

Workshop 3

Outlines

- Improved conditioning
- Optimization
- Scale invariant property
- Imposed Constraints
- Representation
- Normalization

Decomposition of DNNs

normalization
gradient vanishing / explosion
activation vanishing / explosion
网络训练能力差
模型调整方式: 初始化
显式调整方式: normalization

Normalization 定义

- Definition of normalization
 - In statistics: adjustments of values or distributions in statistics
 - In image processing: changing the range of pixel intensity values
 - In data processing: general reduction of data to canonical form
- Definition of normalization in this report
 - Given a set of data $D = \{x^{(i)}\}_{i=1}^N$, the normalization operation is a function $\Phi: x \mapsto \tilde{x}$, which ensures that the transformed data $\tilde{D} = \{\tilde{x}^{(i)}\}_{i=1}^N$ has certain statistical properties.

$$\tilde{x} = \frac{x - \mu}{\sigma}$$

基本 normalization 方式:

三种典型层

BN

B/V motivation

Normalization Operation

- Basic normalization operations
 - Centering
 - Scaling
 - Decorrelating
- Combine above
 - Standardization
 - Whitening

BN

Motivation of Normalizing Input

- Improve the effects of learning
 - Non-parameter models (KNN, Kernel SVM)
 - Distance: Similarity
- Improve optimization efficiency
 - Parametric model (logistic regression)
 - Update parameters iteratively

Normalizing Input Benefits Optimization

- Linear regression: $\mathcal{L}(F(x), \hat{y}) = \frac{1}{2}(Wx - \hat{y})^2$
- Where $C = \frac{1}{N} \sum_{i=1}^N x^T x$ is the covariance matrix
- Gradient: $\frac{\partial \mathcal{L}}{\partial W} = \sum_{i=1}^N x(y - \hat{y})$
- Hessian matrix: $H = \frac{\partial^2 \mathcal{L}}{\partial W \partial W} = C$

Towards Normalizing Activations of DNNs

- Difficulty of analysis for DNNs
 - Nonlinear model
 - X is only linearly connected by $W^{(1)}$; Optimization is over θ , not $W^{(1)}$ only
- What we can exploit?
 - Layer-wise structure
 - $h^{(l)}$ is linearly connected by $W^{(l+1)}$

Normalizing Activations

Normalizing Input Benefits Optimization

- Learning dynamics are controlled by the spectrum of curvature matrix (Hessian H)
 - $\lambda_{\max}(H)$: Optimal learning rate: $\eta = \frac{1}{\lambda_{\max}(H)}$
 - Diverge if $\eta > \frac{1}{\lambda_{\min}(H)}$
- Condition number $K = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}$ control the iterations required for convergence
- Hessian of multiple output: $H = \mathbb{E}_D(x x^T) \otimes I$

Construct Well-Conditioned Landscape

- Denoting $\Sigma_x = \mathbb{E}_{p(x)}(xx^T)$ and $\Sigma_{\nabla h} = \mathbb{E}_{p(x), q(y|x)}(\frac{\partial \ell}{\partial h} \frac{\partial \ell}{\partial h})$
- Criteria
 - 1. The statistics of the layer input (e.g., Σ_x) and output-gradient ($\Sigma_{\nabla h}$) across different layers are equal (across layer)
 - 2. Σ_x and $\Sigma_{\nabla h}$ are well conditioned (in layer)
- Initialization techniques: designed to satisfy Criteria 1 and/or 2 during initialization
 - Axiv-Init [Glorot and Bengio, 2010], He-Init [He et al, 2015]; for Criteria 1
 - Orthogonal Initialization [Saxe et al, 2014]; for Criteria 1 and 2
- General goals of "normalization" in DNNs: controlling the distribution of activations/outputs gradients during training

activation outputs

Normalization by Population Statistics

- Centering the activation
 - Montavon et al, 2014; Wiesler et al, 2014
- Standardizing the activation: centering + scaling
 - Wiesler et al, 2014
- Whitening the activations
 - Desjardins et al 2015; Luo, 2017

Normalization by Population Statistics

- Advantages
 - Well exploit the beneficial property of normalization in optimization
- Drawbacks
 - Training instabilities
 - The estimation is not accurate (sampled data)
 - Internal covariant shift (the distribution of activation varying with training progressing)
 - Can not be used to large networks
 - An inaccurate estimation of population statistics will be amplified as the layers increase

Normalizing activations

- Machine learning/Optimization community:
 - Population statistics of a dataset
- Computer vision community:
 - Local statistics in an sample

$\Sigma_x = \mathbb{E}_{p(x)}(xx^T)$

Timeline: Montavon et al, 2012; Wiesler et al, 2014; Desjardins et al, 2015; Luo, 2017; Krizhevsky et al, 2012; Ren et al, 2017; Ortiz et al, 2020

