深度学习重要算法盘点

同学们，大家好！今天给大家分享深度学习重要算法。

导读
深度学习领域每天都有大量的新研究和论文发表。在这背后，是许多经过时间考验的、被广泛接纳的基础算法、模型和概念。为帮助更好地理解其发展，“数据实战派” 奉上一份对过去几年重磅研究的优质盘点：

原文：Deep Learning’ s Most Important Ideas[1]
作者：Denny Britz（ML 研究员，Google Brain 前成员）
译者：REN

深度学习是一个瞬息万变的领域，层出不穷的论文和新思路可能会令人不知所措。即使是经验丰富的研究人员，也很难准确将研究成果传达给公司的公关部门，继而传达给大众。

对于初学者来说，理解和实现这些技术有利于打下坚实的理论基础，是入门的最佳方法。

在深度学习领域，很多技术都可以跨域多个应用领域，包括计算机视觉，自然语言，语音识别和强化学习等等。在计算机视觉领域使用过深度学习的人，可能很快就能将类似的技术应用到自然语言研究中，即使特定的网络结构有所不同，但其概念，实现方法和代码基本一致。

必须强调的是，本文侧重于计算机视觉，自然语言，语音识别和强化学习领域，但不会详细解释每种深度学习技术，用寥寥数百字解释清楚一篇几十页的论文是不现实的。另外还有一些不容易重现的重要研究，比如 DeepMind 的 AlphaGo 或 OpenAI 的 OpenAI Five（Dota 2 模型），涉及到巨大的工程和运算挑战，因此也不是讨论的重点。

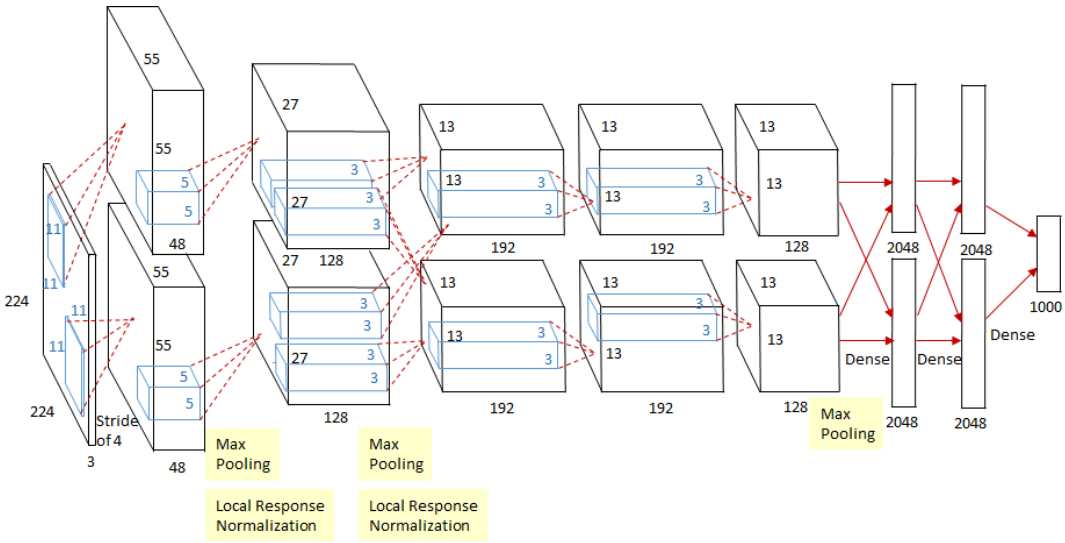
这篇文章的目的，是回顾在深度学习领域影响深远的成果，概述每种技术及其历史背景，尽量引导深度学习新人接触多个领域的基础技术。它们是这个领域最值得信赖的基石，每一个技术都经过了无数次的引用、使用和改进，经得起推敲。

最后分享，文末会附上论文和代码链接。如果想要更好地掌握基础技术和知识，可以尝试先不看参考代码，从零开始用 PyTorch 实现论文中的某些算法。

2012 年：用 AlexNet 和 Dropout 解决 ImageNet 图像分类

AlexNet 通常被认为是近年来引领深度学习和人工智能研究蓬勃发展的基础算法。它是一种深度卷积神经网络（CNN），基于人工智能大牛 Yann LeCun 早年间开发的 LeNet 模型。

