Academic Abstract Analysis & Abstract Revision

学术摘要分析与摘要修订

Qingxin Meng

孟 庆 鑫

December 27, 2018

1 Analysis of 20 Most Influential Papers in Machine Learning

摘要 [1], [2] 来自 DeepMind, 第一作者是 AlphaGo 之父 David Silver.

1 A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver[1,2], **Thomas Hubert**[1], **Julian Schrittwieser**[1], **Ioannis Antonoglou**[1], **Matthew Lai**[1], **Arthur Guez**[1], **Marc Lanctot**[1], **Laurent Sifre**[1], **Dharshan Kumaran**[1], **Thore Graepel**[1], **Timothy Lillicrap**[1], **Karen Simonyan1**[1], **Demis Hassabis**[1]

[1]DeepMind, 6 Pancras Square, London N1C 4AG, UK.

[2]University College London, Gower Street, London WC1E 6BT, UK.

Science 07 Dec 2018: Vol. 362, Issue 6419, pp. 1140-1144 | *DOI: 10.1126/science.aar6404*

Abstract — The game of chess is the longest-studied domain in the history of artificial intelligence. The strongest programs are based on a combination of sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades. By contrast, the AlphaGo Zero program recently achieved superhuman performance in the game of Go by reinforcement learning from self-play.[B] ◀ In this paper, we generalize this approach into a single AlphaZero algorithm that can achieve superhuman performance in many challenging games.[I] ◀ Starting from random play and given no domain knowledge except the game rules, AlphaZero convincingly defeated a world champion program in the games of chess and shogi (Japanese chess), as well as Go.[R] ◀

← B
← I
← R

2 Mastering the game of Go without human knowledge

David Silver✉, **Julian Schrittwieser**, **Karen Simonyan**, **Ioannis Antonoglou**, **Aja Huang**, **Arthur Guez**, **Thomas Hubert**, **Lucas Baker**, **Matthew Lai**, **Adrian Bolton**, **Yutian Chen**, **Timothy Lillicrap**, **Fan Hui**, **Laurent Sifre**, **George van den Driessche**, **Thore Graepel** & **Demis Hassabis**

DeepMind, 5 New Street Square, London EC4A 3TW, UK.

Nature volume 550, pages 354–359 (19 October 2017) | *doi:10.1038/nature24270*

Abstract — A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play.[B] ◀ Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules.[I] ◀ AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo' s own move selections and also the winner of AlphaGo' s games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration.[M] ◀ Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.[R] ◀

← B
← I
← M
← R

摘要 [3] 来自 Google AI, BERT 模型在 NLP 领域取得重大突破, 论文以预印本发布在 arXiv.org.

3 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
Submitted on 11 Oct 2018 | *arXiv:1810.04805v1* [cs.CL]

Abstract — We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.[I] ◄ Unlike recent language representation models (Peters et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers.[M] ◄ As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, *without* substantial task-specific architecture modifications.[C] ◄ ← I ← M ← C
BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE benchmark to **80.4%** (**7.6%** absolute improvement), MultiNLI accuracy to **86.7%** (**5.6%** absolute improvement) and the SQuAD v1.1 question answering Test F1 to **93.2%** (**1.5%** absolute improvement), outperforming human performance by **2.0%**.[R] ◄ ← R

摘要 [4] 是两位华裔学者在政策预测领域的最新成果.

4 Reading China: Predicting policy change with machine learning

Julian TszKin Chan **Weifeng Zhong**
Bates White Economic Consulting American Enterprise Institute
Submitted on 22 Oct 2018, last revised 15 Nov 2018 | *AEI Economics Working Paper Series*
Project website at `https://policychangeindex.com/`

Abstract — For the first time in the literature, we develop a quantitative indicator of the Chinese government' s policy priorities over a long period of time, which we call the Policy Change Index (PCI) for China. The PCI is a leading indicator of policy changes that covers the period from 1951 to the third quarter of 2018, and it can be updated in the future.[I] ◄ It is designed with two building blocks: the full text of the *People' s Daily* —the official newspaper of the Communist Party of China —as input data and a set of machine learning techniques to detect changes in how this newspaper prioritizes policy issues.[M] ◄ Due to the unique role of the *People' s Daily* in China' s propaganda system, detecting changes in this newspaper allows us to predict changes in China' s policies.[R] ◄ The construction of the PCI does not require the understanding of the Chinese text, which suggests a wide range of applications in other settings, such as predicting changes in other (ex-)Communist regimes' policies, measuring decentralization in central-local government relations, quantifying media bias in democratic countries, and predicting changes in lawmakers' voting behavior and in judges' ideological leaning.[C] ◄ ← I ← M ← R ← C

摘要 [5] 来自 Stanford ML Group, Andrew Y. Ng 的最新研究, 论文以预印本发布在 arXiv.org.

