# 2024/06/02

## Exploring Connectivity Patterns: Suppressing Layer-wise Perturbation Growth by Altering Feature Evolution Trajectory

1.发现了robustness-trajectory preference现象

2.给出神经网络扰动层级递增的理论解释

3.设计了新的神经网络模型——suppression connection，通过实验证明suppression connection比residual connection更有鲁棒性

## Enhancing Representation Power of Deep Neural Networks With Negligible Parameter Growth for Industrial Applications

发表时间：2024.4

1.使用numerical method的阶数和ODE的阶数衡量神经网络的表示能力

2.设计了具有更高阶数的numerical method和ODE的网络模型，使得在增加少量参数的同时提高神经网络的表示能力（包括MPD connection和k-step connection）

3.测试了提出的神经网络模型在常见工业领域的表现（缺陷检测、超导体的临界温度预测和带噪声的图像分类）

## Zero Stability Well Predicts Performance of Convolutional Neural Networks

发表时间：2022.6

1.发现Zero Stability能很好地预测模型表现

2.设计了一种具有zero stability的神经网络——ZeroSNet，实验证明这种网络结构比非zero stable的结构有更好的泛化性能，同时在面对噪声时也有更好的鲁棒性

## Activated Gradients for Deep Neural Networks

发表时间：2021.9

提出一种激活函数——gradient activation function（GAF）用于解决病态问题、梯度消失、梯度爆炸和鞍点问题。

# 2024/06/09

## MIAdam: Improving Generalization of Adam by Multiple Integrations

通过引入多重积分使得优化器能够找到平坦极小值点，这种新的优化算法即为MIAdam。

## Deforming the Loss Surface to Affect the Behaviour of the Optimizer

发表时间：2020.9

通过对损失函数进行VDM操作（vertical deformation mapping，提出了AP VDM和LE VDM两种），使得损失函数值小的地方更陡峭，损失函数值大的地方更平缓，从而跳过陡峭的极小值点，找到平缓的极小值点。

## A DIFFUSION THEORY FOR DEEP LEARNING DYNAMICS: STOCHASTIC GRADIENT DESCENT EXPONENTIALLY FAVORS FLAT MINIMA

发表时间：2021.1

1. 给出了梯度噪声协方差矩阵与hessian矩阵之间的关系。证明了噪声的存在使得SGD更偏好平坦极值点。
2. 使用fokker plank方程解释SGD中的参数的概率密度随时间的演化过程。
3. 类比克拉默逃逸问题预测参数从一个极值点逃逸到另一个极值点的时间。

Batch size越小，噪声越大，反之亦然。大噪声和大学习率会使得参数更容易靠近平坦极小值，但不容易收敛，而小噪声小学习率则更容易收敛，但可能会落入尖锐极小值，因此在模型训练初期可以使用小batch size和大学习率ku快速逼近平坦极小值，再使用大batch size和小学习率进行微调。

## Adaptive Inertia: Disentangling the Effects of Adaptive Learning Rate and Momentum

发表时间：2022

期刊/会议：ICML 2022

通过理论分析解释了自适应学习率和动量分别对鞍点逃离和极小值选择的影响。解释了为什么Adam善于逃离鞍点但不善于达到平坦极小值。提出了Adai模型，能够逐参数地动态调整动量的超参数，使得Adai能在加速训练的同时偏好平坦极小值。

## SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION

发表时间：2020

使用同时优化尖锐性和损失的思路，构建了一个新的损失函数，称之为sam算法。

# 2024/06/30

## Improving Generalization Performance by Switching from Adam to SGD

设计了一种名为SWATS的方法来在合适时机由Adam切换为SGD，提升模型泛化性能。

## ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

发表时间：2014

期刊/会议：ICLR

提出了Adam。Adam结合了AdaGrad和RMSProp的优点，引入了bias correction机制。文章给出了Adam收敛性的严格证明。验证了Adam在多个机器学习任务上的有效性。

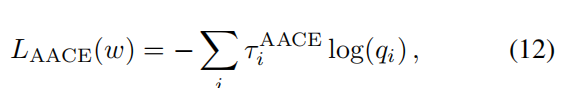
# 2024/07/07

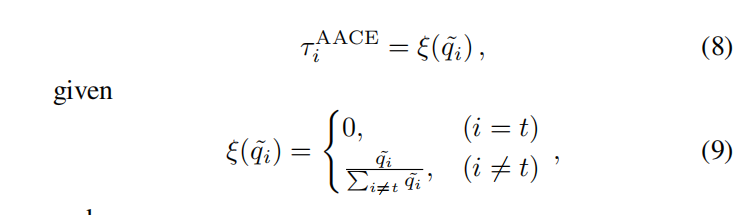
## When Will Gradient Regularization Be Harmful?