AlexNet 结合了 GPU 的强大性能和先进的算法，在对 ImageNet 图像数据集分类时，其表现远远超越了之前的所有算法。它证明了神经网络真的很好用（至少在图像分类上）。AlexNet 也是首次使用 Dropout 技巧的算法之一，为了防止过拟合。此后 Dropout 成为了提高各种深度学习模型泛化能力的重要工具。



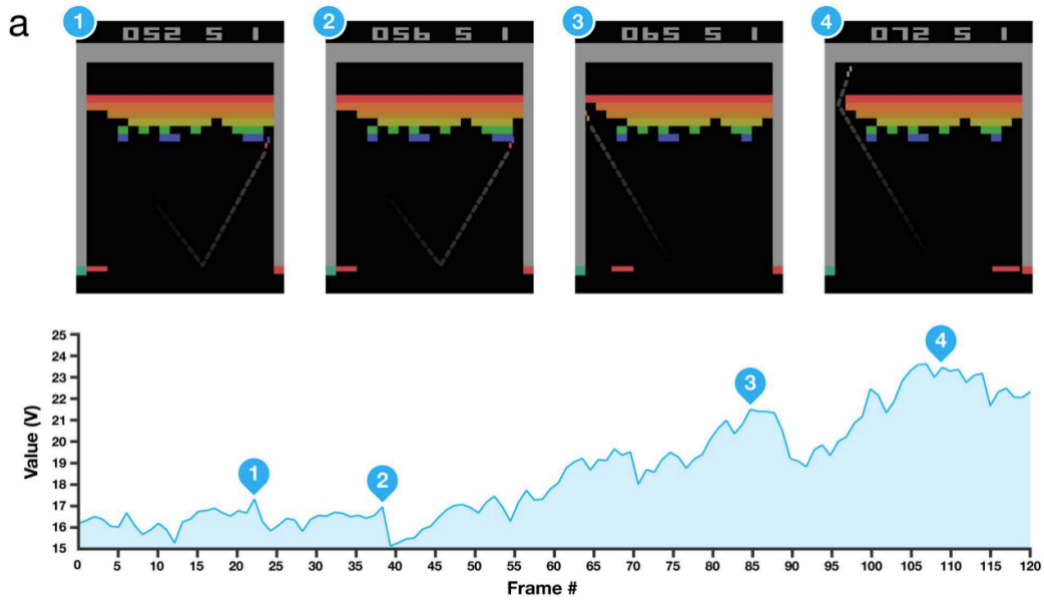
AlexNet 使用的结构，包括一系列卷积层，ReLU 非线性激活函数和最大池化（Max-pooling）已成为公认标准模式，在许多计算机视觉模型结构中都有用到和进一步扩展。

与最新的模型相比，AlexNet 显得异常简单，得益于 PyTorch 等强大的软件库，仅需几行代码即可实现。值得注意的是，目前 AlexNet 的许多实现方法都与最早论文中阐述的有些许不同，目的是为了对卷积神经网络并行运算。

2013 年：利用深度强化学习玩 Atari 游戏

基于在图像识别和 GPU 方面取得的突破，DeepMind 团队成功利用强化学习（RL）训练了一个神经网络，可以通过原始像素输入来玩 Atari 游戏。而且在不知道游戏规则的前提下，相同的神经网络模型还学会了玩 7 种不同的游戏，证明了这种方法的泛化性。

强化学习与监督学习（例如图像分类）的不同之处在于，AI 代理（agent）必须学会在多个时间点上最大化整体奖励，比如赢得一场比赛，而不仅仅是预测分类标签。



由于 AI 智能体直接与环境交互且每个动作都会影响环境，因此训练数据不是独立同分布的（i.i.d.），这使得许多机器学习模型的训练非常不稳定。这可以使用经验回放等技术解决。

尽管没有明显的算法创新，但 DeepMind 的研究巧妙地结合了当时的现有技术，在 GPU 上训练的卷积神经网络，经验回放以及一些数据处理技巧，从而实现了超出大部分人预期的惊艳结果。这使人们有信心继续探索深度强化学习技术，以解决更复杂的任务，由此演变出 AlphaGo 围棋 AI，Dota 2 AI 和星际争霸 2 AI 等等。