5 Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks

Pranav Rajpurkar* **Awni Y. Hannun*** **Masoumeh Haghpanahi*** **Codie Bourn*** **Andrew Y. Ng***
*Stanford University *iRhythm Technologies
Submitted on 6 Jul 2017 | *arXiv:1707.01836v1* [cs.CV]
Stanford Machine Learning Group | *Project website at `https://stanfordmlgroup.github.io/projects/ecg/`*

Abstract — We develop an algorithm which exceeds the performance of board certified cardiologists in detecting a wide range of heart arrhythmias from electrocardiograms recorded with a single-lead wearable monitor.[I] ◄ We build a dataset with more than 500 times the number of unique patients than previously studied corpora. On this dataset, we train a 34-layer convolutional neural network which maps a sequence ← I

of ECG samples to a sequence of rhythm classes. Committees of board-certified cardiologists annotate a gold standard test set on which we compare the performance of our model to that of 6 other individual cardiologists.[M] ◀ We exceed the average cardiologist performance in both recall (sensitivity) and precision ← M
(positive predictive value).[R] ◀ ← R

摘要 [6-14] 是来自 AI-Machine Learning 领域顶级会议 NeurIPS 在 2017, 2018 年的文章, 其中 7 篇为年度 Best paper; 另外的两篇: [10] 来自 Toyota Technological Institute, [11] 来自 NVIDIA 和 MIT, 均为产业界有重大影响的文章.

6 Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes

Hassan Ashtiani
Department of Computing and Software
McMaster University, and
Vector Institute, ON, Canada

Shai Ben-David
School of Computer Science
University of Waterloo
Waterloo, ON, Canada

Nicholas J. A. Harvey
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Christopher Liaw
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Abbas Mehrabian
School of Computer Science
McGill University
Montréal, QC, Canada

Yaniv Plan
Department of Mathematics
University of British Columbia
Vancouver, BC, Canada

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

Abstract — We prove that $\tilde{\Theta}(kd^2/\varepsilon^2)$ samples are necessary and sufficient for learning a mixture of k Gaussians in $\mathbb{R}^d$, up to error ε in total variation distance. This improves both the known upper bounds and lower bounds for this problem. For mixtures of axis-aligned Gaussians, we show that $\tilde{O}(kd/\varepsilon^2)$ samples suffice, matching a known lower bound.

The upper bound is based on a novel technique for distribution learning based on a notion of *sample compression*. Any class of distributions that allows such a sample compression scheme can also be learned with few samples. Moreover, if a class of distributions has such a compression scheme, then so do the classes of *products* and *mixtures* of those distributions.[I] ◀ The core of our main result is showing that the ← I
class of Gaussians in $\mathbb{R}^d$ has an efficient sample compression.[R] ◀ ← R

7 Neural Ordinary Differential Equations

Ricky T. Q. Chen* **Yulia Rubanova*** **Jesse Bettencourt*** **David Duvenaud**
University of Toronto, Vector Institute, Toronto, Canada.

Submitted on 19 Jun 2018, last revised 22 Oct 2018 | *arXiv:1806.07366v3* [cs.LG]
32nd Conference on Neural Information Processing Systems (NeurIPS 2018 Best Paper), Montréal, Canada.

Abstract — We introduce a new family of deep neural network models. Instead of specifying a discrete sequence of hidden layers, we parameterize the derivative of the hidden state using a neural network. The output of the network is computed using a blackbox differential equation solver. These continuous-depth models have constant memory cost, adapt their evaluation strategy to each input, and can explicitly trade numerical precision for speed. We demonstrate these properties in continuous-depth residual networks and continuous-time latent variable models. We also construct continuous normalizing flows, a generative model that can train by maximum likelihood, without partitioning or ordering the data dimensions. For training, we show how to scalably backpropagate through any ODE solver, without access to its internal operations. This allows end-to-end training of ODEs within larger models.[I] ◀ ← I

8 Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

Kevin Scaman **Francis Bach** **Sébastien Bubeck** **Yin Tat Lee** **Laurent Massoulié**

Submitted on 1 Jun 2018 | *arXiv:1806.00291v1* [math.OC]
32nd Conference on Neural Information Processing Systems (NeurIPS 2018 Best Paper), Montréal, Canada.