梯度正则化（通常指L2正则化）与自适应优化器（Adam）同时使用是不合适的，原因在于训练初期的梯度正则化过大，因此作者提出了三种梯度正则化的warm up策略（r-warmup, λ-warmup, zero-warmup）解决这一问题。

## ADAPTIVE ADVERSARIAL CROSS-ENTROPY LOSS FOR SHARPNESS-AWARE MINIMIZATION

设计了一种名为ADAPTIVE ADVERSARIAL CROSS-ENTROPY（AACE）的损失函数来改进SAM的性能：





同时在计算扰动向量时不对梯度进行正则化。通过这两点保证扰动向量在极值点不减小，使得其探索更广的空间，并且防止了梯度消失。使用Wide ResNet、PyramidNet模型在CIFAR-100、CIFAR-10、Fashion-MNIST、Food101几个数据集上进行测试。

## ImageNet Classification with Deep Convolutional Neural Networks

发表时间：2012

期刊/会议：NeurIPS

提出了AlexNet，一种由5个卷积层和两个全连接层构成的模型。首次提出了深度卷积神经网络的概念，是卷积神经网络在图像分类领域的奠基之作。其中将输出向量相近的图片进行对比的做法颇具启发性，对神经网络的可解释性研究有所贡献。

## Deep Residual Learning for Image Recognition

发表时间：2015

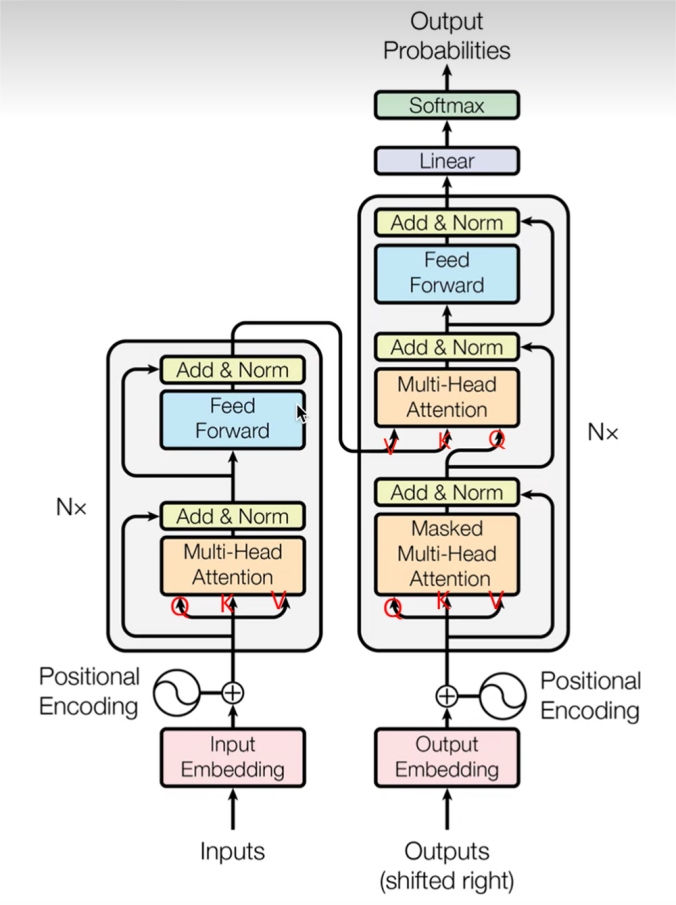
期刊/会议：CVPR

1. 将残差连接用于深度网络模型，本质上是在每一层的输入与输出之间加入了一个恒等变换，使得每一层在最坏情况下也不会降低整个模型的表现，可以理解为一种特殊的正则化，降低了整个模型的复杂度，使得大幅堆网络深度成为可能，是这方面的开山之作。
2. 引入了bottleneck design，即把通道数很高的数据通过1\*1卷积核转换为低通道数据，卷积后再转换为高通道数据（与输入通道数相等），以此大幅降低了计算量。

## Attention is all you need

发表时间：2017

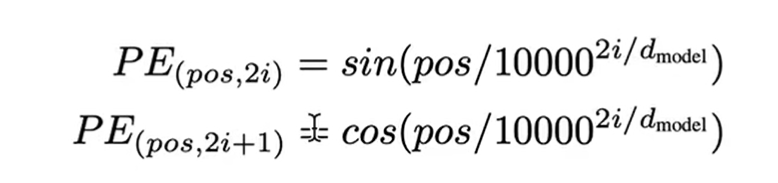
期刊/会议：NeurIPS



提出了Transformer模型，使人们看到了在RNN之外的另一种能良好处理序列数据的模型，为研究打开了新思路，此后各种新模型纷纷涌现。

Transformer是一种seq2seq模型，虽然Transformer之后被用于自然语言处理、计算机视觉、语音识别等等几乎所有领域，但其本身是针对机器翻译问题提出的。Transformer由encoder和decoder两个模块构成。其中encoder由以下几个部分构成：

1. **input word embedding：**由稀疏的one-hot编码进入一个不带bias的前馈神经网络得到一个向量。
2. **position encoding：**通过正余弦函数表示单词的位置信息。

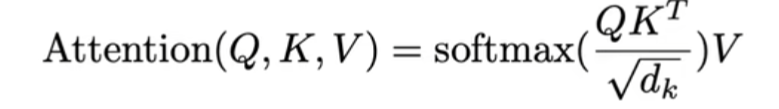


1. **multi-head self-attention：**由多组Q、K、V组成，每组Q、K、V各自计算一个attention向量，拼接起来后放入一个FFN中得到最终的输出向量。
2. **feed forward network：**每个单独位置各有一个，不同位置共享参数。

**decoder由以下几个部分构成：**

1. **output word embedding：**与encoder类似，不过输入为另一个序列单词的向量。
2. **masked multi-head self-attention：**保证预测一个单词时只能使用此单词以前的单词信息，不能使用它本身和之后的单词信息。
3. **multi-head cross-attention：**用decoder的masked multi-head self-attention输出作为Q，encoder的multi-head self-attention输出作为K和V进行计算。
4. **feed forward network：**每个单独位置各有一个，不同位置共享参数。
5. **softmax：**计算出目标单词的概率。

在encoder和decoder中，虽然都存在名为 multi-head self-attention的结构，但它们的实现方式略有不同，具体可以分为三种，其中encoder一种，decoder两种。