Local Normalization in a Sample

- Local normalization
 - Local contrast normalization [Jarrett et al, ICCV 2009]
 - Local response normalization [Krizhevsky et al, NeurIPS 2012]
 - Divisive normalization [Ren et al, ICLR 2017]
 - Local context normalization [Ortiz et al, CVPR 2020]

Given an example $X \in \mathbb{R}^{C \times H \times W}$

Batch Normalization (BN)

Input: Values of x over a mini-batch: $S = \{x_{1 \dots m}\}$
 Output: $\{y_i = BN_{\gamma, \beta}(x_i)\}$

```


$$\begin{aligned}
    \mu_B &\leftarrow \frac{1}{m} \sum_{i=1}^m x_i && // \text{mini-batch mean} \\
    \sigma_B^2 &\leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 && // \text{mini-batch variance} \\
    \bar{x}_i &\leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} && // \text{normalize} \\
    y_i &\leftarrow \gamma \bar{x}_i + \beta && // \text{scale and shift}
  \end{aligned}$$


```

Extra learnable scale and bias

Normalize over mini-batch data

Back-propagate through the transformation

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^m \frac{\partial \ell}{\partial x_i} \cdot (x_i - \mu_B) \cdot \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \cdot (\sigma_B^2 + \epsilon)^{-1/2}$$

$$\frac{\partial \ell}{\partial \mu_B} = \left(\sum_{i=1}^m \frac{\partial \ell}{\partial x_i} \cdot \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \frac{2}{m} \sum_{i=1}^m (x_i - \mu_B)$$

$$\frac{\partial \ell}{\partial \sigma_B^2} = \frac{2}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \cdot \frac{\partial \ell}{\partial \sigma_B^2}$$

Batch Normalization (BN)

- Extension to CNN input
 - Jarrett et al, 2009; Gulyehre and Bengio, 2013

Spatial location as an example for normalization

$X \in \mathbb{R}^{N \times C \times H \times W}$

Other Partitions for Normalization

- MLP input: $X \in \mathbb{R}^{N \times C}$

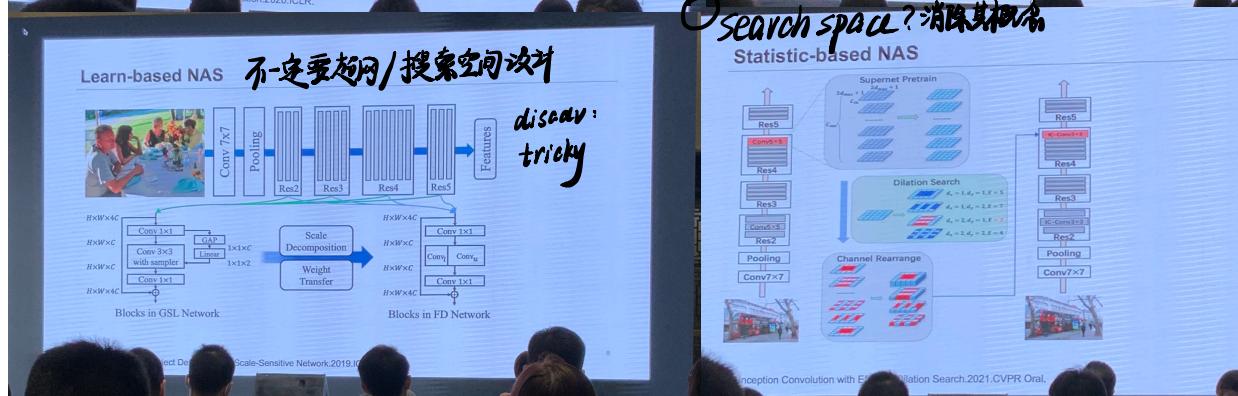
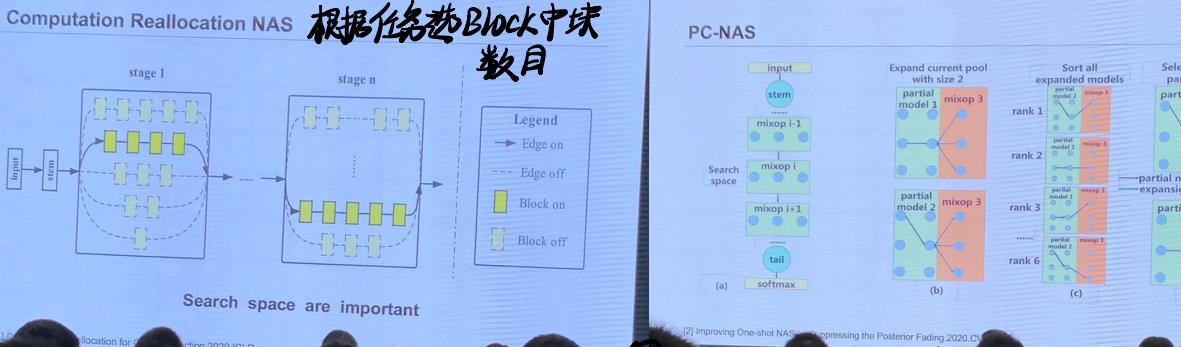
Batch Norm

