Interesting paper titles

Enmei Tu @ Shanghai Jiao Tong Univerity
UoW, Hamilton, NZ
Dec 2021


10

How simple!

An **embarrassingly simple** approach to zero-shot learning

ICML 2015

An embarrassingly simple approach to zero-shot learning

Bernardino Romera-Paredes BERNARDINO.ROMERAPAREDES@ENG.OX.AC.UK
University of Oxford, Department of Engineering Science, Parks Road, Oxford, OX1 3PJ, UK

Philip H. S. Torr PHILIP.TORR@ENG.OX.AC.UK
University of Oxford, Department of Engineering Science, Parks Road, Oxford, OX1 3PJ, UK

Abstract

Zero-shot learning consists in learning how to recognise new concepts by just having a description of them. Many sophisticated approaches have been proposed to address the challenges this problem comprises. In this paper we describe a zero-shot learning approach that can be implemented in just one line of code, yet it is able to outperform state of the art approaches on standard datasets. The approach is based on a more general framework which models the relationships between features, attributes, and classes as a two linear layers network, where the weights of the top layer are not learned but are given by the environment. We further provide a learning bound on the generalisation error of this kind of approaches, by casting them as domain adaptation methods. In experiments carried out on three standard real datasets, we found that our approach is able to perform significantly better than the state of art on all of them, obtaining a ratio of improvement up to 17%.

ity of the description of the categories to be distinguished makes it unfeasible to obtain training instances for each of them, e.g. when a user wants to recognise a particular model of dress.

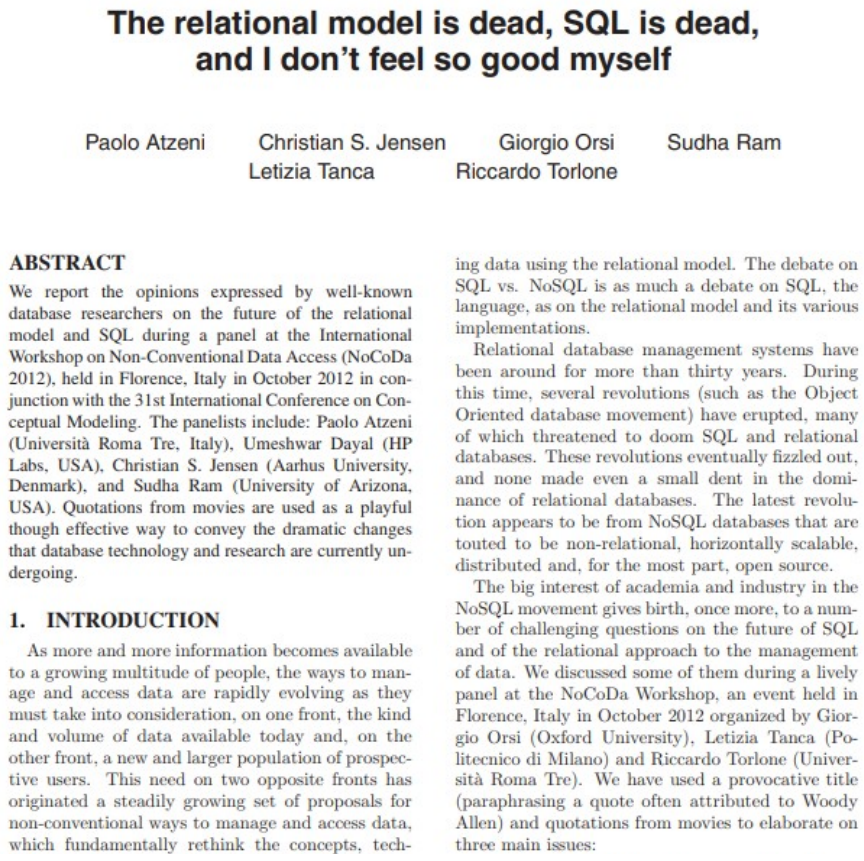
There is an increasing interest in the study of zero-shot learning (ZSL) approaches with the aim of solving this problem. ZSL consists in recognising new categories of instances without training examples, by providing a high-level description of the new categories that relate them to categories previously learned by the machine. This can be done by means of leveraging an intermediate level: the attributes that provide semantic information about the categories to classify. This paradigm is inspired by the way human beings are able to identify a new object by just reading a description of it, leveraging similarities between the description of the new object and previously learned concepts. Similarly, zero-shot learning approaches are designed to learn this intermediate semantic layer, the attributes, and apply them at inference time to predict new classes, provided with their description in terms of these attributes. Hereafter we use the term signature to refer to the attribute description of a class.


9

I don’t feel so good!