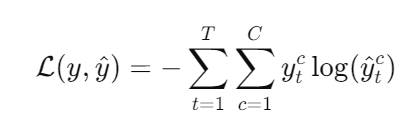
此处也有一个[batch\_size, padding\_length, padding\_length]的mask矩阵，因为输入的句子是经过padding的，真正存在的单词不应该与padding的单词之间有关联。Q与K相乘后会得到一个表示各单词间关联程度的矩阵，这个矩阵与mask矩阵element-wise相乘后再进行softmax操作，就会得到句子中归一化的各个单词间的相关程度，且屏蔽掉padding的单词。

encoder的multi-head self-attention只计算序列自身的关联性，而且可以一次性输入全部数据，不牵扯因果关系，因此较为简单。

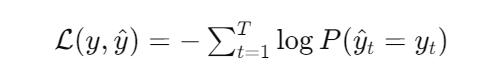
decoder的第一种multi-head self-attention名为masked multi-head self-attention。虽然它也只计算序列自身的相关性，但由于因果关系，当前的单词只能和之前的单词有关，不能和它本身和之后的有关。

decoder的第二种multi-head self-attention将decoder 的masked multi-head self-attention作为Q，而encoder的输出作为K和V，涉及到两个不同的序列。

基于不同的任务，Transformer会使用不同的损失函数，在seq2seq任务中，通常使用交叉熵损失函数：



其中T是目标序列长度，C是目标序列单词表的长度。也可以简化为：



## A Gentle Introduction to Graph Neural Networks

发表时间：2021/09/02

期刊/会议：Distill

提出了GNN（Graph Neural Network）。一个GNN的输入是一个图，输出是一个同结构的图，但是节点和边的属性会发生变化。在预测一个顶点的属性时，若该顶点没有表示自身的向量，可以将其所有邻接顶点和全局向量取出，做一个pooling，即可得到该顶点的向量，此过程称为汇聚（aggregate）。同样的，也可以由边向量汇聚出点向量，由点向量汇聚出边向量，由边或点向量汇聚出全局向量。

**信息传递：**为了更好地leverage图的机构信息，在处理边或点对应的向量时，不是直接将其向量送入MLP，而是将其与相邻元素的向量相加后，再输入MLP中进行更新，这个过程有些类似于卷积。

在对整个图进行多次信息传递后，在输入MLP时也就可以不再进行信息传递。

## Generative Adversarial Nets

发表时间：2014/06/10

期刊/会议：NIPS

提出了Generative Adversarial Nets（GAN），开创了使用有监督学习的损失函数来做无监督学习的方法，GAN在后来也成为几年中最热门的研究领域之一。

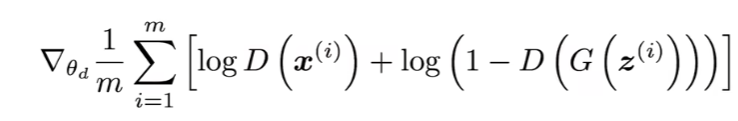
GAN最主要应用于图片生成领域，其由一个生成器和一个判别器构成，生成器与判别器都是一个简单的MLP。生成器的输入是一个符合高斯分布的向量，输出则是一个较长的图片向量。判别器的输入是一个图片向量，输出则是对该图片是否是来自数据集的一个分类。生成器的目标是不断接近于数据集的分布以骗过判别器，判别器的任务则是分辨出图片是否来自于数据集。理论上最后的结果是生成器生成的图片完全拟合于数据集，使得判别器无法分辨，因而判别器对所有图片的分类概率都是0.5。

GAN通过以下损失函数来训练生成器和判别器：

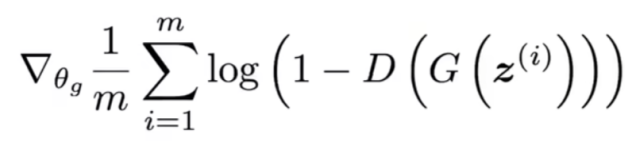


其中D代表判别器，G代表生成器，x是数据集，z是随机向量。

在一轮迭代中，先使用SGD更新k次判别器的参数：



再更新一次生成器的参数：



## DECOUPLED WEIGHT DECAY REGULARIZATION

发表时间：2019

期刊/会议：ICLR

提出了L2正则化与weight decay对Adam之类的自适应梯度算法的差异，将weight decay从梯度计算中解耦，以提高Adam的性能。使用了一种基于贝叶斯滤波视角对解耦权重衰减的解释。

# 2024/07/14

## Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

发表时间：2011

期刊/会议：JMLR

提出了Adagrad算法。Adagrad算法使用累计的梯度的平方和的开方来自适应调整学习率，但这种调整方法会使得学习率随着训练的进行不断减小，可能会造成梯度消失的问题。本文的目标实际上是解决在线学习中的优化问题，作者使用了次梯度方法来解决非光滑的优化问题，Adagrad实际上是次梯度方法的一种变体。文中使用了regret bound model来衡量算法的性能，并给出了严格的数学证明，得到了很强的regret guarantees。