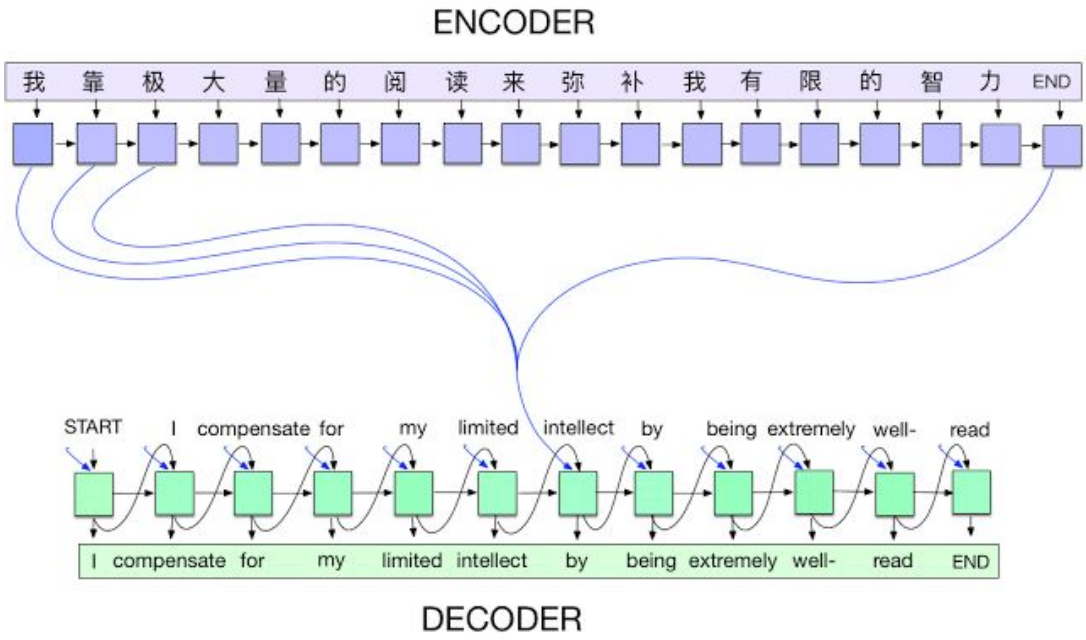
此后，Atari 系列游戏成为了强化学习研究的基准。最初的算法只能在 7 个游戏中超越人类，但未来几年中，更先进的模型开始在越来越多的游戏中击败人类。其中一款名为 “蒙特祖玛的复仇” 的游戏因需要长期规划而闻名，也被认为是最难解决的游戏之一，于 2018 年被攻克。

今年 4 月，AI 终于在 Atari 的全部 57 款游戏中超越了人类。

2014 年：采用注意力机制的编码器 - 解码器网络

在自然语言处理领域，尽管有长短期记忆网络（LSTM）和编码器 - 解码器网络（Encoder-Decoder），能够处理语言建模和翻译任务，但其实直到 2014 年注意力机制（Attention Mechanism）的问世，才获得了跨越式的进步。

在处理语言时，每个标记（token）—— 可能是字符，单词或介于两者之间的某种东西 —— 都会被输入一个循环神经网络（RNN）之中。例如 LSTM，该网络可以记住之前一定时间之内的输入值。