The relational model is dead, SQL is dead, and **I don’t feel so good myself**

SIGMOD 2013



The relational model is dead, SQL is dead, and I don’t feel so good myself

Paolo Atzeni Christian S. Jensen Giorgio Orsi Sudha Ram
Letizia Tanca Riccardo Torlone

ABSTRACT

We report the opinions expressed by well-known database researchers on the future of the relational model and SQL during a panel at the International Workshop on Non-Conventional Data Access (NoCoDa 2012), held in Florence, Italy in October 2012 in conjunction with the 31st International Conference on Conceptual Modeling. The panelists include: Paolo Atzeni (Università Roma Tre, Italy), Umeshwar Dayal (HP Labs, USA), Christian S. Jensen (Aarhus University, Denmark), and Sudha Ram (University of Arizona, USA). Quotations from movies are used as a playful though effective way to convey the dramatic changes that database technology and research are currently undergoing.

1. INTRODUCTION

As more and more information becomes available to a growing multitude of people, the ways to manage and access data are rapidly evolving as they must take into consideration, on one front, the kind and volume of data available today and, on the other front, a new and larger population of prospective users. This need on two opposite fronts has originated a steadily growing set of proposals for non-conventional ways to manage and access data, which fundamentally rethink the concepts, teching data using the relational model. The debate on SQL vs. NoSQL is as much a debate on SQL, the language, as on the relational model and its various implementations.

Relational database management systems have been around for more than thirty years. During this time, several revolutions (such as the Object Oriented database movement) have erupted, many of which threatened to doom SQL and relational databases. These revolutions eventually fizzled out, and none made even a small dent in the dominance of relational databases. The latest revolution appears to be from NoSQL databases that are touted to be non-relational, horizontally scalable, distributed and, for the most part, open source.

The big interest of academia and industry in the NoSQL movement gives birth, once more, to a number of challenging questions on the future of SQL and of the relational approach to the management of data. We discussed some of them during a lively panel at the NoCoDa Workshop, an event held in Florence, Italy in October 2012 organized by Giorgio Orsi (Oxford University), Letizia Tanca (Politecnico di Milano) and Riccardo Torlone (Università Roma Tre). We have used a provocative title (paraphrasing a quote often attributed to Woody Allen) and quotations from movies to elaborate on three main issues:


8

We tried, but it didn't work!

The Thing That We Tried Didn't Work Very Well: Deictic Representation in Reinforcement Learning

UAI 2002

The Thing That We Tried Didn't Work Very Well: Deictic Representation in Reinforcement Learning

Sarah Finney
AI Lab
MIT
Cambridge, MA 02139

Natalia H. Gardiol
AI Lab
MIT
Cambridge, MA 02139

Leslie Pack Kaelbling
AI Lab
MIT
Cambridge, MA 02139

Tim Oates
Dept of Computer Science
Univ. of Maryland, BC
Baltimore, MD 21250

Abstract

Most reinforcement learning methods operate on propositional representations of the world state. Such representations are often intractably large and generalize poorly. Using a deictic representation is believed to be a viable alternative: they promise generalization while allowing the use of existing reinforcement-learning methods. Yet, there are few experiments on learning with deictic representations reported in the literature. In this paper we explore the effectiveness of two forms of deictic representation and a naïve propositional representation in a simple blocks-world domain. We find, empirically, that the deictic representations actually worsen learning performance. We conclude with a discussion of possible causes of these results and strategies for more effective learning in domains with objects.

One strategy that has been successful in the planning world [8] is to *propositionalize* what is essentially a relational domain. That is, to make an attribute vector with a single Boolean attribute for each possible instance of the properties and relations in the domain. There are some fairly serious potential problems with such a representation, including the fact that it does not give much basis for generalization over objects. Additionally, the number of bits to be considered grows exponentially with the number of objects in the world, even if the task to be accomplished does not become more complicated. An alternative to this full-propositional representation is to create a *deictic*-propositional representation that, intuitively, affords more possibility for appropriate generalization.

The word deictic was introduced into the artificial intelligence vernacular by Agre and Chapman [1], who were building on Ullman's work on visual routines [15]. A deictic expression is one that "points" to something: its meaning is relative to the agent that uses it and the context in which it is used. *The-book-that-I-am-*


7

Hessian?