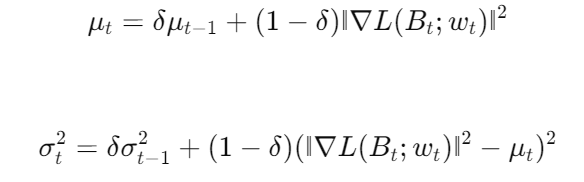
## AN ADAPTIVE POLICY TO EMPLOY SHARPNESS-AWARE MINIMIZATION

发表时间：2023

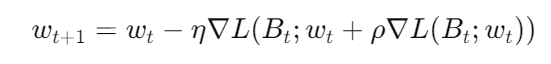
期刊/会议：ICLR

2024/07/14

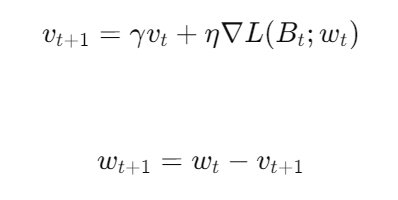
提出了一种自适应策略——AE-SAM来在SAM和SGDM之间切换，核心思想是根据损失函数景观的尖锐程度来选择优化策略，在尖锐区域使用SAM而在平坦区域使用SGDM。AE-SAM引入了一个尖锐度的量化度量来判断切换的时机，首先计算梯度的2范数的平方和其方差的指数移动平均：



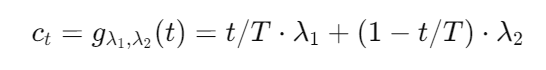
若，则使用SAM：



否则使用SGDM：



为了使算法在训练早期阶段主要使用SGDM，而在训练后期逐渐增加SAM的使用，阈值ct会随着迭代次数线性减少：



λ 1和 λ 2是预设的阈值范围，T 是总迭代次数。在模型训练初期，SGDM收敛较快，且开销较小，可找到一个较好的初始解，模型训练后期，参数接近局部最优解，SAM对平坦极值的偏好可以更好地微调参数。

# 2024/07/21

## Scaling Distributed Machine Learning with the Parameter Server

发表时间：2014

期刊/会议：OSDI

2024/07/25

提出了一个分布式系统训练机器学习模型的框架，拥有弹性扩容、容灾等特性。机器学习任务由于其大的计算量和较低的数据交互，对于分布式系统的常规任务来说是相对简单的任务。但在大语言模型兴起的时代，用分布式系统训练大模型又重新变为一个值得研究的领域。

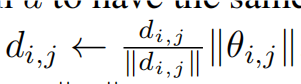
# 2024/07/28

## Visualizing the Loss Landscape of Neural Nets

发表时间：2018

期刊/会议：NIPS

2024/07/29

提出了一种名为Filter-Wise Normalization的可视化loss landscape的方法，相比于传统可视化loss landscape的方法（包括1D和2D方法），这种方法能更准确地描述最小值附近的尖锐程度，也与泛化性能呈现出很好的相关性。Filter-Wise Normalization就是对卷积层的滤波器进行的正则化。

同时本文还给出了最小值附近的Hessian矩阵的最小特真值的热度图，进一步观察平坦程度。最后利用PCA方法给出了一种更好地绘制优化器优化轨迹的方法，修正了之前方法的错误。

# 2024/08/04

## An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

发表时间：2021

期刊/会议：ICLR

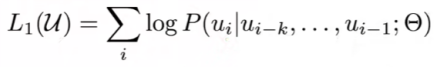
2024/7/30

提出了Vision Transformer的模型，打通了NLP和CV之间的鸿沟。虽然之前也有将自注意力机制用于CV的研究，但本文提出尽量基于原始的Transformer模型不做改动，这便使得CV和NLP的模型能够统一起来。提出了将图片划分为多个16\*16的patch的方法来缩短Transformer所要处理的序列长度，从而使得模型的训练复杂度降低到可以接受的范围。研究发现VIT在小数据集（ImageNet）上的表现不如CNN，但在大型数据集上的表现超过CNN。如何将VIT用于小样本数据也是值得研究的方向。

## Improving Language Understanding by Generative Pre-Training

发表时间：2018/06

2024/08/02

提出了GPT模型，将之前在CV领域运用的比较成熟的预训练-微调模型应用到NLP领域。将Transformer的解码器部分作为模型架构，将作为目标函数最大化。在多个任务集上的表现都优于之前的方法。GPT选择使用预测下一个单词的任务进行训练，相比于BERT完形填空的任务要难一些，但理论上随着模型的增大上限也更高。

## Language Models are Unsupervised Multitask Learners

发表时间：2019

2024/08/02

提出了GPT2。使用zero shot（无样本）方法使得通过prompt让同一个模型不经过调整就可以应对不同的任务，即直接使用预训练的模型对子任务做预测，不对子任务提供任何训练样本。使用更大的数据集训练了更大的GPT模型，取得了还不错的结果。

## Language Models are Few-Shot Learners

发表时间：2020

提出了GPT3。在模型架构上相比GPT2基本没有变化。详细对比了Few Shot、One Shot和Zero Shot之间模型的性能差距。使用大型数据集训练了一个更大的模型，最终取得了极为优秀的结果，证明了语言大模型确实是可以大力出奇迹的。

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

发表时间：2019

发表会议：NAACL-HLT 2019

提出了BERT（Bidirectional Encoder Representations from Transformers）模型。使用了Transformer的编码器架构，提出使用双向的序列处理方法。使用完形填空和下一句预测来训练模型。训练了一个较大模型，在诸多任务上表现较好。但在机器翻译等生成式任务上有局限性。

## Masked Autoencoders Are Scalable Vision Learners

发表时间：2021

发表期刊/会议：CVPR 2022

提出了MAE模型。使用和VIT类似的transformer架构，但是解码器部分较为简单。使用和VIT类似的方法将图片分割为patch序列，每个图片中只保留原来的1/4的patch，这使得任务变得比较困难。使用自监督方法进行预训练。模型经过微调之后取得了比之前框架更好的效果。