Abstract — In this work, we consider the distributed optimization of non-smooth convex functions using a network of computing units. We investigate this problem under two regularity assumptions: (1) the

Lipschitz continuity of the *global* objective function, and (2) the Lipschitz continuity of *local* individual functions. Under the *local regularity* assumption, we provide the first optimal first-order decentralized algorithm called *multi-step primal-dual* (MSPD) and its corresponding optimal convergence rate. A notable aspect of this result is that, for non-smooth functions, while the dominant term of the error is in $O(1/\sqrt{t})$, the structure of the communication network only impacts a second-order term in $O(1/t)$, where t is time. In other words, the error due to limits in communication resources decreases at a fast rate even in the case of non-strongly-convex objective functions. Under the *global regularity* assumption, we provide a simple yet efficient algorithm called *distributed randomized smoothing* (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.[I] ◄ ← I

9 Non-delusional Q-learning and value-iteration

Tyler Lu **Dale Schuurmans** **Craig Boutilier**
Google AI
32nd Conference on Neural Information Processing Systems (NeurIPS 2018 Best Paper), Montréal, Canada.

Abstract — We identify a fundamental source of error in Q-learning and other forms of dynamic programming with function approximation. Delusional bias arises when the approximation architecture limits the class of expressible greedy policies.[I] ◄ Since standard Q-updates make globally uncoordinated ← I action choices with respect to the expressible policy class, inconsistent or even conflicting Q-value estimates can result, leading to pathological behaviour such as over/under-estimation, instability and even divergence.[B] ◄ To solve this problem, we introduce a new notion of policy consistency and define a lo- ← B cal backup process that ensures global consistency through the use of information sets—sets that record constraints on policies consistent with backed-up Q-values. We prove that both the model-based and model-free algorithms using this backup remove delusional bias, yielding the first known algorithms that guarantee optimal results under general conditions. These algorithms furthermore only require polynomially many information sets (from a potentially exponential support). Finally, we suggest other practical heuristics for value-iteration and Q-learning that attempt to reduce delusional bias.[I] ◄ ← I

10 Exploring Generalization in Deep Learning

Behnam Neyshabur **Srinadh Bhojanapalli** **David Mcallester** **Nati Srebro**
Toyota Technological Institute at Chicago
Submitted on 16 Oct 2017, last revised 22 Feb 2018 | *arXiv:1710.05468v3* [stat.ML]
32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Abstract — With a goal of understanding what drives generalization in deep networks, we consider several recently suggested explanations, including norm-based control, sharpness and robustness. We study how these measures can ensure generalization, highlighting the importance of scale normalization, and making a connection between sharpness and PAC-Bayes theory. We then investigate how well the measures explain different observed phenomena.[I] ◄ ← I

11 Video-to-Video Synthesis

Ting-Chun Wang[*] **Ming-Yu Liu**[*] **Jun-Yan Zhu**[⋆] **Nikolai Yakovenko**[*]
Andrew Tao[*] **Jan Kautz**[*] **Bryan Catanzaro**[*]
[*]NVIDIA [⋆]MIT CSAIL
Submitted on 20 Aug 2018, last revised 3 Dec 2018 | *arXiv:1808.06601v2* [cs.CV]
32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Abstract — We study the problem of video-to-video synthesis, whose goal is to learn a mapping function from an input source video (e.g., a sequence of semantic segmentation masks) to an output photorealistic video that precisely depicts the content of the source video. While its image counterpart, the image-to-image translation problem, is a popular topic, the video-to-video synthesis problem is less explored in the literature. Without modeling temporal dynamics, directly applying existing image synthesis approaches

to an input video often results in temporally incoherent videos of low visual quality.[B] ◀ In this paper, we ← B
propose a video-to-video synthesis approach under the generative adversarial learning framework.[I] ◀ ← I
Through carefully-designed generators and discriminators, coupled with a spatio-temporal adversarial
objective, we achieve high-resolution, photorealistic, temporally coherent video results on a diverse set of
input formats including segmentation masks, sketches, and poses.[M] ◀ Experiments on multiple bench- ← M
marks show the advantage of our method compared to strong baselines. In particular, our model is capable
of synthesizing 2K resolution videos of street scenes up to 30 seconds long, which significantly advances
the state-of-the-art of video synthesis.[R] ◀ Finally, we apply our method to future video prediction, out- ← R
performing several competing systems.[C] ◀ Code, models, and more results are available at our `website`. ← C