Learning to learn by **gradient descent by gradient descent**

NeurIP 2016

Learning to learn by gradient descent by gradient descent

Marcin Andrychowicz[1], **Misha Denil**[1], **Sergio Gómez Colmenarejo**[1], **Matthew W. Hoffman**[1], **David Pfau**[1], **Tom Schaul**[1], **Brendan Shillingford**[1,2], **Nando de Freitas**[1,2,3]

[1]Google DeepMind [2]University of Oxford [3]Canadian Institute for Advanced Research

marcin.andrychowicz@gmail.com
{mdenil,sergomez,mwhoffman,pfau,schaul}@google.com
brendan.shillingford@cs.ox.ac.uk, nandodefreitas@google.com

Abstract

The move from hand-designed features to learned features in machine learning has been wildly successful. In spite of this, optimization algorithms are still designed by hand. In this paper we show how the design of an optimization algorithm can be cast as a learning problem, allowing the algorithm to learn to exploit structure in the problems of interest in an automatic way. Our learned algorithms, implemented by LSTMs, outperform generic, hand-designed competitors on the tasks for which they are trained, and also generalize well to new tasks with similar structure. We demonstrate this on a number of tasks, including simple convex problems, training neural networks, and styling images with neural art.

Not Hessian?

Learning to Learn **without Gradient Descent by Gradient Descent**

ICML 2017

Learning to Learn without Gradient Descent by Gradient Descent

Yutian Chen [1] Matthew W. Hoffman [1] Sergio Gómez Colmenarejo [1] Misha Denil [1] Timothy P. Lillicrap [1]
Matt Botvinick [1] Nando de Freitas [1]

Abstract

We learn recurrent neural network optimizers trained on simple synthetic functions by gradient descent. We show that these learned optimizers exhibit a remarkable degree of transfer in that they can be used to efficiently optimize a broad range of derivative-free black-box functions, including Gaussian process bandits, simple control objectives, global optimization benchmarks and hyper-parameter tuning tasks. Up to the training horizon, the learned optimizers learn to trade-off exploration and exploitation, and compare favourably with heavily engineered Bayesian optimization packages for hyper-parameter tuning.

dient descent, evolutionary strategies, simulated annealing, and reinforcement learning.

For instance, one can learn to learn by gradient descent by gradient descent, or learn local Hebbian updates by gradient descent (Andrychowicz et al., 2016; Bengio et al., 1992). In the former, one uses supervised learning at the meta-level to learn an algorithm for supervised learning, while in the latter, one uses supervised learning at the meta-level to learn an algorithm for unsupervised learning.

Learning to learn can be used to learn both models and algorithms. In Andrychowicz et al. (2016) the output of meta-learning is a trained recurrent neural network (RNN), which is subsequently used as an optimization algorithm to fit other models to data. In contrast, in Zoph and Le (2017) the output of meta-learning can also be an RNN model, but


6

What happened next!

We built a fake news & click-bait filter: **What Happened Next Will Blow Your Mind!**

RANLP 2017

We Built a Fake News & Click-bait Filter: What Happened Next Will Blow Your Mind!

Georgi Karadzhov[1], Pepa Gencheva[1], Preslav Nakov[2], and Ivan Koychev[1]

[1]Sofia University "St. Kliment Ohridski", Bulgaria
[2]Qatar Computing Research Institute, HBKU, Qatar
{georgi.m.karadjov, pepa.k.gencheva}@gmail.com,
pnakov@hbku.edu.qa, koychev@uni-sofia.bg

Abstract

It is completely amazing! Fake news and click-baits have totally invaded the cyber space. Let us face it: everybody hates them for three simple reasons. Reason #2 will absolutely amaze you. What these can achieve at the time of election will completely blow your mind! Now, we all agree, this cannot go on, you know, somebody has to stop it. So, we did this research on fake news/click-bait detection and trust us, it is totally great research, it really is! Make no mistake. This is the best research ever! Seriously, come have a look, we have it all: neural networks, attention mechanism, sentiment lexicons, author profiling, you name it. Lexical features, semantic features, we absolutely have it all. And we have totally tested it, trust us! We have results, and numbers, really big numbers. The best numbers ever! Oh, and analysis, absolutely top notch analysis. Interested? Come read the shocking truth about fake news and click-bait in the Bulgarian cyber space. You won't believe what we have found!

While the motives behind these two types of fake news are different, they constitute a growing problem as they constitute a sizable fraction of the online news that users encounter on a daily basis. With the recent boom of Internet, mobile, and social networks, the spread of fake news increases exponentially. Using on-line methods for spreading harmful content makes the task of keeping the Internet clean significantly harder as it is very easy to publish an article and there is no easy way to verify its veracity. Currently, domains that consistently spread misinformation are being banned from various platforms, but this is a rather inefficient way to deal with fake news as websites that specialize in spreading misinformation are reappearing with different domain names. That is why our method is based purely on text analysis,[1] without taking into account the domain name or website's reliability as a source of information. Our work is focused on exploring various stylistic and lexical features in order to detect misleading content, and on experiments with neural network architectures in order to evaluate how deep learning can be used for detecting fake news. Moreover, we created various language-specific resources that could be used in future work on fake news and clickbait detection for Bulgarian, includ-

What happened next!

We used Neural Networks to Detect Clickbaits: **You won't believe what happened Next!**

ECIR 2017

We used Neural Networks to Detect Clickbaits: You won't believe what happened Next!