## Momentum Contrast for Unsupervised Visual Representation Learning

发表时间：2020

发表期刊/会议：CVPR 2020

提出了Moco算法，使得无监督的对比学习预训练出的模型能够和有监督学习训练出的模型性能相媲美，弥合了对比学习领域有监督学习和无监督学习的差距。Moco提出的队列式batch有效地缓解了训练数据大小受限的问题，同时momentum式的参数更新也缓解了参数一致性的问题。本文将Moco与end to end式和memory bank式的对比损失的机制与Moco式的进行比较，通过实验说明Moco有效地缓解了以上两种方式存在的两种问题，取得了良好效果。

## Learning Transferable Visual Models From Natural Language Supervision

发表时间：2021

发表期刊/会议：ICML

2024/08/09

提出了CLIP模型。CLIP模型使用文本和图片之间的对比学习作为代理任务，具体而言，使用一个图片编码器和文本编码器提取出图片和文本的特征，然后分别计算图片相对于文本的余弦相似度和文本相对于图片的余弦相似度，其中将图片和文本相对应的作为正类，其他的均为负类，将这两种相似度的平均值作为损失函数进行训练。作者自己构建了一个十分巨大的数据集。结果zero shot的CLIP就可以和许多精心设计的模型的分类精度相媲美。值得注意的是one shot的CLIP相对于zero shot的CLIP的分类精度竟然大幅下降，这显示出与人类十分不同的学习特性。CLIP的最大贡献在于突破了以往分类任务对于给定类别的依赖性，使得使用同一个CLIP模型可以进行任意给定的分类任务，这一特点源于CLIP模型对于语义和图片之间关联性的深度理解。但是CLIP也存在很多局限性，它在泛化能力上还有待研究，例如它在MNIST上的准确率只有88%，这甚至比不过一个简单的MLP模型。

## Two-Stream Convolutional Networks for Action Recognition in Videos

发表时间：2014

发表期刊/会议：NIPS

2024/08/09

提出了双流网络。使用两个卷积神经网络分别抽取视频的空间和光流信息，最后根据两个网络的输出进行视频分类。使得深度学习在视频分类领域的准确率能够和当时最好的手工分类方法相当，证明了深度学习在视频分类中的有效性。成为深度学习在视频分类领域的开创性工作。

## Limited Data, Unlimited Potential: A Study on ViTs Augmented by Masked Autoencoders

发表时间：2023

发表期刊/会议：IEEE/CVF

2024/08/12

提出了一种将分类任务作为基础任务和自监督任务作为辅助任务相结合的训练方法，使得VIT的性能在小数据集上有了显著提高。其中基础任务就是原始的分类任务，使用交叉熵作为损失函数，辅助任务使用了MAE中的patch预测任务，是一种自监督任务，使用均方误差作为损失函数。

# 2024/08/18

## Highly accurate protein structure prediction with Alphafold

发表时间：2021

发表期刊/会议：Nature

2024/08/14

提出了Alphaflod2，在Alphafold1的基础上构建了一个非常庞大的模型，使得蛋白质结构的预测误差降低到了原子级别。

## Evaluating Large Language Models Trained on Code

发表时间：2021

2024/08/14

Openai使用GPT模型，从github上下载大量代码进行训练，以达到能自动写代码的效果。创造了一个名为human eval的数据集，包含了docstring和对应的单元测试，用以检测生成的代码能否通过这些测试。虽然这篇工作新意不高，但是试图解决一个较大的问题，也就是自动生成代码。

## End-to-End Object Detection with Transformers

发表时间：2020

发表期刊/会议：ECCV

2024/08/15

提出了DETR框架。使用Transformer作为模型结构。提出了使用可学习的object query代替以往目标检测任务中的anchor。使用二分图匹配替代了原有的nms，使得模型能够端到端地进行目标检测。DETR因为其简单性而被后续众多工作进一步研究，成为一个里程碑式的工作。

## ViLT: Vison-and-Language Transformer Without Convolution or Region Supervision

发表时间：2021

发表期刊/会议：ICML

2024/09/20

使用一个简单的线性层完成对图像数据和文本数据的特征提取，大大提高了计算效率，使用Transformer模型实现特征的融合，在获得有竞争力的表现的基础上降低了特征提取的复杂度。是对传统VLP模型效率的一种提升。

# 2024/10/20

## Coevolutionary Neurodynamic Optimization with Gradient Norm Penalty for Generalization Improvement

发表时间：2022

发表期刊/会议：IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

2024/10/17

介绍了一种改进的协同进化神经动力学优化方法，称为带梯度范数惩罚的协同进化神经动力学优化（CNOGNP）。神经动力学方法提供了局部精确搜索能力，而PSO具有强大的全局搜索能力。协同进化神经动力学优化（CNO）结合了神经动力学方法和PSO的优点，拥有较强的局部和全局搜索能力，在此基础上，文章引入了梯度范数惩罚机制来增强该方法的泛化性能。

## Noise-Suppressing Neural Algorithm for Solving Time-Varying System of Linear Equations: A Control-Based Approach

发表时间：2019

发表期刊/会议： IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS

2024/10/17

在传统ZNN的基础上结合控制理论的观点，提出了一种能够在多种噪声的存在下求解离散时变线性方程组的方法。本文将连续时间递归神经网络和离散时间神经算法的建立从传统的优化（或计算）角度扩展到控制理论角度。理论分析和数值实验证明了本文的方法能够在噪声的存在下保证求解性能。