12 Safe and Nested Subgame Solving for Imperfect-Information Games

Noam Brown **Tuomas Sandholm**
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15217
Submitted on 8 May 2017, last revised 16 Nov 2017 | *arXiv:1705.02955v3* [cs.AI]
31st Conference on Neural Information Processing Systems (NIPS 2017 Best Paper), Long Beach, CA, USA.

Abstract — In imperfect-information games, the optimal strategy in a subgame may depend on the strat-
egy in other, unreached subgames. Thus a subgame cannot be solved in isolation and must instead con-
sider the strategy for the entire game as a whole, unlike perfect-information games. Nevertheless, it is
possible to first approximate a solution for the whole game and then improve it in individual subgames.
This is referred to as *subgame solving*.[B] ◀ We introduce subgame-solving techniques that outperform prior ← B
methods both in theory and practice. We also show how to adapt them, and past subgame-solving tech-
niques, to respond to opponent actions that are outside the original action abstraction; this significantly
outperforms the prior state-of-the-art approach, action translation. Finally, we show that subgame solv-
ing can be repeated as the game progresses down the game tree, leading to far lower exploitability. These
techniques were a key component of *Libratus*, the first AI to defeat top humans in heads-up no-limit Texas
hold'em poker.[I] ◀ ← I

13 A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum Gatsby Unit, UCL
Wenkai Xu Gatsby Unit, UCL
Zoltán Szabó CMAP, École Polytechnique
Kenji Fukumizu The Institute of Statistical Mathematics
Arthur Gretton Gatsby Unit, UCL
Submitted on 22 May 2017, last revised 24 Oct 2017 | *arXiv:1705.07673v2* [stat.ML]
31st Conference on Neural Information Processing Systems (NIPS 2017 Best Paper), Long Beach, CA, USA.

Abstract — We propose a novel adaptive test of goodness-of-fit, with computational cost linear in the
number of samples.[I] ◀ We learn the test features that best indicate the differences between observed ← I
samples and a reference model, by minimizing the false negative rate. These features are constructed via
Stein's method, meaning that it is not necessary to compute the normalising constant of the model.[M] ◀ We ← M
analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative,
our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice
of parameters for that test.[R] ◀ In experiments, the performance of our method exceeds that of the earlier ← R
linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and
where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time
two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.[C] ◀ ← C

14 Variance-based Regularization with Convex Objectives

Hongseok Namkoong **John C. Duchi**
Stanford University
Submitted on 8 Oct 2016, last revised 14 Dec 2017 | *arXiv:1610.02581v3* [stat.ML]
31st Conference on Neural Information Processing Systems (NIPS 2017 Best Paper), Long Beach, CA, USA.

Abstract — We develop an approach to risk minimization and stochastic optimization that provides a
convex surrogate for variance, allowing near-optimal and computationally efficient trading between ap-
proximation and estimation error.[I] ◄ Our approach builds off of techniques for distributionally robust ← I
optimization and Owen's empirical likelihood, and we provide a number of finite-sample and asymptotic
results characterizing the theoretical performance of the estimator. In particular, we show that our proce-
dure comes with certificates of optimality, achieving (in some scenarios) faster rates of convergence than
empirical risk minimization by virtue of automatically balancing bias and variance.[M] ◄ We give corrob- ← M
orating empirical evidence showing that in practice, the estimator indeed trades between variance and
absolute performance on a training sample, improving out-of-sample (test) performance over standard
empirical risk minimization for a number of classification problems.[R] ◄ ← R

摘要 [15] 是计算机视觉领域最有影响力的文章之一, 来自计算机视觉领域 Top 1 会议 CVPR.