Ankesh Anand[1], Tanmoy Chakraborty[2], and Noseong Park[3]

[1] Indian Institute of Technology, Kharagpur, India
anandank@mila.quebec,
[2] University of Maryland, College Park, USA
tanchak@umiacs.umd.edu
[3] University of North Carolina, Charlotte, USA
npark2@uncc.edu

Abstract. Online content publishers often use catchy headlines for their articles in order to attract users to their websites. These headlines, popularly known as *clickbaits*, exploit a user's curiosity gap and lure them to click on links that often disappoint them. Existing methods for automatically detecting clickbaits rely on heavy feature engineering and domain knowledge. Here, we introduce a neural network architecture based on *Recurrent Neural Networks* for detecting clickbaits. Our model relies on distributed word representations learned from a large unannotated corpora, and character embeddings learned via Convolutional Neural Networks. Experimental results on a dataset of news headlines show that our model outperforms existing techniques for clickbait detection with an accuracy of 0.98 with F1-score of 0.98 and ROC-AUC of 0.99.

Keywords: Clickbait Detection, Deep Learning, Neural Networks

1 Introduction

"Clickbait" is a term used to describe a news headline which will tempt a user to follow by using provocative and catchy content. They purposely withhold the information required to understand what the content of the article is, and often exaggerate the article to create misleading expectations for the reader. Some of the example of clickbaits are:

- "The Hot New Phone Everybody Is Talking About"
- "Youll Never Believe Who Tripped and Fell on the Red Carpet"


5

BERT, speak!

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

NAACL 2018

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

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model 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.

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 score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks,

BERT, speak!

Bert has a mouth, and it must speak: Bert as a markov random field language model

NAACL 2019

BERT has a Mouth, and It Must Speak:
BERT as a Markov Random Field Language Model

Alex Wang
New York University
alexwang@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Abstract

We show that BERT (Devlin et al., 2018) is a Markov random field language model. This formulation gives way to a natural procedure to sample sentences from BERT. We generate from BERT and find that it can produce high-quality, fluent generations. Compared to the generations of a traditional left-to-right language model, BERT generates sentences that are more diverse but of slightly worse quality.

2 BERT as a Markov Random Field

Let $X = (x_1, \ldots, x_T)$ be a sequence of random variables x_i, each of which is categorical in that it can take one of M items from a vocabulary $V = \{v_1, \ldots, v_M\}$. These random variables form a fully-connected graph with undirected edges, indicating that each variable x_i is dependent on all the other variables.

Joint Distribution To define a Markov random field (MRF), we start by defining a potential over

BERT, speak!

Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords

EMNL 2021

Putting Words in BERT's Mouth:
Navigating Contextualized Vector Spaces with Pseudowords

Taelin Karidi[1] **Yichu Zhou**[2] **Nathan Schneider**[3] **Omri Abend**[1] **Vivek Srikumar**[2]
[1]Hebrew University of Jerusalem, {taelin.karidi, omri.abend}@mail.huji.ac.il
[2]University of Utah, {flyaway, svivek}@cs.utah.edu
[3]Georgetown University, nathan.schneider@georgetown.edu

Abstract

We present a method for exploring regions around individual points in a contextualized vector space (particularly, *BERT space*), as a way to investigate how these regions correspond to word senses.

By inducing a contextualized "pseudoword" as a stand-in for a static embedding in the input layer, and then performing masked prediction of a word in the sentence, we are able to investigate the geometry of the BERT-space in a controlled manner around individual instances. Using our method on a set of carefully constructed sentences targeting ambiguous English words, we find substantial regu-


4

One for all!

One ring to multiplex them all

Optical Physics 2017

One ring to multiplex them all

High-speed communication systems that use optical fibres often require hundreds of lasers. An approach that replaces these lasers with a single, ring-shaped optical device offers many technical advantages. SEE LETTER P.274

VICTOR TORRES-COMPANY

Optical-fibre communication systems form the backbone of the Internet. Current systems rely on a technology called wavelength-division multiplexing to transmit digital information at high speeds. On the transmitter side, this technology combines (multiplexes) many optical channels into a single optical fibre. Each channel uses a laser of a different frequency, and hundreds of lasers are typically needed to occupy the bandwidth

A microresonator frequency comb is an optical device that allows light of many optical frequencies to be generated in a micrometre-scale platform (Fig. 1). Tobias Kippenberg, one of the current paper's co-authors, helped to pioneer this technology about a decade ago[2]. The device essentially consists of a light source, called a pump laser, and a microresonator — a set-up also known as an optical cavity, which is used to trap light at certain 'resonance' frequencies. The frequency of the pump laser is closely tuned to a particular resonance of

One for all!

