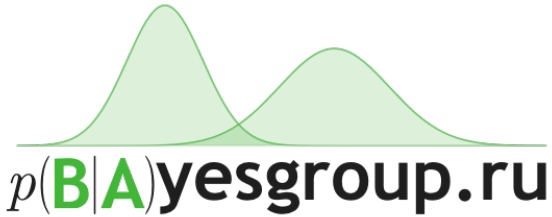
Научная работа



Слайды подготовлены на основе доклада Арсения Ашуха и Дмитрия Молчанова.

Из чего состоит научная работа?

- Чтение статей
- Проведение исследований
- Написание текстов
- Проведение презентаций

Сегодня, о том как нужно и как НЕ нужно:

- проводить исследование (научный метод)
- читать/писать научные тексты
- делать презентации, строить графики

Научный метод

- Любое научное утверждение
  - должно быть доказано
  - может быть опровергнуто
  - должно дополняться соображениями о том, каким образом выполняется утверждение
  - должно сопровождаться указаниями на его собственные «слабые места»
- Не следует увеличивать число сущностей без необходимости
- Эксперименты должны опровергать неверные гипотезы
- Эксперимент не должен быть подогнан под гипотезу

http://elementy.ru/nauchno-populyarnaya_biblioteka/430627

Научный метод

- Любое научное утверждение
  - должно быть доказано
  - может быть опровергнуто
  - должно дополняться соображениями о том, каким образом выполняется утверждение
  - должно сопровождаться указаниями на его собственные «слабые места»
- Не следует увеличивать число сущностей без необходимости
- Эксперименты должны опровергать неверные гипотезы
- Эксперимент не должен быть подогнан под гипотезу

http://elementy.ru/nauchno-populyarnaya_biblioteka/430627

О чем не забывать в исследовании

- Изучение литературы
- Сравнение с существующими методами

- Статистическая значимость результатов
- Ablation study
- Мотивационные примеры

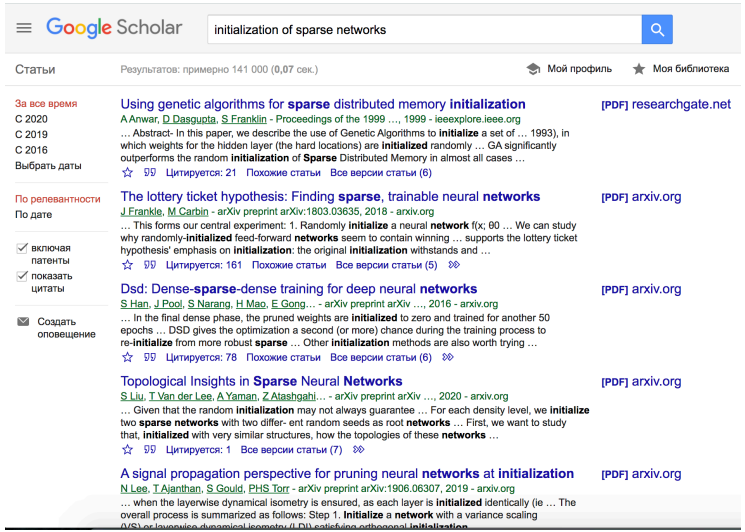
- Воспроизводимость результатов
- Сохранение всех результатов (моделей, метрик и тп)

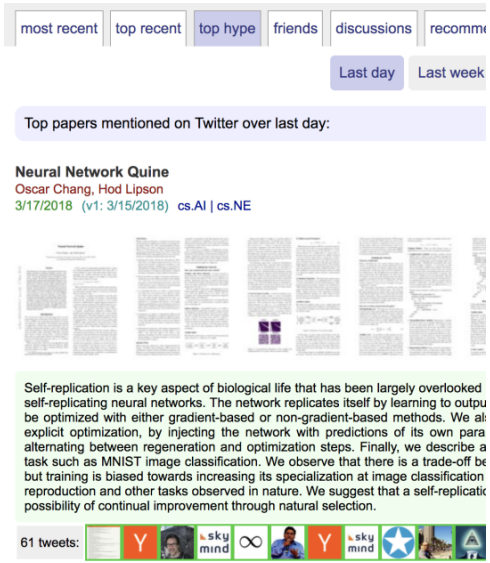
Откуда брать статьи

Ведущие конференции по ML:

- ICML, NIPS, ICLR – общие
- ACL, NAACL, EMNLP – обработка естественного языка
- ICCV, CVPR – компьютерное зрение
- и т.д.