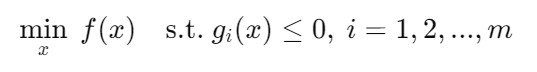
## 40.A Survey of Neurodynamic Optimization

发表时间：2024

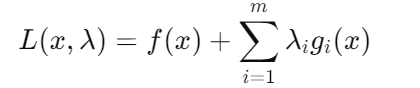
发表期刊/会议：IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE

2024/10/18

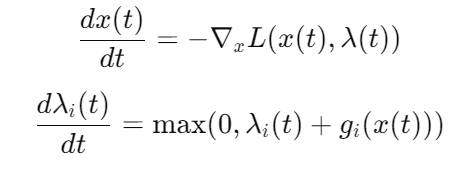
假设我们有一个标准的优化问题：



其中，f(x) 是目标函数，g\_i(x) 是不等式约束条件。这类问题可以通过拉格朗日乘子法表示为：



神经动力学优化通过递归神经网络动态调整x和λ，逐步逼近最优解。神经动力学优化使用的微分方程模型可以视为一种递归结构。状态变量x(t)和λ(t) 随时间t演化,状态变量 x(t) 和 λ(t) 随着时间t 的变化满足以下微分方程：



这种模型的动力学设计确保神经网络的状态不断趋向于优化问题的解。

## 41.ADABATCH: ADAPTIVE BATCH SIZES FOR TRAINING DEEP NEURAL NETWORKS

发表时间：2018

发表期刊/会议：arxiv

2024/10/19

提出了一种逐步扩大batch size来提高训练效率的方法，在cifar10hecifar100数据集上作者使用SGDM，每20个epoch将batch size增大为2倍，并将学习率降低为0.75倍，相应的，对于没有使用自适应batch size的SGDM，作者则将学习率每20个epoch降低为0.375倍。对于Imagenet数据集，作者使用SGDM，每30个epoch将batch size增大为2倍，并将学习率降低为0.2倍，对于固定batch size的baseline，则将学习率降低为0.1倍。这种方法可以在不太影响准确率的基础上降低训练时间。

# 2024/10/27

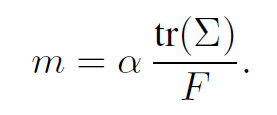
## 42.Coupling Adaptive Batch Sizes with Learning Rates

发表时间：2017

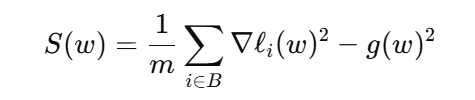
发表期刊/会议：33rd Conference on Uncertainty in Artificial Intelligence (UAI)

2024/10/24

提出了种动态调整批量大小的方法CABS，该方法根据梯度方差自适应调整，并与学习率耦合。其调整batch size 的方法为：



其中m为自适应batch size，tr（Σ）为梯度的方差矩阵的迹，F为目标函数。对于tr（Σ），在每个mini batch中计算梯度方差的估计值S，然后使用指数移动平均对方差估计进行平滑处理：



黑色的钟表

低可信度描述已自动生成

CABS方法得到的自适应batch size与学习率大小和梯度方差成正比而与损失值成反比。也就是在学习率较大时，batch size也较大，这一策略的直觉是：如果在小批量下使用过大的学习率，可能会使得梯度噪声的影响较大，导致优化过程不稳定；在梯度的方差较大时，更大的批量大小会带来更稳定的梯度估计；损失值较大时，通过减小批量大小可以使得参数跳出局部极小值。不过计算梯度方差的估计值带来的计算开销较大，而且本文也没有结合学习率衰减进行实验。

## 43.Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

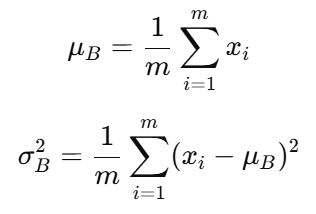
发表时间：2015

发表期刊/会议：ICLR

2024/10/26

神经网络中每一层的输出数据在训练过程中会发生变化（称为内部协变量偏移Internal Covariate Shift），导致收敛速度变慢。作者提出批量归一化（Batch Normalization, BN）技术，通过在每一小批训练样本上归一化输入，从而稳定每一层输出数据的分布，此方法可以使模型用更大的学习率训练，减少对参数初始化的敏感性，减少drop out的敏感性，提高训练速度。

对于每一层的每一小批数据（mini-batch），其中m为小批的大小。对这个小批数据进行归一化处理时，BN会计算该小批的均值和方差：



每个输入值xi都使用计算得到的均值和方差进行标准化，使其具有零均值和单位方差：

图示

描述已自动生成

归一化后的输出通过一个可学习的缩放参数𝛾和偏移参数𝛽进行线性变换，以保持网络的表达能力：



这样，通过调整 γ和 𝛽，可以让该层学习到不同的分布。此步骤的目的是确保在不同层、不同批次中的输入不会因为归一化失去其原有的分布信息。

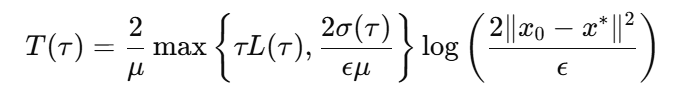
## 44. Adaptive Learning of the Optimal Batch Size of SGD

发表时间：2020

发表期刊/会议：NeurIPS OPT ML Workshop

2024/10/27

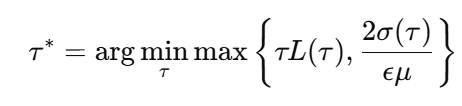
提出了一种基于梯度平滑性（L-smoothness）和梯度噪声（σ）的计算最优batch size的策略。在每次迭代中，SGD的总迭代复杂度是批量大小和迭代次数的乘积。文章基于梯度平滑性和梯度噪声，推导了一个关于批量大小 τ 的迭代复杂度下界公式：



其中 T(τ)表示总迭代复杂度,𝜇表示函数的强凸性常数，L(τ) 和 σ(τ) 分别是平滑性常数和噪声，依赖于批量大小𝜏。公式中的𝜖代表收敛到最优解的距离。

为了找到最优批量大小，作者分析了如何使得总迭代复杂度公式最小化。由于公式中的

𝜏L(τ) 和 σ(τ) 是 τ 的分段线性函数，最优批量大小𝜏∗ 可以通过最小化下面的函数来求得：



这个最小化过程考虑了批量大小对平滑性和噪声的不同影响，从而推导出在给定采样策略下的最优批量大小。

# 2024/11/03

## 45.GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks

发表时间：2018

发表期刊/会议：International Conference on Learning Representations

2024/11/01

提出了一种对于多任务的梯度正则化策略，将多任务训练时的梯度变为同样的分布，防止了多任务训练中某一任务对于参数影响过大的问题。

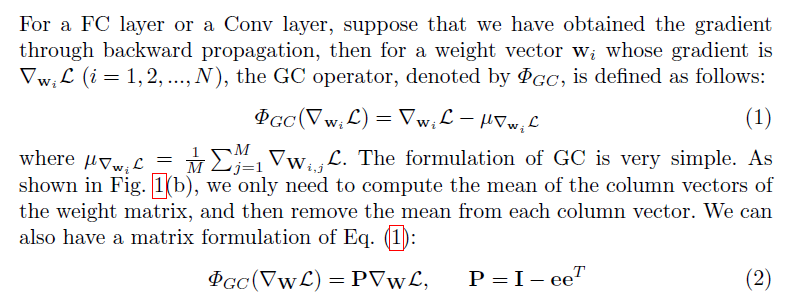
## 46.Gradient centralization: A new optimization technique for deep neural networks

发表时间：2020

发表期刊/会议：ECCV

2024/11/01

提出了对梯度进行中心化来提高泛化性的方法，即在每次更新时给当前对每一层的梯度减去当前批次该梯度的均值。



## 47.ADDING GRADIENT NOISE IMPROVES LEARNING FOR VERY DEEP NETWORKS

发表时间：2016

发表期刊/会议：arxiv

2024/11/1

在训练过程中给梯度增加噪声，并使得噪声的方差衰减，从而可以得到更好的泛化性能。

# 2024/11/17

## 48.Zeroing neural networks: A survey

发表时间：2017

发表期刊/会议：Neurocomputing

2024/11/13

文章介绍了ZNN是一种求解时变问题的有效方法，并对比了不同的ZNN变体，给出了ZNN的稳定性和收敛性分析，并介绍了ZNN的应用领域。文章指出ZNN有连续时间ZNN模型和离散时间ZNN模型两类。ZNN在方程求解、机器人控制和混沌系统等领域都有应用。

# 2024/12/01

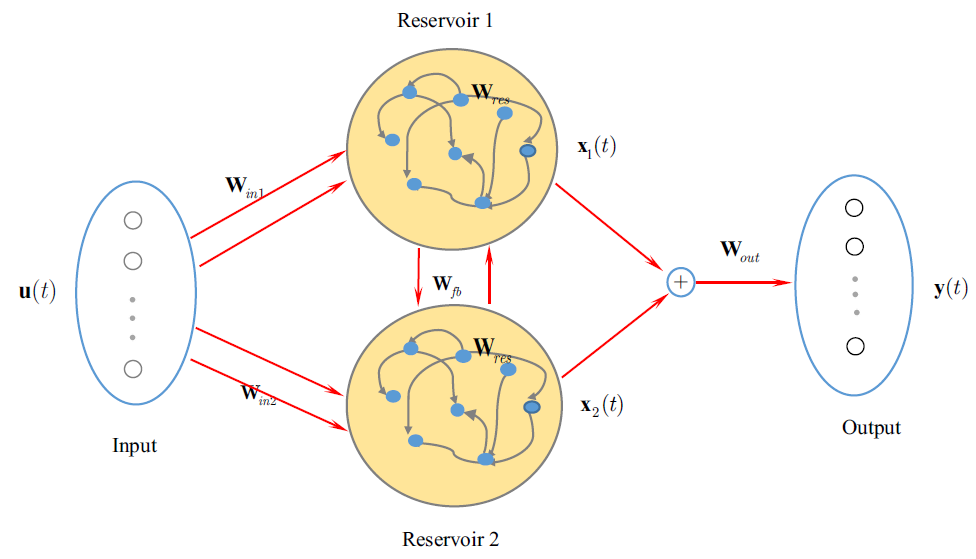
## 49.An echo state network with interacting reservoirs for modeling and analysis of nonlinear systems

发表时间：2024

发表期刊/会议：Nonlinear Dynamics

2024/11/26

文章提出一种名为InterESN的有多存储池的ESN模型。该模型使用一个Winter的稀疏矩阵使得两个存储池之间得以交互，具体交互方式为：



文本, 信件

描述已自动生成

其中f1和f2作者使用了cos和sin而不是常用的tanh。同时作者给出了InterESN的ESP性质证明和memory capacity的证明。

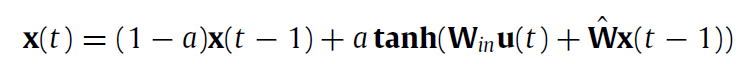
## 50.Design of deep echo state networks