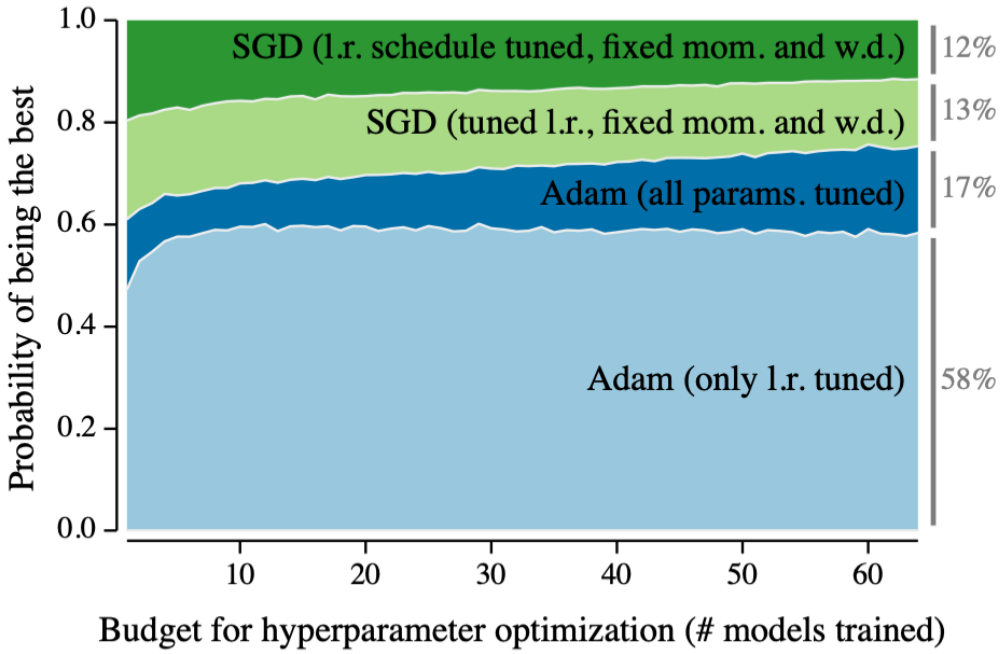
换句话说，句子与时间序列非常相似，每个标记都是一个时间步长。这些循环神经网络模型通常很难处理长时间的相关性，因为会 “忘记” 较早的输入值，而且使用梯度下降来优化这些模型同样非常困难。

新的注意力机制有助于缓解这一问题。它通过引入 “快捷连接（shortcut connections）”，使神经网络可以自适应选择 “回顾” 前面的输入值（将输入加到输出上）。在生成特定输出时，这些连接允许网络决定哪些输入很重要。翻译模型是一个经典例子，在生成翻译后的输出字 / 词时，模型会将其映射到一个或多个特定的输入字 / 词。

2014 年：Adam 优化器

训练神经网络需要使用优化器使损失函数（比如平均分类错误）最小化。优化器负责弄清楚如何调整网络参数，实现学习目标。

大多数优化器都基于随机梯度下降（SGD）及其变种。许多优化器本身都包含可调参数，例如学习率（learning rate）。为特定问题找到正确的参数配置，不仅可以减少训练时间，还可以找到更好的损失函数局部最小值，得到更好的训练结果。



大型研究实验室经常运行昂贵的超参数搜索，需要设计非常复杂的学习率变化计划，以便从优化器中获得最大收益。有时候，他们找到的最终结果超过了现有基准，但这是花费了大量资金对优化器进行优化的结果。类似的细节经常在论文中被忽略，导致没有相同预算来优化其优化器的研究人员找不到最优解。

Adam 优化器使用了自适应矩估计方法，对随机目标函数执行一阶梯度优化并自动调整学习率。结果非常可靠，并且对超参数选择不太敏感。

简而言之，Adam 不需要像其他优化器一样进行大量的调整。尽管调整得非常好的 SGD 优化器可以得到更好的结果，但是 Adam 让研究更容易完成，因为如果无法得到预想中的结果，科研人员至少可以排除优化器调整不当这一原因。

2014/2015 年：生成式对抗网络（GAN）

生成式模型的目标是创建逼真的数据样本，例如栩栩如生的假人脸图片。因为这类模型必须对全部数据分布进行建模（像素很多），而不仅仅是分类图片，所以它们通常很难训练。生成式对抗网络（GAN）就是这样一种模型。

Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and "deconvolutional" generator)