15 Deep Neural Networks are Easily Fooled:
High Confidence Predictions for Unrecognizable Images

Anh Nguyen **Jason Yosinski** **Jeff Clune**
University of Wyoming Cornell University University of Wyoming
Submitted on 5 Dec 2014, last revised 2 Apr 2015 | *arXiv:1412.1897v4* [cs.CV]
Date Added to IEEE *Xplore*: 15 October 2015 | *DOI: 10.1109/CVPR.2015.7298640*
2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ©IEEE

Abstract — Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on
a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are
now able to classify objects in images with near-human-level performance, questions naturally arise as
to what differences remain between computer and human vision. A recent study[30] revealed that chang-
ing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as
something else entirely (e.g. mislabeling a lion a library).[B] ◄ Here we show a related result: it is easy ← B
to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe
to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static
is a lion).[I] ◄ Specifically, we take convolutional neural networks trained to perform well on either the ← I
ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that
DNNs label with high confidence as belonging to each dataset class.[M] ◄ It is possible to produce images ← M
totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which
we call "fooling images" (more generally, fooling examples).[R] ◄ Our results shed light on interesting ← R
differences between human vision and current DNNs, and raise questions about the generality of DNN
computer vision.[C] ◄ ← C

摘要 [16-20] 来自深度学习/神经网络方向具有重要影响力的 5 篇文章.

16 Dance Dance Convolution

Chris Donahue[*] **Zachary C. Lipton**[⋆] **Julian McAuley**[⋆]
[*]UCSD Department of Music, San Diego, CA [⋆]UCSD Department of Computer Science, San Diego, CA
Submitted on 20 Mar 2017, last revised 21 Jun 2017 | *arXiv:1703.06891v3* [cs.LG]
Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, PMLR 70, 2017.

Abstract — Dance Dance Revolution (DDR) is a popular rhythm-based video game. Players perform steps
on a dance platform in synchronization with music as directed by on-screen step charts. While many step
charts are available in standardized packs, players may grow tired of existing charts, or wish to dance
to a song for which no chart exists.[B] ◄ We introduce the task of learning to choreograph. Given a raw ← B
audio track, the goal is to produce a new step chart.[I] ◄ This task decomposes naturally into two subtasks: ← I
deciding when to place steps and deciding which steps to select. For the step placement task, we combine
recurrent and convolutional neural networks to ingest spectrograms of low-level audio features to predict
steps, conditioned on chart difficulty. For step selection, we present a conditional LSTM generative model
that substantially outperforms n-gram and fixed-window approaches.[M] ◄ ← M

17 Don't Decay the Learning Rate, Increase the Batch Size

Samuel L. Smith **Pieter-Jan Kindermans** **Chris Ying** **Quoc V. Le**
Google Brain

Submitted on 1 Nov 2017, last revised 24 Feb 2018 | *arXiv:1711.00489v2* [cs.LG]
6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada.

Abstract — It is common practice to decay the learning rate.[B] ◄ Here we show one can usually obtain ← B
the same learning curve on both training and test sets by instead increasing the batch size during training. This procedure is successful for stochastic gradient descent (SGD), SGD with momentum, Nesterov momentum, and Adam. It reaches equivalent test accuracies after the same number of training epochs, but with fewer parameter updates, leading to greater parallelism and shorter training times.[I] ◄ We can ← I
further reduce the number of parameter updates by increasing the learning rate ϵ and scaling the batch size $B \propto \epsilon$. Finally, one can increase the momentum coefficient m and scale $B \propto 1/(1-m)$, although this tends to slightly reduce the test accuracy.[M] ◄ Crucially, our techniques allow us to repurpose exist- ← M
ing training schedules for large batch training with no hyper-parameter tuning. We train ResNet-50 on ImageNet to 76.1% validation accuracy in under 30 minutes.[R] ◄ ← R

18 Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Alec Radford indico Research, Boston, MA
Luke Metz indico Research, Boston, MA
Soumith Chintala Facebook AI Research, New York, NY

Submitted on 19 Nov 2015, last revised 7 Jan 2016 | *arXiv:1511.06434v2* [cs.LG]
4th International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico.

Abstract — In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention.[B] ◄ In this work we hope to help bridge the gap between the success of CNNs for supervised ← B
learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning.[I] ◄ Training on various image datasets, we show con- ← I
vincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.[R] ◄ ← R

19 U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger✉ **Philipp Fischer** **Thomas Brox**
Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany

Submitted on 18 May 2015 | *arXiv:1505.04597v1* [cs.CV]
18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Germany.