Откуда брать статьи



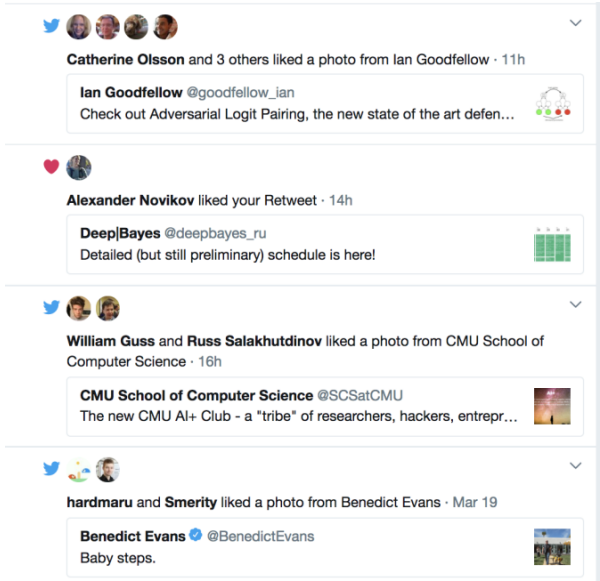


[Google Scholar]

[Arxiv-sanity]

Откуда брать статьи

/r/machinelearning



https://www.reddit.com/r/MachineLearning/comments/5jjzny/d_deep_learning_twitter_loop/

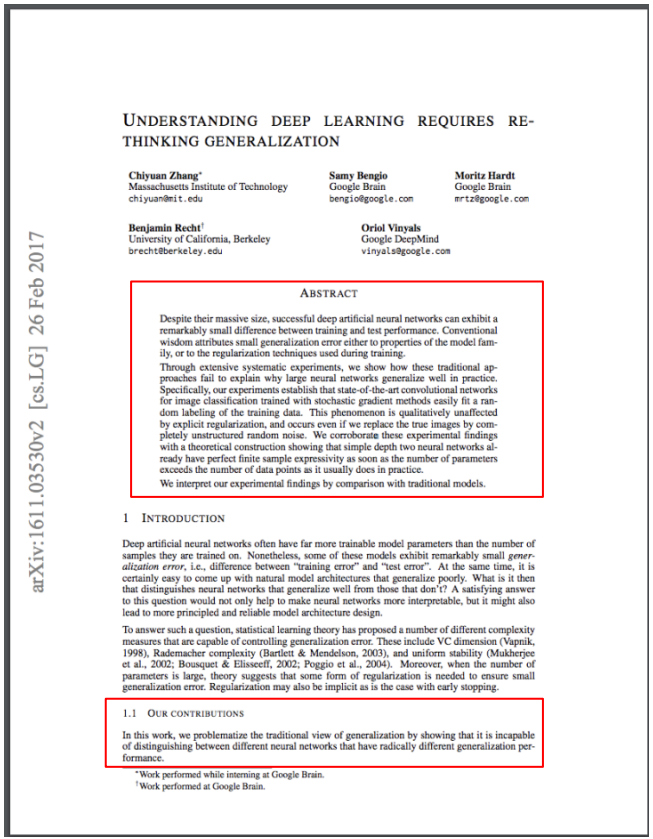
Как читать статьи

Очень популярный обзор: http://ccr.sigcomm.org/online/files/p83-keshavA.pdf

Основная идея – есть несколько этапов, сколько из них проходить зависит от целей:

1. Заголовок + абстракт
2. Введение, заключение, названия секций + возможно схемы
3. Беглое прочтение всей статьи
4. Доскональное прочтение всей статьи и аппендикса

Как читать статьи



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

ABSTRACT

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

We interpret our experimental findings by comparison with traditional models.

1 INTRODUCTION

Deep artificial neural networks often have far more trainable model parameters than the number of samples they are trained on. Nonetheless, some of these models exhibit remarkably small *generalization error*, i.e., difference between "training error" and "test error". At the same time, it is certainly easy to come up with natural model architectures that generalize poorly. What is it then that distinguishes neural networks that generalize well from those that don't? A satisfying answer to this question would not only help to make neural networks more interpretable, but it might also lead to more principled and reliable model architecture design.

To answer such a question, statistical learning theory has proposed a number of different complexity measures that are capable of controlling generalization error. These include VC dimension (Vapnik, 1998), Rademacher complexity (Bartlett & Mendelson, 2003), and uniform stability (Mukherjee et al., 2002; Bousquet & Elisseeff, 2002; Poggio et al., 2004). Moreover, when the number of parameters is large, theory suggests that some form of regularization is needed to ensure small generalization error. Regularization may also be implicit as is the case with early stopping.

1.1 OUR CONTRIBUTIONS

In this work, we problematize the traditional view of generalization by showing that it is incapable of distinguishing between different neural networks that have radically different generalization performance.