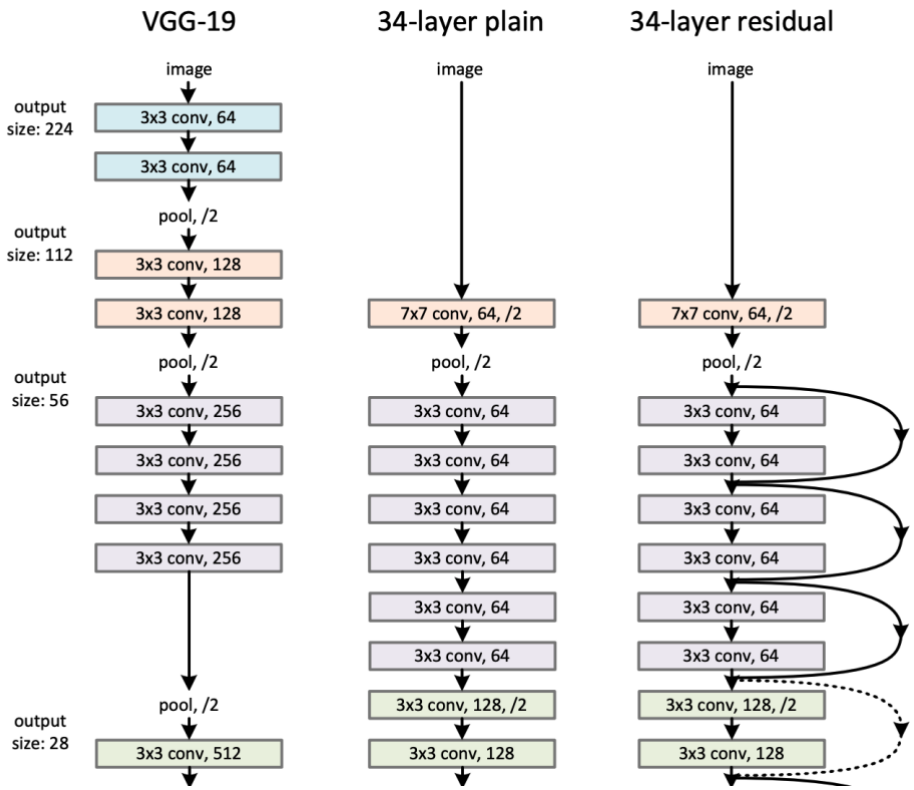
GAN 的基本思想是训练两个神经网络：生成器和判别器。生成器的目标是生成样本，蒙蔽判别器，而判别器则是要区分真实图像和（生成器生成的）虚假图像。随着训练的进行，鉴别器的表现会越来越好，但生成器也会变得更好，生成的图像也更加逼真。

初代 GAN 只能产生模糊的低分辨率图像，并且训练起来非常不稳定。但是随着人们不断努力，诞生了诸如 DCGAN，Wasserstein GAN，CycleGAN，StyleGAN 等多种多样的模型，现在已经可以生成高分辨率的逼真图像和视频。

2015 年：残差网络（ResNet）

自 2012 年以来，研究人员在 AlexNet 的基础上添砖加瓦，发明了性能更好的基于卷积神经网络的模型，例如 VGGNet 和 Inception 等等。ResNet 是其中最有代表性的一个，标志着下一个迭代。

目前，ResNet 的变体通常用作各种任务的基准模型，也被用来构建更复杂的模型。



除了在 ILSVRC 2015 分类挑战中获得第一名之外，ResNet 的过人之处还在于它的模型深度：论文中提到的最深 ResNet 有 1000 层，并且仍然表现良好，尽管在基准任务上比其 101 和 152 层对应的网络稍差。由于梯度消失，训练这种非常深的网络是一个极具挑战性的优化问题，几乎没有研究人员认为训练如此深的网络可以带来良好的稳定结果。

ResNet 使用了 “身份快捷连接（identity shortcut connections）” 连接来帮助实现梯度流动。解释这些连接的一种方法是，ResNet 只需要学习从一层到另一层的 “增量 delta” ，这通常比学习完整的（传递）要容易得多。

2017 年：Transformer

引入注意力机制的 Seq2Seq 模型已经有很好的表现，但缺点在于需要顺序计算，很难做到并行。这让研究人员很难将它们扩大到非常长的序列，即使引入了注意力机制，该模型在构建复杂的长期相关关系时仍然相形见绌。大多数的 “工作” 似乎都在循环层中完成。