Abstract — There is large consent that successful training of deep networks requires many thousand annotated training samples.[B] ◄ In this paper, we present a network and training strategy that relies on the ← B
strong use of data augmentation to use the available annotated samples more efficiently.[I] ◄ The archi- ← I
tecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization.[M] ◄ We show that such a network can be trained end-to-end from very few images ← M
and outperforms the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Using the same network trained on transmitted light microscopy images (phase contrast and DIC) we won the ISBI cell tracking challenge 2015 in these categories by a large margin. Moreover, the network is fast. Segmentation of a 512×512 image takes less than a second on a recent GPU.[R] ◄ The full implementation (based on Caffe) and the trained ← R
networks are available at `http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net`.

20 Deep Learning: A Bayesian Perspective

Nicholas Polson
Booth School of Business,
University of Chicago

Vadim Sokolov
Volgenau School of Engineering,
George Mason University

Submitted on 1 Jun 2017, last revised 14 Nov 2017 | *arXiv:1706.00473v4* [stat.ML]
Bayesian Anal. Volume 12, Number 4 (2017), 1275-1304. | *DOI: 10.1214/17-BA1082*

Abstract — Deep learning is a form of machine learning for nonlinear high dimensional pattern matching and prediction.[B] ◀ By taking a Bayesian probabilistic perspective, we provide a number of insights into more efficient algorithms for optimisation and hyper-parameter tuning.[I] ◀ Traditional high-dimensional data reduction techniques, such as principal component analysis (PCA), partial least squares (PLS), reduced rank regression (RRR), projection pursuit regression (PPR) are all shown to be shallow learners. Their deep learning counterparts exploit multiple deep layers of data reduction which provide predictive performance gains. Stochastic gradient descent (SGD) training optimisation and Dropout (DO) regularization provide estimation and variable selection. Bayesian regularization is central to finding weights and connections in networks to optimize the predictive bias-variance trade-off. To illustrate our methodology, we provide an analysis of international bookings on Airbnb.[R] ◀ Finally, we conclude with directions for future research.[C] ◀

← B
← I
← R
← C

20 篇摘要的 BIMRC 结构统计表

Title	**Publication**	**B**	**I**	**M**	**R**	**C**
A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play	Science	√	√		√	
Mastering the game of Go without human knowledge	Nature	√	√	√	√	
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	arXiv		√	√	√	√
Reading China: Predicting policy change with machine learning	AEI		√	√	√	√
Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks	arXiv		√	√	√	
Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes	NeurIPS⋆		√		√	
Neural Ordinary Differential Equations	NeurIPS⋆		√			
Optimal Algorithms for Non-Smooth Distributed Optimization in Networks	NeurIPS⋆		√			
Non-delusional Q-learning and value-iteration	NeurIPS⋆	√	√			
Exploring Generalization in Deep Learning	NeurIPS		√			
Video-to-Video Synthesis	NeurIPS	√	√	√	√	√
Safe and Nested Subgame Solving for Imperfect-Information Games	NIPS⋆	√	√			
A Linear-Time Kernel Goodness-of-Fit Test	NIPS⋆		√	√	√	√
Variance-based Regularization with Convex Objectives	NIPS⋆		√	√	√	
Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images	CVPR	√	√	√	√	√
Dance Dance Convolution	ICML	√	√	√		
Don't Decay the Learning Rate, Increase the Batch Size	ICLR	√	√	√	√	
Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks	ICLR	√	√		√	
U-Net: Convolutional Networks for Biomedical Image Segmentation	MICCAI	√	√	√	√	
Deep Learning: A Bayesian Perspective	ISBA	√	√		√	√

⋆ 表示该文章为年度 *Best paper.*

20 篇摘要的特点总结:

- BIMRC 结构适用于有应用对象的文章, 纯粹算法的构造/证明倾向以 Introduction 为主.
- 名词化结构不多; 从句结构很多; 主动句式比较多.
- 语义逻辑最重要, 语言表达是锦上添花.