One Model To Learn Them All

Arxiv 2017, cited 276

One Model To Learn Them All

Łukasz Kaiser
Google Brain
lukaszkaiser@google.com

Aidan N. Gomez*
University of Toronto
aidan@cs.toronto.edu

Noam Shazeer
Google Brain
noam@google.com

Ashish Vaswani
Google Brain
avaswani@google.com

Niki Parmar
Google Research
nikip@google.com

Llion Jones
Google Research
llion@google.com

Jakob Uszkoreit
Google Research
usz@google.com

Abstract

Deep learning yields great results across many fields, from speech recognition, image classification, to translation. But for each problem, getting a deep model to work well involves research into the architecture and a long period of tuning. We present a single model that yields good results on a number of problems spanning multiple domains. In particular, this single model is trained concurrently on ImageNet, multiple translation tasks, image captioning (COCO dataset), a speech recognition corpus, and an English parsing task. Our model architecture incorporates building blocks from multiple domains. It contains convolutional layers, an attention mechanism, and sparsely-gated layers. Each of these computational blocks is crucial for a subset of the tasks we train on. Interestingly, even if a block is not crucial for a task, we observe that adding it never hurts performance and in most cases improves it on all tasks. We also show that tasks with less data benefit largely from joint training with other tasks, while performance on large tasks degrades only slightly if at all.

One for all!

One Big Net For Everything

Arxiv 2018

Juergen Schmidhuber – The ...

One Big Net For Everything

Technical Report

Jürgen Schmidhuber
The Swiss AI Lab, IDSIA
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Università della Svizzera italiana
Scuola universitaria professionale della Svizzera italiana
Galleria 2, 6928 Manno-Lugano, Switzerland

24 February 2018
Earlier drafts: 23 Aug, 30 Aug, 4 Sep, 31 Oct, 25 Nov, 14 Dec 2017

Abstract

I apply recent work on "learning to think" (2015) [77] and on POWERPLAY (2011) [75] to the incremental training of an increasingly general problem solver, continually learning to solve new tasks without forgetting previous skills. The problem solver is a single recurrent neural network (or similar general purpose computer) called ONE. ONE may sometimes grow or shrink, e.g., by adding or pruning neurons and connections, as proposed in 1965 [27, 26]. ONE is unusual in the sense that it is trained in various ways, e.g., by black box optimization / reinforcement learning / artificial evolution as well as supervised / unsupervised learning. For example, ONE may learn through neuroevolution to control a robot through environment-changing actions, and learn through unsupervised gradient descent to predict future inputs and vector-valued reward signals [55, 56, 60] as suggested in 1990. User-given tasks can be defined through extra goal-defining input patterns, also proposed in 1990 [79, 57, 58, 80]. Suppose ONE has already learned many skills. Now a copy of ONE can be re-trained to learn a new skill, e.g., through slow trial and error-based neu-

One for all!

One Big Net For Everything

Arxiv 2018

Is it a human brain?

Juergen Schmidhuber – The ...

One Big Net For Everything

Technical Report

Jürgen Schmidhuber
The Swiss AI Lab, IDSIA
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Università della Svizzera italiana
Scuola universitaria professionale della Svizzera italiana
Galleria 2, 6928 Manno-Lugano, Switzerland

24 February 2018
Earlier drafts: 23 Aug, 30 Aug, 4 Sep, 31 Oct, 25 Nov, 14 Dec 2017

Abstract

I apply recent work on "learning to think" (2015) [77] and on POWERPLAY (2011) [75] to the incremental training of an increasingly general problem solver, continually learning to solve new tasks without forgetting previous skills. The problem solver is a single recurrent neural network (or similar general purpose computer) called ONE. ONE may sometimes grow or shrink, e.g., by adding or pruning neurons and connections, as proposed in 1965 [27, 26]. ONE is unusual in the sense that it is trained in various ways, e.g., by black box optimization / reinforcement learning / artificial evolution as well as supervised / unsupervised learning. For example, ONE may learn through neuroevolution to control a robot through environment-changing actions, and learn through unsupervised gradient descent to predict future inputs and vector-valued reward signals [55, 56, 60] as suggested in 1990. User-given tasks can be defined through extra goal-defining input patterns, also proposed in 1990 [79, 57, 58, 80]. Suppose ONE has already learned many skills. Now a copy of ONE can be re-trained to learn a new skill, e.g., through slow trial and error-based neu-


3

Fast, Faster!

RCNN

CVPR 2014

Rich feature hierarchies for accurate object detection and semantic segmentation

Tech report (v5)

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik
UC Berkeley
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: *Regions with CNN features. We also compare R-CNN to OverFeat, a recently proposed sliding-window detector based on a similar CNN architecture. We find that R-CNN outperforms OverFeat by a large margin on the 200-class ILSVRC2013 detection dataset. Source code for the complete system is available at* http://www.cs.berkeley.edu/~rbg/rcnn.