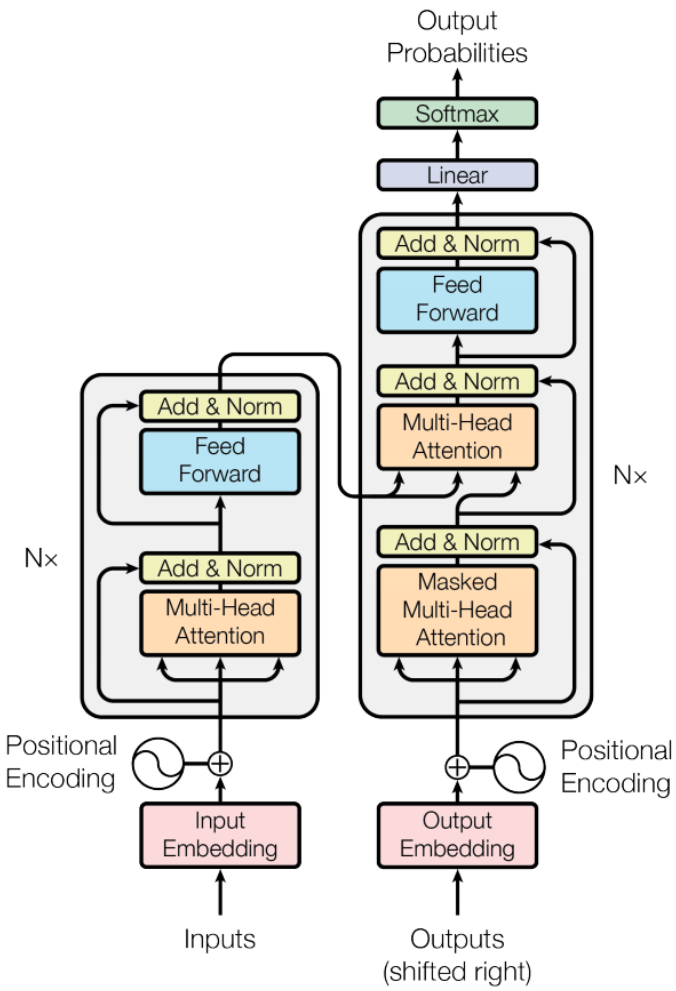


Figure 1: The Transformer - model architecture.

Transformer 的出现解决了这些问题，方法是完全消除循环，用多个前馈自我注意力层代替，然后并行处理序列中的所有单词或符号。由于输入和输出之间的路径较短，更容易通过梯度下降优化，进而实现快速训练且易于扩展。此外，Transformer 还会使用位置编码技术，将输入值的顺序告诉神经网络。

Transformer 的效果超出了所有人的预料。在接下来的几年中，它成为了绝大多数自然语言处理任务和其他序列任务的标准架构，甚至还用到了计算机视觉领域中。

2018 年：BERT 和微调自然语言处理模型

预训练是指训练模型执行某些任务，然后将学到的参数作为初始参数，用于其他类似任务中。这符合人们的直觉：一个已经学会将图像分类为猫或狗的模型，应该已经掌握了有关图像和毛茸茸的动物的通用知识。所以微调该模型并对狐狸分类时，人们希望它比从零学习的模型做得更好。

类似地，学会预测句子中下一个单词的模型应该已经学会了有关人类语言模式的通用知识。人们希望它在翻译或情绪分析等相关任务中起点更高。
预训练和微调已在计算机视觉领域作为标准使用许久，但将其运用在自然语言处理中更具挑战性。大多数表现最好的结果仍来自完全监督模型。随着 Transformer 的出现，研究人员终于可以更方便的开展预训练，由此诞生了 ELMo，ULMFiT 和 OpenAI GPT 之类的模型。

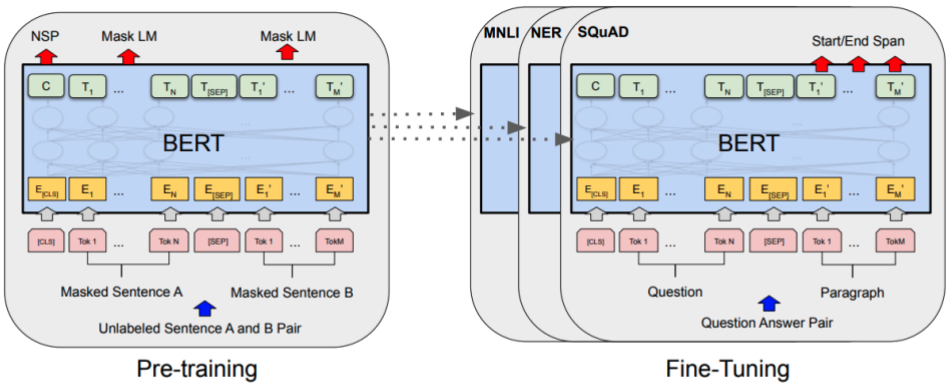


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

BERT 是这种思路的最新成果，许多人认为它开启了自然语言处理研究的新纪元。该模型在预训练中会对句子中故意被删掉的单词进行预测，还会判断两个句子彼此相连的可性能，而不是单纯地预训练如何预测下一个单词。