2 Abstract Revision

下面是一篇我发表在 ICDH 2014 上的文章的摘要:

A New Edge Detection Method Based on Hausdorff Distance

You Lu **Qingxin Meng** **Jiangtao Guo**
School of Computer Science and Technology, China University of Petroleum, Beijing
Date Added to IEEE *Xplore*: 29 December 2014 | *DOI: 10.1109/ICDH.2014.15*
5th International Conference on Digital Home (ICDH 2014), ©IEEE

Abstract — Most edge detection methods are based on first-order or second-order differential. These are local methods. Using Hausdorff distance to quantify the strength of the edge is a method with a holistic property. Firstly, down sample the image, and split the image into two sets. Secondly, get the feature image by assigning a value for each point using the scalar field map constructed by Hausdorff distance. Since the edge features have local properties, in this paper, we constructed a map which can get local feature images using sub image and combined them into a feature image. Finally we present a method to get the edge image by feature image. We experimentally verify the feasibility of this method.

摘要问题详细剖析:

- 第一、二句: Most edge detection methods are based on first-order or second-order differential. These are local methods.
  重点是强调局部化方法, “基于一阶、二阶微分”作为插入语, 两句话合为一句更合适; 另外“局部化方法”用 localization method.
  修改后: Most edge detection algorithms, first-order or second-order differential based, are localization methods.

- 第三句: Using Hausdorff distance to quantify the strength of the edge is a method with a holistic property.
  用 Hausdorff 距离这种做法并不是共识, 是本文提出来的方法, 用“我们介绍了...”这种句式比较合适; 边缘强度用 edge intensity, 而 the strength of the edge 说法不专业.
  修改后: Here we introduce an algorithm based on Hausdorff distance, different from localization methods, to quantify the edge intensity.

- Firstly, ... Secondly, ... Finally ... 这是中文式的描述过程的逻辑, 英文语法不对.
  这段话原本的语义是: 构造一个基于 Hausdorff 距离的映射 $H:\left(S^1,S^2\right)\longmapsto F$, 其中 S^1,S^2 是通过下采样分割原图像得到的两个图像, F 是特征图像; 再提出一种方法, 从 F 提取边缘.
  Since ... 一句要表达的是: 受下采样间隔起始点的影响, S^1,S^2 各不相同, 在构造映射 H 的时候需要“整合”各种不同的情况, 实际上就是求数学期望.
  原文很难理解, 逻辑混乱, 整段话需要重写: 逻辑结构是算法分两个部分, 第一部分是构造映射, 第二部分是提取边缘. 然后再说考虑数学期望的情况.
  即: We establish the algorithm in two stages: 1)we construct a scalar field map from two down sampled images to one featured image by Hausdorff distance; and 2)we present a method to get the edge image from featured image. Due to the dependency of interval starting point in down sample, the featured image is replaced by its expectation.

- 最后一句时态和语序都有问题.
  应改写为: We verified the feasibility of this algorithm through experiments.

修改后的摘要及摘要中文译文如下:

A New Edge Detection Method Based on Hausdorff Distance

You Lu **Qingxin Meng** **Jiangtao Guo**
School of Computer Science and Technology, China University of Petroleum, Beijing
Date Added to IEEE *Xplore*: 29 December 2014 | *DOI: 10.1109/ICDH.2014.15*
5th International Conference on Digital Home (ICDH 2014), ©IEEE

Abstract — Most edge detection algorithms, first-order or second-order differential based, are localiza-
tion methods.[B] ◀ Here we introduce an algorithm based on Hausdorff distance, different from localiza- ← B
tion methods, to quantify the edge intensity.[I] ◀ We establish the algorithm in two stages: 1)we construct ← I
a scalar field map from two down sampled images to one featured image by Hausdorff distance; and 2)we
present a method to get the edge image from featured image. Due to the dependency of interval starting
point in down sample, the featured image is replaced by its expectation.[M] ◀ We verified the feasibility of ← M
this algorithm through experiments.[R] ◀ ← R

摘要译文 — 多数边缘检测算法（基于一阶、二阶微分）是局部化方法.[B] ◀ 这里我们介绍一种基于 ← B
Hausdorff 距离的非局部化算法以量化边缘强度.[I] ◀ 我们分两个阶段建立算法: 1) 通过 Hausdorff ← I
距离构造一个从两个下采样图像到一个特征图像的标量场映射; 2) 提出了一种从特征图像中获取
边缘图像的方法. 由于下采样依赖于间隔起始点, 特征图像要替换为它的数学期望.[M] ◀ 我们通过 ← M
实验验证了该算法的可行性.[R] ◀ ← R