*Work performed while interning at Google Brain.
†Work performed at Google Brain.

Variational Dropout Sparsifies Deep Neural Networks

Dmitry Molchanov [1 2 *] **Arsenii Ashukha** [3 4 *] **Dmitry Vetrov** [3 1]

Abstract

We explore a recently proposed Variational Dropout technique that provided an elegant Bayesian interpretation to Gaussian Dropout. We extend Variational Dropout to the case when dropout rates are unbounded, propose a way to reduce the variance of the gradient estimator and report first experimental results with individual dropout rates per weight. Interestingly, it leads to extremely sparse solutions both in fully-connected and convolutional layers. This effect is similar to automatic relevance determination effect in empirical Bayes but has a number of advantages. We reduce the number of parameters up to 280 times on LeNet architectures and up to 68 times on VGG-like networks with a negligible decrease of accuracy.

1. Introduction

Deep neural networks (DNNs) are a widely popular family of models which is currently state-of-the-art in many important problems (Szegedy et al., 2016; Silver et al., 2016). However, DNNs often have many more parameters than the number of the training instances. This makes them prone to overfitting (Hinton et al., 2012; Zhang et al., 2016) and necessitates using regularization. A commonly used regularizer is Binary Dropout (Hinton et al., 2012) that prevents co-adaptation of neurons by randomly dropping them during training. An equally effective alternative is Gaussian Dropout (Srivastava et al., 2014) that multiplies the outputs of the neurons by Gaussian random noise.

Dropout requires specifying the dropout rates which are the probabilities of dropping a neuron. The dropout rates are typically optimized using grid search. To avoid the exponential complexity of optimizing multiple hyperparameters, the dropout rates are usually shared for all layers. Recently it was shown that dropout can be seen as a special case of Bayesian regularization (Gal & Ghahramani, 2015; Kingma et al., 2015). It is an important theoretical result that justifies dropout and at the same time allows us to tune individual dropout rates for each weight, neuron or layer in a Bayesian way.

Instead of injecting noise we can regularize a model by reducing the number of its parameters. This technique is especially attractive in the case of deep neural networks. Modern neural networks contain hundreds of millions of parameters (Szegedy et al., 2015; He et al., 2015) and require a lot of computational and memory resources. It restricts us from using deep neural networks when those resources are limited. Inducing sparsity during training of DNNs leads to regularization, compression, and acceleration of the resulting model (Han et al., 2015a; Scardapane et al., 2016).

Sparse Bayesian Learning (Tipping, 2001) provides a principled framework for training of sparse models without the manual tuning of hyperparameters. Unfortunately, this approach does not extend straightforwardly to DNNs. During past several years, a number of papers (Hoffman et al., 2013; Kingma & Welling, 2013; Rezende et al., 2014) on scalable variational inference have appeared. These techniques make it possible to train Bayesian Deep Neural Networks using stochastic optimization and provide us an opportunity to transfer Bayesian regularization techniques from simple models to DNNs.

In this paper, we study Variational Dropout (Kingma et al., 2015) in the case when each weight of a model has its individual dropout rate. We propose Sparse Variational Dropout that extends Variational Dropout to all possible values of dropout rates and leads to a sparse solution. To achieve this goal, we provide a new approximation of the KL-divergence term in Variational Dropout objective that is tight on the full domain. We also propose a way to greatly reduce the variance of the stochastic gradient estimator and show that it leads to a much faster convergence and a better value of the objective function. We show theoretically that

[*]Equal contribution [1]Yandex, Russia [2]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, Russia [3]National Research University Higher School of Economics, Moscow, Russia [4]Moscow Institute of Physics and Technology, Moscow, Russia. Correspondence to: Dmitry Molchanov <dmitry.molchanov@skolkovotech.ru>, Arsenii Ashukha <ars.ashuha@gmail.com>, Dmitry Vetrov <vetrovd@yandex.ru>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Какие тексты приходится писать?

- Статьи
- Курсовые работы / дипломные работы / диссертации / ...
- Рецензии
- Ответы на рецензии (ребатл)
- Технические отчеты для компаний / грантов / ...
- Заявки на гранты
- …

Какие приходится делать доклады?

- Презентация статьи на семинаре
- Презентация статьи на конференции
- Защита курсовой / диплома / диссертации

Основные различия – ограничение по времени и подробность изложения

С чего начать подготовку?