R-CNN: *Regions with CNN features*

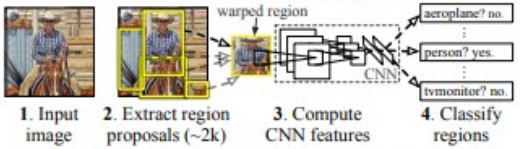


Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of **53.7% on PASCAL VOC 2010**. For comparison, [39] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%. On the 200-class **ILSVRC2013 detection dataset, R-CNN's mAP is 31.4%**, a large improvement over OverFeat [34], which had the previous best result at 24.3%.

archical, multi-stage processes for computing features that are even more informative for visual recognition.

Fukushima's "neocognitron" [19], a biologically-inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process.

Fast, Faster!

Fast RCNN

ICCV 2015

Fast R-CNN

Ross Girshick
Microsoft Research
rbg@microsoft.com

Abstract

This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) *for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. Fast R-CNN trains the very deep VGG16 network 9× faster than R-CNN, is 213× faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3× faster, tests 10× faster, and is more accurate. Fast R-CNN is implemented in Python and C++ (using Caffe) and is available under the open-source MIT License at https://github.com/rbgirshick/fast-rcnn.*

1. Introduction

Recently, deep ConvNets [14, 16] have significantly improved image classification [14] and object detection [9, 19] accuracy. Compared to image classification, object detection is a more challenging task that requires more complex methods to solve. Due to this complexity, current approaches (*e.g.*, [9, 11, 19, 25]) train models in multi-stage while achieving top accuracy on PASCAL VOC 2012 [7] with a mAP of 66% (vs. 62% for R-CNN).[1]

1.1. R-CNN and SPPnet

The Region-based Convolutional Network method (R-CNN) [9] achieves excellent object detection accuracy by using a deep ConvNet to classify object proposals. R-CNN, however, has notable drawbacks:

1. **Training is a multi-stage pipeline.** R-CNN first fine-tunes a ConvNet on object proposals using log loss. Then, it fits SVMs to ConvNet features. These SVMs act as object detectors, replacing the softmax classifier learnt by fine-tuning. In the third training stage, bounding-box regressors are learned.
2. **Training is expensive in space and time.** For SVM and bounding-box regressor training, features are extracted from each object proposal in each image and written to disk. With very deep networks, such as VGG16, this process takes 2.5 GPU-days for the 5k images of the VOC07 trainval set. These features require hundreds of gigabytes of storage.
3. **Object detection is slow.** At test-time, features are extracted from each object proposal in each test image. Detection with VGG16 takes 47s / image (on a GPU).

Fast, Faster!

Faster RCNN

NeurIP 2015

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren* **Kaiming He** **Ross Girshick** **Jian Sun**
Microsoft Research
{v-shren, kahe, rbg, jiansun}@microsoft.com

Abstract

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [7] and Fast R-CNN [5] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a *Region Proposal Network* (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model [19], our detection system has a frame rate of 5fps (*including all steps*) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image. Code is available at `https://github.com/ShaoqingRen/faster_rcnn`.

1 Introduction

Recent advances in object detection are driven by the success of region proposal methods (*e.g.*, [22]) and region-based convolutional neural networks (R-CNNs) [6]. Although region-based CNNs were computationally expensive as originally developed in [6], their cost has been drastically reduced thanks to sharing convolutions across proposals [7, 5]. The latest incarnation, Fast R-CNN [5], achieves near real-time rates using very deep networks [19], *when ignoring the time spent on region proposals*. Now, proposals are the computational bottleneck in state-of-the-art detection systems.

Fast, Faster!

Faster RCNN

NeurIP 2015

Searched, no
Fastest RCNN yet!

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren* **Kaiming He** **Ross Girshick** **Jian Sun**
Microsoft Research
{v-shren, kahe, rbg, jiansun}@microsoft.com

Abstract

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [7] and Fast R-CNN [5] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a *Region Proposal Network* (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model [19], our detection system has a frame rate of 5fps (*including all steps*) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image. Code is available at `https://github.com/ShaoqingRen/faster_rcnn`.

1 Introduction

Recent advances in object detection are driven by the success of region proposal methods (*e.g.*, [22]) and region-based convolutional neural networks (R-CNNs) [6]. Although region-based CNNs were computationally expensive as originally developed in [6], their cost has been drastically reduced thanks to sharing convolutions across proposals [7, 5]. The latest incarnation, Fast R-CNN [5], achieves near real-time rates using very deep networks [19], *when ignoring the time spent on region proposals*. Now, proposals are the computational bottleneck in state-of-the-art detection systems.