Layer Norm

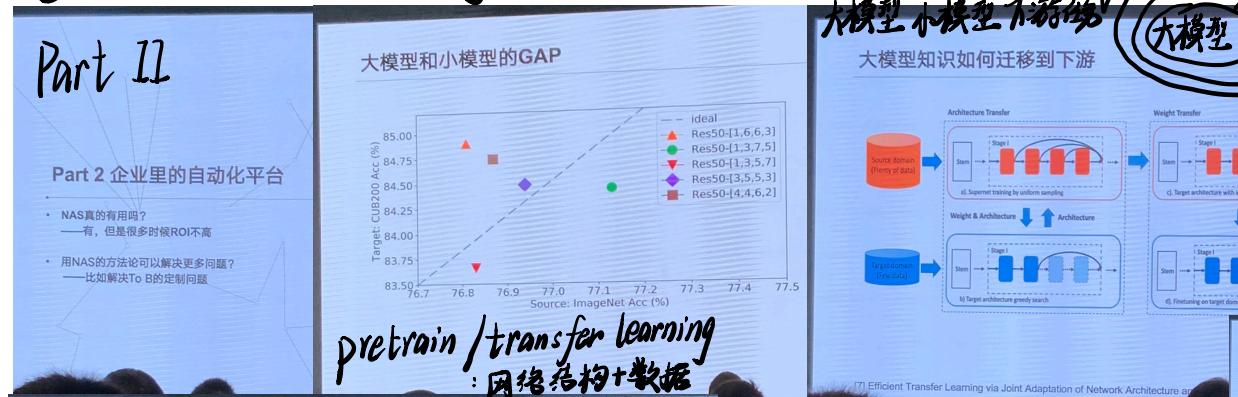
Group Norm

Batch Group Norm

Workshop 4 全生命周期的自动化机器学习

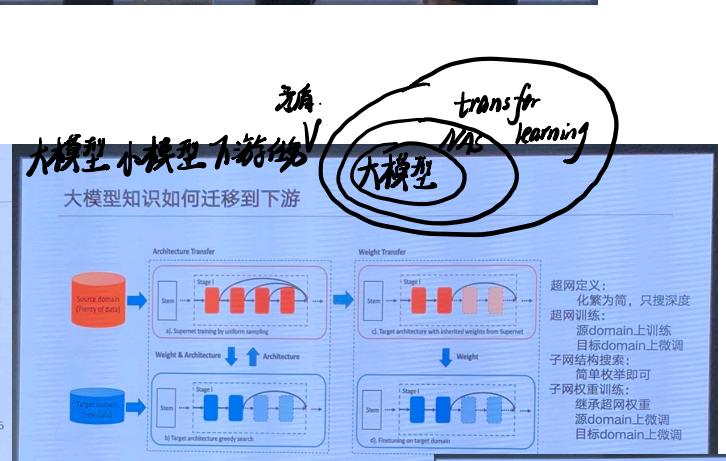


Beyond NAs } Auto sampling
} Auto Aug.



The diagram illustrates the user-centered automation model production line process:

- User Requirements:** The process begins with a user providing requirements. A callout box notes: "我们的产品要在arm上用，可以比10系列快点，比11系列慢20%，主要用在XXX场景，有YYY等bad cases可能会出现。没人比我更懂我的需求。"
- Platform Pre-processing:** The user requirements are used to generate a platform configuration. This step involves:
 - 选择类别 (Select Category)
 - 选择平台及 Latency要求 (Select Platform and Latency Requirements)
 - 根据平台样例从 Domain Pools 勾选Domains (Select Domains based on platform examples from Domain Pools)
- Platform Configuration:** The configuration includes:
 - 根据平台, Latency要求查表选择指定类别下的优先结构 (Based on platform, Latency requirements, select priority structures under specified categories)
 - 选择核心Domain 数据及用户指定 Domains构成 Fast-FT数据集 (Select core Domain data and user-specified Domains to form a Fast-FT data set)
- Packaging:** The final step involves packaging the selected structure and data into a Fast-FT format.



Spotlight
1. Condense V2