- Изучить материал, чтобы понять примерный план доклада
- Понять стартовые и конечные точки (принять во внимание уровень подготовки аудитории и временные рамки доклада)
- Составить список основных пунктов, которые слушатели должны вынести из вашего доклада
- Составить краткий план доклада – основные пункты и переходы между ними
- Составить подробный план доклада (что будет на каждом слайде)
- Сделать презентацию (по ходу будут вноситься правки в план)

Структура доклада

- Титульный слайд
- (Optional) План доклада
- Постановка проблемы и мотивация
- Постепенный ввод необходимых обозначений и понятий
- Описание основной идеи работы, желательно на примерах
- (Optional) Более подробное описание, технические детали
- Эксперименты
- Заключение

Как делать презентации

- Презентация помогает рассказчику, дополняет рассказ
- Что говорить с этим слайдом → что писать на слайде
- Один слайд на 1-2 минуты
  - Не стоит делать больше 20 слайдов в презентации на полчаса
- Читаемый шрифт (здесь 18pt) (мельче (14) делать (12) не стоит (10))
- Несколько простых слайдов лучше одного сложного
- Цвета – рамки – стрелки

Хорошо в меру

Кликабельные ссылки на источники можно ставить в footnote: youtube.com/watch?v=dQw4w9WgXcQ

Как делать презентации

Оглавление

1. Часть 1
2. Часть 2
3. Часть 3

- Простой белый фон лучше картинки
- Простой шаблон лучше перегруженного
- Анимированные переходы – зло
- Не забудьте выйти из мессенджеров и закрыть вкладки
  - Можно зайти с чистого пользователя
- И НЕ НАДО ПИСАТЬ КАПСОМ
  - Да, даже заголовки

Как делать презентации

Длинные абзацы текста в презентации на устном докладе считаются дурным тоном. Они гораздо тяжелее читаются, и слушателю приходится отвлекаться от, собственно, рассказа докладчика, чтобы все-таки прочитать эту стену текста. Обычно информацию можно донести более эффективно.

- Списки читаются гораздо проще
- На слайды нужно выносить ключевые идеи
- Не обязательно писать полные предложения

How to make a presentation

- Slides in English on a Russian talk are still OK
- Easy to reuse slides
- You can put the presentation on your homepage
- Не мешайте языки вместе

Формулы в презентации

- Идеи первичны
- Формулы должны иллюстрировать доклад, а не быть самоцелью
- Стараться опускать тяжелые выкладки

$$\mathrm{LogN}_{[a,b]}(x \mid \mu, \sigma^2) = \frac{1}{Zx\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2 \right),\ \log x \in [a, b],\ x > 0$$

$$\mathrm{Var}\left[x \sim \mathrm{LogN}_{[a,b]}(x \mid \mu, \sigma^2) \right] = \frac{1}{Z} \int_{e^a}^{e^b} \frac{(x - \mathbb{E}x)^2}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2 \right) dx =$$

$$= \frac{1}{Z} \left[\int_{e^a}^{\infty} \frac{(x - \mathbb{E}x)^2}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2 \right) dx - \int_{e^b}^{\infty} \frac{(x - \mathbb{E}x)^2}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2 \right) dx \right]$$

Формулы в презентации

- Идеи первичны
- Формулы должны иллюстрировать доклад, а не быть самоцелью
- Стараться опускать тяжелые выкладки
- Все формулы надо проговаривать
- Большие формулы ⇒ используем выделение

$$\mathbb{E}_{q(\widetilde{W}\,|\,\phi)} \log p(y\,|\,x, \widetilde{W}) - D_{KL}(q(\widetilde{W}\,|\,\phi)\,||\,p(\widetilde{W})) \to \max_{\phi}$$

Data-term
e.g. cross-entropy loss

Regularizer
i.e. KL-distance

Как готовиться к выступлению

- Продумать, о чем говорить на каждом слайде
  - Можно писать план в speaker notes
- Доклад до 15 минут можно выучить наизусть
- Продумать интонации – нужно рассказывать, а не монотонно зачитывать
- Особенно поначалу очень важно репетировать
  - Не смотря в написанный текст
  - Стоя и говоря вслух
  - С диктофоном / камерой / другом / …
  - С часами (!)

Как не волноваться на выступлении

- Никак :)
- Со стороны сложно заметить волнение
- Обычно цель аудитории – понять доклад, а не ругать / высмеивать докладчика
- Можно выучить первую и последнюю минуты доклада наизусть
- Или весь доклад целиком (если он <15 минут)
- С каждым разом становится немного проще

Заключение

- Умение работать с научными статьями, выступать с докладами и писать тексты – важная часть научной (и не только) работы
- Это достаточно технические навыки, которые можно и нужно развивать
- Именно для этого мы все тут и собрались

Спасибо за внимание!

Время для вопросов!

- Так тоже не надо делать
- Лучше остановиться на слайде с заключением