发表时间：2018

发表期刊/会议：Neural Networks

2024/12/01

提出了基于信号频率分析的DeepESN设计方法，以解决如何确定深度递归网络的层数的问题。该方法优点在于无需多次尝试不同的ESN层数，减少了超参数确定的计算量。一个传统的ESN的状态转移方程为：



其中α为泄露参数（leaky parameter）。而一个深度ESN的结构为：

图示

描述已自动生成

每一层的状态转移方程为：

文本, 信件

描述已自动生成

文本

描述已自动生成是输入矩阵，是第l层的存储池矩阵， 是各层x纵向堆叠起来的向量，文本

描述已自动生成是输出层向量。

本文提出的基于频谱的层数确定方法目标是通过分析储层状态信号的频谱特性（尤其是频谱中心和频谱扩展）来判断增加新层是否能有效提升网络对时间动态的捕捉能力。当新层的加入对频谱特性的影响变得微乎其微时，算法终止，确定最终的层数。

# 2024/12/08

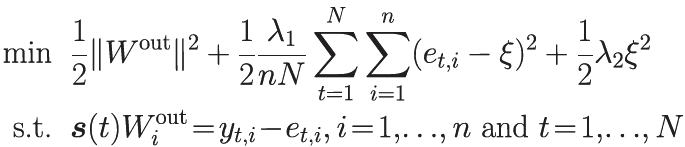
## 51. Echo State Network With Probabilistic Regularization for Time Series Prediction

发表时间：2023

发表期刊/会议：IEEE/CAA JOURNAL OF AUTOMATICA SINICA

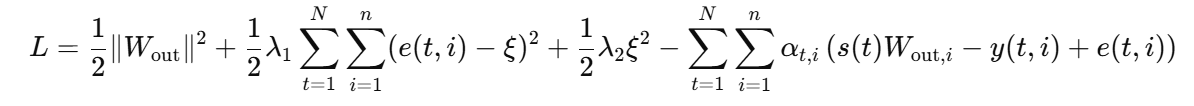
2024/12/7

本文提出一种目标函数用于训练Wout，这种目标函数通过最小化建模误差的均值和方差来提高模型的抗噪性。为最小化建模误差的均值和方差，需解决以下优化问题：



其中目标函数的第一项是对权重大小的惩罚项，第二项最小化建模误差e(t)与一个全局误差因子ξ之间的差异。这个误差因子ξ是用来协调训练样本之间的误差的。通过调整ξ，PRESN能够控制误差的分布，使其更符合真实系统的误差分布，第三项最小化误差的均值，即通过将误差的均M(e(t)与真实系统的误差均M(ϵ)对齐，确保模型不仅在方差上与实际系统一致，而且能够处理偏差。约束条件的含义是，模型的输出应该等于真实目标值 y(t,i)减去误差项e(t,i)。

为了求解上述目标函数和约束条件，PRESN 使用了拉格朗日乘子法来优化目标函数。拉格朗日函数的形式如下：



从而可以得出Wout的解为：

文本

描述已自动生成

## 52. Growing Echo State Network With an Inverse-Free Weight Update Strategy

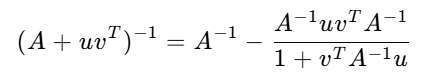
发表时间：2023

发表期刊/会议：IEEE TRANSACTIONS ON CYBERNETICS

2024/12/07

本文提出一种求解Wout的方法，使用Sherman-Morrison公式和Schur补降低了矩阵求逆的计算量。

Sherman-Morrison公式允许在不直接计算矩阵逆的情况下更新矩阵的逆：



Schur补是在分块矩阵中，当某个块可逆时，可以用其补块来表示整个矩阵的逆。

文本

中度可信度描述已自动生成

D的Schur补为，这个概念在IFESN中用于处理递归网络中的动态更新，从而避免了直接对整个状态矩阵进行求逆。

## 53. A Collective Neurodynamic Approach to Constrained Global Optimization

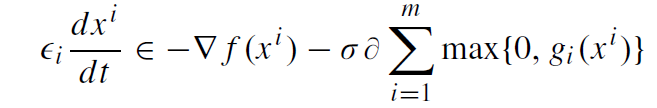
发表时间：2017

发表期刊/会议：IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

2024/12/08

这篇文章提出了一种新的集体神经动力学方法，用于求解带约束的全局优化问题。其核心思想是模拟大脑风暴的工作机制，结合神经网络与粒子群优化技术，以提高优化问题求解的效率与精度。本文中单层RNN被用来寻找优化问题的KKT点，文章结合了PSO算法，使用多个单层RNN在不同区域同时对优化问题的解进行寻找，并不断更新全局最优点。文章也给出了该算法的收敛性证明，这种集体神经动力学方法能够以概率 1 计算出全局最优解。

每个点的更新方法为：

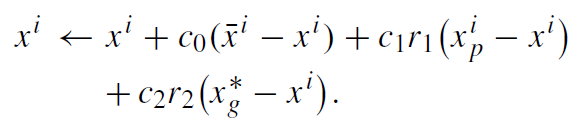


接着更新局部最优解和全局最优解：

文本, 信件

描述已自动生成

根据局部最优解和全局最优解再更新各个粒子的状态：



计算粒子的多样性：

钟表的特写

描述已自动生成

若Div<δ，则需要施加小波变异：

钟表的特写

中度可信度描述已自动生成

手机屏幕截图

中度可信度描述已自动生成

重复迭代，直到满足迭代次数或其它停止条件时停止迭代。