完成这些任务不需要标记数据，因此它可以在任何文本上训练，这意味着取之不尽的训练数据。该预训练模型已经学习了一些语言方面的常规属性，之后对其进行微调就能用来解决监督式任务，例如问题回答或预测情绪。

BERT 在各种各样的任务中都表现出色，还有 HuggingFace 一类的公司，允许人们轻松获取和微调用于各种自然语言处理任务的 BERT 类模型。时至今日，在 BERT 的基础上出现了 XLNet，RoBERTa 和 ALBERT 等更加先进的模型。

2019/2020 年及未来：巨大的语言模型和自我监督式学习

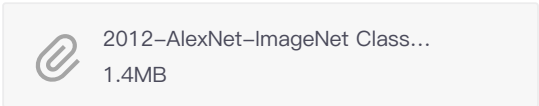
纵观深度学习历史及其发展趋势，人们不难发现，可以更好地并行运算，拥有更多数据和更多模型参数的算法一次又一次地击败了所谓的 “更聪明的技术”。这种趋势似乎一直持续至今，OpenAI 放出了拥有 1750 亿个参数的巨大语言模型 GPT-3，尽管它只有简单的训练目标和标准网络结构，却显示出无可比拟的强大泛化能力。

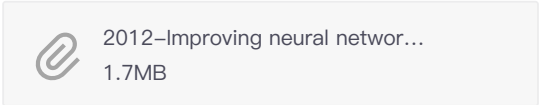
同样的趋势还出现在自我监督学习方法上，比如 SimCLR，它们可以更好地利用未标记的数据。 随着模型变大和训练速度变快，那些可以更有效地利用网络上大量未标记的数据，并将学习到的通用知识转移到其他任务上的模型将变得越来越有价值。

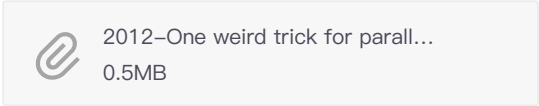
原文：Deep Learning’ s Most Important Ideas
链接：https://www.kdnuggets.com/2020/09/deep-learnings-most-important-ideas.html
作者：Denny Britz（ML 研究员，Google Brain 前成员）

文中涉及的论文

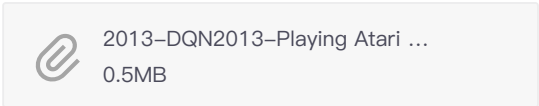
2012 年：用 AlexNet 和 Dropout 解决 ImageNet 图像分类
论文：
·ImageNet Classification with Deep Convolutional Neural Networks (2012)
·Improving neural networks by preventing co-adaptation of feature detectors (2012)
·One weird trick for parallelizing convolutional neural networks (2014)







2013 年：利用深度强化学习玩 Atari 游戏
论文：
·Playing Atari with Deep Reinforcement Learning (2013)



2014 年：采用注意力机制的编码器 – 解码器网络

论文：
·Sequence to Sequence Learning with Neural Networks
·Neural Machine Translation by Jointly Learning to Align and Translate

2014–Sequence to Sequence L…
0.1MB

2014–Neural Machine Translati…
0.4MB

2014 年：Adam 优化器
论文：
·Adam: A Method for Stochastic Optimization

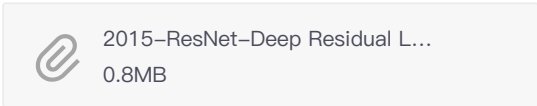
2014–Adam–A Method for Sto…
0.6MB

2014/2015 年：生成式对抗网络（GAN）
论文：
·Generative Adversarial Networks
·Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

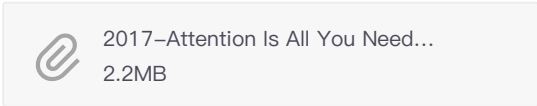
2014–Generative Adversarial N…
0.5MB

2015–Unsupervised Represent…
7.5MB

2015 年：残差网络（ResNet）
论文：
·Deep Residual Learning for Image Recognition



2017 年：Transformer
论文：
·Attention is All You Need



2018 年：BERT 和微调自然语言处理模型
论文：
·BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding