh0103}@gmail.com, {fengjs, yansc}@sea.com

Abstract

Visual recognition has been dominated by convolutional neural networks (CNNs) for years. Though recently the prevailing vision transformers (ViTs) have shown great potential of self-attention based models in ImageNet classification, their performance is still inferior to that of the latest SOTA CNNs if no extra data are provided. In this work, we try to close the performance gap and demonstrate that attention-based models are indeed able to outperform CNNs. We find a major factor limiting the performance of ViTs for ImageNet classification is their low efficacy in encoding fine-level features into the token representations. To resolve this, we introduce a novel outlook attention *and present a simple and general architecture, termed Vision Outlooker (VOLO). Unlike self-attention that focuses on global dependency modeling at a coarse level, the outlook attention efficiently encodes finer-level features and contexts into tokens, which is shown to be critically beneficial to recognition performance but largely ignored by the self-attention. Experiments show that our VOLO achieves 87.1% top-1 accuracy on ImageNet-1K classification, which is the first model exceeding 87% accuracy on this competitive benchmark, without using any extra training data. In addition, the pre-trained VOLO transfers well to downstream tasks, such as semantic segmentation. We achieve 84.3% mIoU score on the cityscapes validation set and 54.3% on the ADE20K validation set. Code is available at* https://github.com/sail-sg/volo.

1. Introduction

Modeling in visual recognition, which was long dominated by convolutional neural networks (CNNs), has recently been revolutionized by Vision Transformers (ViTs) [14, 51, 68]. Different from CNNs that aggregate and transform features via local and dense convolutional kernels, ViTs directly model long-range dependencies of local patches (*a.k.a.* tokens) through the self-attention mechanism which is with greater flexibility in modeling visual contents. Despite the remarkable effectiveness on visual recognition [37, 32, 52, 79], the performance of ViT models still lags behind that of the state-of-the-art CNN models. For instance, as shown in Table 1, the state-of-the-art transformer-based CaiT [52] attains 86.5% top-1 accuracy on ImageNet, which however is still 0.3% lower compared with the 86.8% top-1 accuracy achieved by the CNN-based NFNet-F5 [2] with SAM and augmult [15, 16].

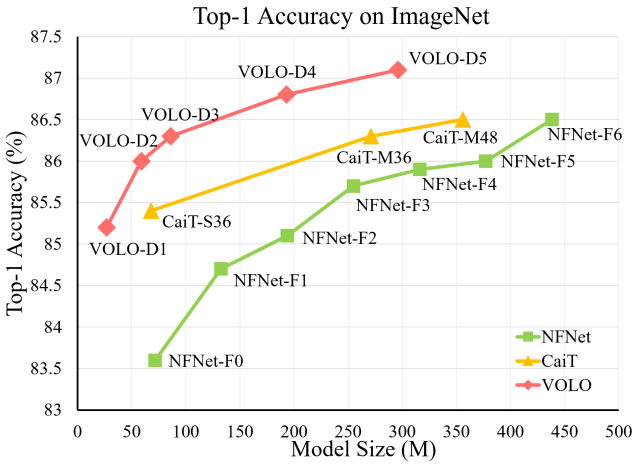


Figure 1. ImageNet top-1 accuracy of state-of-the-art CNN-based and Transformer-based models. All the results are obtained based on the best test resolutions, without using any extra training data. Our VOLO-D5 achieves the best accuracy, outperforming the latest NFNet-F6 w/ SAM [2, 15] and CaiT-M48 w/ KD [22, 69], while using much less training parameters. To our best knowledge, VOLO-D5 is the first model exceeding 87% top-1 accuracy on ImageNet.

In this work we try to close such performance gap. We find one major factor limiting ViTs from outperforming CNNs is their low efficacy in encoding fine-level features and contexts into token representations, which are critical for achieving compelling visual recognition performance. Fine-level information can be encoded into tokens by finer-grained image tokenization, which however would lead to a token sequence of greater length that increases quadratically the complexity of the self-attention mechanism of ViTs.

*Equal contribution.

FaceGuard: Proactive Deepfake Detection

Yuankun Yang[1]*, Chenyue Liang[2]*, Hongyu He[3], Xiaoyu Cao[3], Neil Zhenqiang Gong[3]

[1]Fudan University, 17307110068@fudan.edu.cn
[2]Chinese Academy of Sciences, lllcy_cheryl@outlook.com
[3]Duke University, {hongyu.he, xiaoyu.cao, neil.gong}@duke.edu

Abstract

Existing deepfake-detection methods focus on *passive* detection, i.e., they detect fake face images via exploiting the artifacts produced during deepfake manipulation. A key limitation of passive detection is that it cannot detect fake faces that are generated by new deepfake generation methods. In this work, we propose *FaceGuard*, a *proactive* deepfake-detection framework. FaceGuard embeds a watermark into a real face image before it is published on social media. Given a face image that claims to be an individual (e.g., Nicolas Cage), FaceGuard extracts a watermark from it and predicts the face image to be fake if the extracted watermark does not match well with the individual's ground truth one. A key component of FaceGuard is a new deep-learning-based watermarking method, which is 1) robust to normal image post-processing such as JPEG compression, Gaussian blurring, cropping, and resizing, but 2) fragile to deepfake manipulation. Our evaluation on multiple datasets shows that FaceGuard can detect deepfakes accurately and outperforms existing methods.

1 Introduction

As deep learning becomes more and more powerful, deep learning based *deepfake generation methods* can produce more and more realistic-looking deepfakes [8, 18, 19, 20, 30, 35, 41, 42, 51, 56]. In this work, we focus on fake faces because faces are key ingredients in human communications. Moreover, we focus on *manipulated* fake faces, in which a deepfake generation method replaces a target face as a source face (known as *face replacement*) or changes the facial expressions of a target face as those of a source face (known as *face reenactment*). For instance, in the well-known Trump-Cage deepfakes example [34], Trump's face (target face) is replaced as Cage's face (source face). Fake faces can be used to assist the propagation of fake news, rumors, and disinformation on social media (e.g., Facebook, Twitter, and Instagram). Therefore, fake faces pose growing concerns to the integrity of online information, highlighting the urgent needs for deepfake detection.

Existing deepfake detection mainly focuses on *passive* detection, which exploits the artifacts in fake faces to detect them after they have been generated. Specifically, given a face image, a passive detector extracts various features from it and classifies it to be real or fake based on the features. The features can be manually designed based on some heuristics [2, 14, 22, 23, 27, 50] or automatically extracted by a deep neural network based feature extractor [1, 6, 11, 14, 28, 29, 37, 38, 48, 54]. Passive detection faces a key limitation [7], i.e., it cannot detect fake faces that are generated by new deepfake generation methods that were not considered when training the passive detector. As new deepfake generation methods are continuously developed, this limitation poses significant challenges to passive deepfake detection.

Our work: In this work, we propose *FaceGuard*, a *proactive* deepfake-detection framework. FaceGuard addresses the limitation of passive detection via proactively embedding watermarks into real face images before they are manipulated by deepfake generation methods. Figure 1 illustrates the difference between passive detection and FaceGuard. Specifically, before posting an individual's real face image on social media, FaceGuard embeds a watermark (i.e., a binary vector in our work) into it. The watermark is human imperceptible, i.e., a face image and its watermarked version look visually the same to human eyes. For instance, the watermark can be embedded into an individual's face image using the individual's smartphone. Suppose a face image is claimed to be an individual, e.g., the manipulated

*The first two authors made equal contributions. They performed this research when they were remote interns in Gong's group.

Theoretical Linear Convergence of Unfolded ISTA and its Practical Weights and Thresholds

Xiaohan Chen[*]
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA
chernxh@tamu.edu

Jialin Liu[*]
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095, USA
liujl11@math.ucla.edu

Zhangyang Wang
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA
atlaswang@tamu.edu

Wotao Yin
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095, USA
wotaoyin@math.ucla.edu

Abstract

In recent years, unfolding iterative algorithms as neural networks has become an empirical success in solving sparse recovery problems. However, its theoretical understanding is still immature, which prevents us from fully utilizing the power of neural networks. In this work, we study unfolded ISTA (Iterative Shrinkage Thresholding Algorithm) for sparse signal recovery. We introduce a weight structure that is necessary for asymptotic convergence to the true sparse signal. With this structure, unfolded ISTA can attain a linear convergence, which is better than the sublinear convergence of ISTA/FISTA in general cases. Furthermore, we propose to incorporate thresholding in the network to perform support selection, which is easy to implement and able to boost the convergence rate both theoretically and empirically. Extensive simulations, including sparse vector recovery and a compressive sensing experiment on real image data, corroborate our theoretical results and demonstrate their practical usefulness. We have made our codes publicly available.[2].

1 Introduction

This paper aims to recover a sparse vector x^* from its noisy linear measurements:

$$b = Ax^* + \varepsilon, \tag{1}$$

where $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $\varepsilon \in \mathbb{R}^m$ is additive Gaussian white noise, and we have $m \ll n$. (1) is an ill-posed, highly under-determined system. However, it becomes easier to solve if x^* is assumed to be sparse, i.e. the cardinality of support of x^*, $S = \{i|x_i^* \neq 0\}$, is small compared to n.

A popular approach is to model the problem as the LASSO formulation (λ is a scalar):

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\|b - Ax\|_2^2 + \lambda\|x\|_1 \tag{2}$$

and solve it using iterative algorithms such as the iterative shrinkage thresholding algorithm (ISTA) [1]:

$$x^{k+1} = \eta_{\lambda/L}\Big(x^k + \frac{1}{L}A^T(b - Ax^k)\Big), \quad k = 0, 1, 2, \ldots \tag{3}$$

[*]These authors contributed equally and are listed alphabetically.

[2]https://github.com/xchen-tamu/linear-lista-cpss

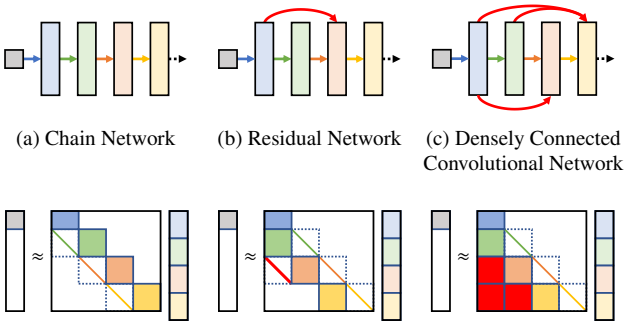
32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Dataless Model Selection with the Deep Frame Potential

Calvin Murdock[1] Simon Lucey[1,2]
[1]Carnegie Mellon University [2]Argo AI
{cmurdock,slucey}@cs.cmu.edu

Abstract

Choosing a deep neural network architecture is a fundamental problem in applications that require balancing performance and parameter efficiency. Standard approaches rely on ad-hoc engineering or computationally expensive validation on a specific dataset. We instead attempt to quantify networks by their intrinsic capacity for unique and robust representations, enabling efficient architecture comparisons without requiring any data. Building upon theoretical connections between deep learning and sparse approximation, we propose the deep frame potential: a measure of coherence that is approximately related to representation stability but has minimizers that depend only on network structure. This provides a framework for jointly quantifying the contributions of architectural hyper-parameters such as depth, width, and skip connections. We validate its use as a criterion for model selection and demonstrate correlation with generalization error on a variety of common residual and densely connected network architectures.



(a) Chain Network (b) Residual Network (c) Densely Connected Convolutional Network

(d) Induced Dictionary Structures for Sparse Approximation

Figure 1: Why are some deep neural network architectures better than others? In comparison to (a) standard chain connections, skip connections like those in (b) ResNets [13] and (c) DenseNets [15] have demonstrated significant improvements in training effectiveness, parameter efficiency, and generalization performance. We provide one possible explanation for this phenomenon by approximating network activations as (d) solutions to sparse approximation problems with different induced dictionary structures.

1. Introduction

Deep neural networks have dominated nearly every benchmark within the field of computer vision. While this modern influx of deep learning originally began with the task of large-scale image recognition [18], new datasets, loss functions, and network configurations have quickly expanded its scope to include a much wider range of applications. Despite this, the underlying architectures used to learn effective image representations are generally consistent across all of them. This can be seen through the community's quick adoption of the newest state-of-the-art deep networks from AlexNet [18] to VGGNet [28], ResNets [13], DenseNets [15], and so on. But this begs the question: why do some deep network architectures work better than others? Despite years of groundbreaking empirical results, an answer to this question still remains elusive.

Fundamentally, the difficulty in comparing network architectures arises from the lack of a theoretical foundation for characterizing their generalization capacities. Shallow machine learning techniques like support vector machines [6] were aided by theoretical tools like the VC-dimension [31] for determining when their predictions could be trusted to avoid overfitting. Deep neural networks, on the other hand, have eschewed similar analyses due to their complexity. Theoretical explorations of deep network generalization [24] are often disconnected from practical applications and rarely provide actionable insight into how architectural hyper-parameters contribute to performance.

Building upon recent connections between deep learning and sparse approximation [26, 23], we instead interpret feed-forward deep networks as algorithms for approximate inference in related sparse coding problems. These problems aim to optimally reconstruct zero-padded input images as sparse, nonnegative linear combinations of atoms from architecture-dependent dictionaries, as shown in Fig. 1. We propose to indirectly analyze practical deep network architectures with complicated skip connections, like residual networks (ResNets) [13] and densely connected convolu-

Fast Fourier Convolution

Lu Chi[1], **Borui Jiang**[2], **Yadong Mu**[1*]
[1]Wangxuan Institute of Computer Technology, [2]Center for Data Science
Peking University
{chilu,jbr,myd}@pku.edu.cn

Abstract

Vanilla convolutions in modern deep networks are known to operate locally and at fixed scale (*e.g.*, the widely-adopted 3×3 kernels in image-oriented tasks). This causes low efficacy in connecting two distant locations in the network. In this work, we propose a novel convolutional operator dubbed as *fast Fourier convolution* (FFC), which has the main hallmarks of non-local receptive fields and cross-scale fusion within the convolutional unit. According to spectral convolution theorem in Fourier theory, point-wise update in the spectral domain globally affects all input features involved in Fourier transform, which sheds light on neural architectural design with non-local receptive field. Our proposed FFC is inspired to capsulate three different kinds of computations in a single operation unit: a local branch that conducts ordinary small-kernel convolution, a semi-global branch that processes spectrally stacked image patches, and a global branch that manipulates image-level spectrum. All branches complementarily address different scales. A multi-branch aggregation step is included in FFC for cross-scale fusion. FFC is a generic operator that can directly replace vanilla convolutions in a large body of existing networks, without any adjustments and with comparable complexity metrics (*e.g.*, FLOPs). We experimentally evaluate FFC in three major vision benchmarks (ImageNet for image recognition, Kinetics for video action recognition, MSCOCO for human keypoint detection). It consistently elevates accuracies in all above tasks by significant margins.

1 Introduction

Deep neural networks have been the prominent driving force for recent dramatic progress in several research domains. The goal of this paper is the exposition of a novel convolutional unit codenamed *fast Fourier convolution* (FFC). Motivating our design of FFC, we consider two desiderata. First, one of the core concepts in deep convolutional neural networks (CNNs) is *receptive field* that is deeply rooted in the visual cortex architecture. In convolutional networks, receptive field refers to the image part that is accessible by one filter. A majority of modern networks have adopted the architecture of deeply stacking many convolutions with small receptive field (3×3 in ResNet [11] for images or $3 \times 3 \times 3$ in C3D [27] for videos). This still ensures that all image parts are visible to high layers, since stacking convolutional layers can increase the receptive field either linearly or exponentially (*e.g.*, using atrous convolutions [2]). However, for context-sensitive tasks such as human pose estimation, large receptive field in convolutions is highly desired. Recent endeavor on enlarging receptive field includes deformable convolution [9] and non-local neural networks [31].

Secondly, CNNs typically admit a chain-like topology. Neural layers provide different levels of feature abstraction. The idea of cross-scale fusion has celebrated its success in various scenarios. For example, one can tailor and send high-level semantics to shallower layers for guiding more accurate spatial detection, as shown in the seminal work of FPN [18]. Recent studies have considered

*Corresponding author.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.