2

Attention!

Attention is all you need

NeurIPS 2017

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Attention!

Attention is **not** all you need

CoRR 2021

Attention is not *all* you need:
pure attention loses rank doubly exponentially with depth

Yihe Dong [1] **Jean-Baptiste Cordonnier** [2] **Andreas Loukas** [3]

Abstract

Attention-based architectures have become ubiquitous in machine learning. Yet, our understanding of the reasons for their effectiveness remains limited. This work proposes a new way to understand self-attention networks: we show that their output can be decomposed into a sum of smaller terms—or paths—each involving the operation of a sequence of attention heads across layers. Using this path decomposition, we prove that self-attention possesses a strong inductive bias towards "token uniformity". Specifically, without skip connections or multi-layer perceptrons (MLPs), the output converges doubly exponentially to a rank-1 matrix. On the other hand, skip connections and MLPs stop the output from degeneration. Our experiments verify the convergence results on standard transformer architectures.

attention layers. Surprisingly, we find that *pure* self-attention networks (SANs), i.e., transformers with skip connections and multi-layer perceptrons (MLPs) disabled, lose expressive power *doubly exponentially* with respect to network depth. More specifically, we prove that the output converges with a cubic rate to a rank one matrix that has identical rows. While we derive the convergence bounds in part by using properties of stochastic matrices, our results go beyond what one would expect based on standard results. In particular, by leveraging the cascading effects of specifically stacking self-attention modules, we show exponentially faster convergence than what standard theory prescribes. Furthermore, while previous studies have considered the rank of individual self-attention matrices (Wang et al., 2020; Katharopoulos et al., 2020; Cordonnier et al., 2020b), our results are the first to address conditions under which the *entire* network converges to rank *one*.

This raises the question, why do transformers work? Our

Attention!

Attention is **indeed** all you need

INLG 2021

Attention Is Indeed All You Need:
Semantically Attention-Guided Decoding for Data-to-Text NLG

Juraj Juraska and **Marilyn Walker**
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz
{jjuraska,mawalker}@ucsc.edu

Abstract

Ever since neural models were adopted in data-to-text language generation, they have invariably been reliant on extrinsic components to improve their semantic accuracy, because the models normally do not exhibit the ability to generate text that reliably mentions all of the information provided in the input. In this paper, we propose a novel decoding method that extracts interpretable information from encoder-decoder models' cross-attention, and uses it to infer which attributes are mentioned in the generated text, which is subsequently used to rescore beam hypotheses. Using this decoding method with T5 and BART, we show on three datasets its ability to dramatically reduce semantic errors in the generated outputs, while maintaining their state-of-the-art quality.

1 Introduction

Task-oriented dialogue systems require high semantic fidelity of the generated responses in order fully utilize the model's knowledge. The method we propose extracts interpretable information from the model's cross-attention mechanism at each decoding step, and uses it to infer which slots have been correctly realized in the output. Coupled with beam search, we use the inferred slot realizations to rescore the beam hypotheses, preferring those with the fewest missing or incorrect slot mentions.

To summarize our contributions, the proposed semantic attention-guided decoding method, or SEA-GUIDE for short: **(1)** drastically reduces semantic errors in the generated text (shown on the E2E, ViGGO, and MultiWOZ datasets); **(2)** is domain- and model-independent for encoder-decoder architectures with cross-attention, as shown on different sizes of T5 and BART; **(3)** works out of the box, but is parameterizable, which allows for further optimization; **(4)** adds only a small performance overhead over beam search decoding; and **(5)** perhaps most importantly, requires no model modifications, no additional training data or data preprocessing

Attention!

Attention is **indeed** all you need

INLG 2021

Attention Is Indeed All You Need:
Semantically Attention-Guided Decoding for Data-to-Text NLG

Juraj Juraska and **Marilyn Walker**
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz
{jjuraska,mawalker}@ucsc.edu

Attention is indeed not all you need?

ICML 2022/2023/... ?

Abstract

Ever since neural models were adopted in data-to-text language generation, they have invariably been reliant on extrinsic components to improve their semantic accuracy, because the models normally do not exhibit the ability to generate text that reliably mentions all of the information provided in the input. In this paper, we propose a novel decoding method that extracts interpretable information from encoder-decoder models' cross-attention, and uses it to infer which attributes are mentioned in the generated text, which is subsequently used to rescore beam hypotheses. Using this decoding method with T5 and BART, we show on three datasets its ability to dramatically reduce semantic errors in the generated outputs, while maintaining their state-of-the-art quality.

1 Introduction

Task-oriented dialogue systems require high semantic fidelity of the generated responses in order fully utilize the model's knowledge. The method we propose extracts interpretable information from the model's cross-attention mechanism at each decoding step, and uses it to infer which slots have been correctly realized in the output. Coupled with beam search, we use the inferred slot realizations to rescore the beam hypotheses, preferring those with the fewest missing or incorrect slot mentions.

To summarize our contributions, the proposed semantic attention-guided decoding method, or SEA-GUIDE for short: **(1)** drastically reduces semantic errors in the generated text (shown on the E2E, ViGGO, and MultiWOZ datasets); **(2)** is domain- and model-independent for encoder-decoder architectures with cross-attention, as shown on different sizes of T5 and BART; **(3)** works out of the box, but is parameterizable, which allows for further optimization; **(4)** adds only a small performance overhead over beam search decoding; and **(5)** perhaps most importantly, requires no model modifications, no additional training data or data preprocessing


1

Chicken, Chicken, Chicken!

Delivery of roxarsone via **chicken** diet → **chicken** → **chicken** manure → soil → rice plant

Elsevier STE (IF 7.96), 2016

Delivery of roxarsone via chicken diet → chicken → chicken manure → soil → rice plant

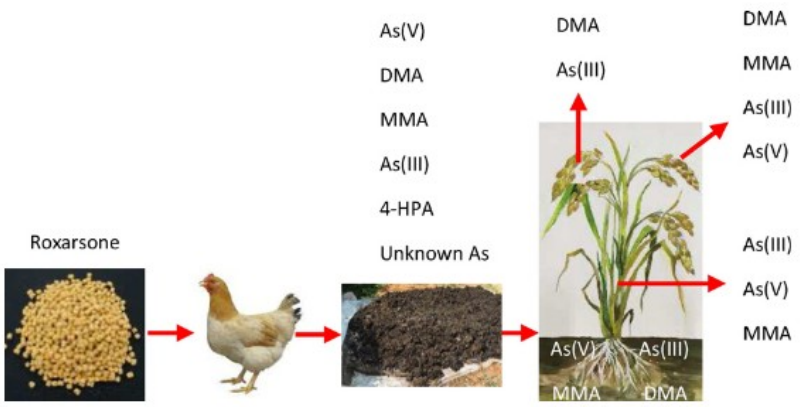
Lixian Yao [a,*], Lianxi Huang [b], Zhaohuan He [b], Changmin Zhou [b], Weisheng Lu [a], Cuihua Bai [a]

[a] College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China
[b] Institute of Agricultural Resources and Environment, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China

HIGHLIGHTS

- We examined the conversion of roxarsone during its delivery via human food chain.
- Roxarsone is transformed to more toxic As species in rice plant and paddy soil.
- The quantitative delivery of ROX via human food chain is firstly reported.
- Animal manure bearing roxarsone metabolites should not be used in rice production.

GRAPHICAL ABSTRACT



Chicken, Chicken, Chicken!

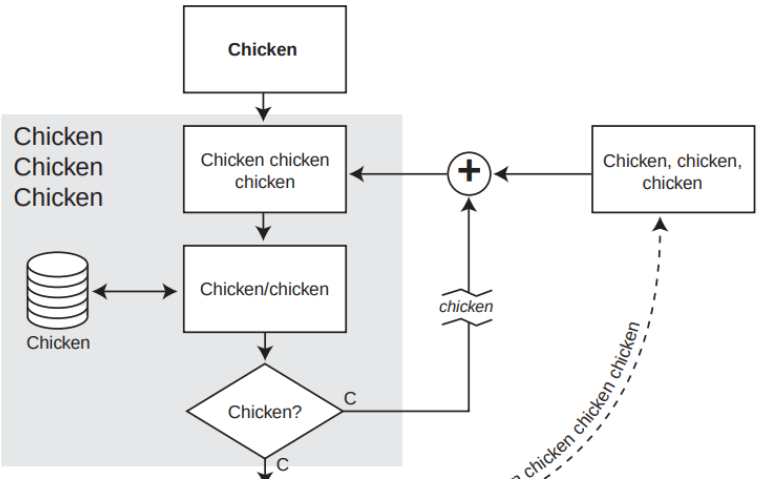
Chicken Chicken Chicken: Chicken Chicken

Doug Zongker

University of Washington

Chicken

Chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken *chicken* chicken chicken chicken chicken chicken. Chicken chicken chicken chicken chicken chicken chicken chicken. Chicken, chicken chicken chicken, chicken chicken, chicken chicken chicken chicken "chicken chicken" chicken "chicken chicken" chicken. Chicken, chicken chicken chicken chicken chicken chicken (chicken chicken) chicken chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken. Chicken chicken chicken chicken chicken chicken chicken chicken, chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken.



AAAS 2016 Presentation: https://www.youtube.com/watch?v=yL_-1d9OSdk
Paper: https://isotropic.org/papers/chicken.pdf