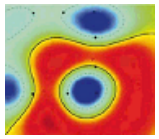
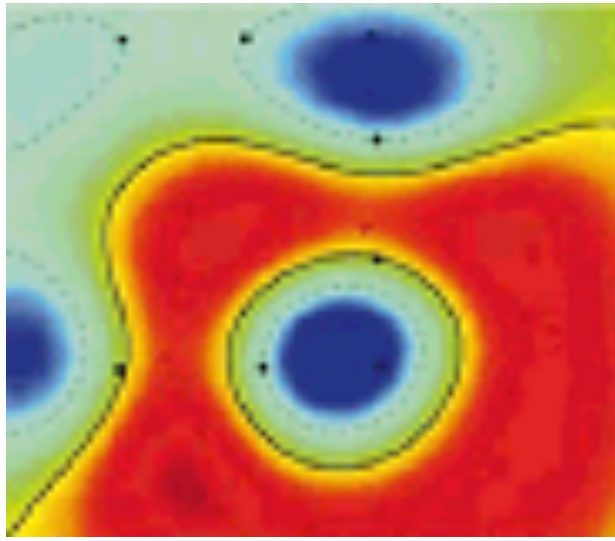


WiSe 2024/25

Deep Learning 1

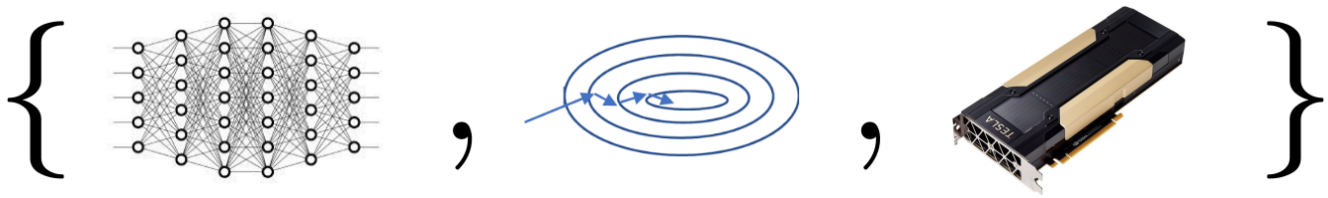
Technische Universität Berlin



Lecture 5

Overfitting & Robustness (1)

Recap Lectures 1–4



Lectures 1–4:

- With flexible neural network architectures, powerful optimization techniques, and fast machines, we have means of producing functions that can accurately fit large amount of highly nonlinear data.

Question:

- Do the learned neural networks generalize to new data, e.g. will it be able to correctly classify *new* images?

The data on which we train the model also matter!

A Bit of Theory

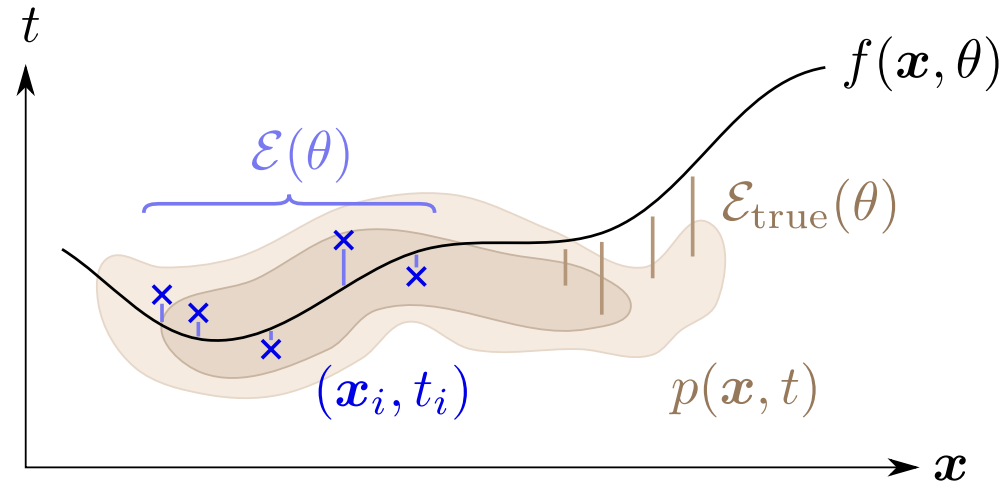
So far, we have only considered the error we use to optimize the model (aka. the training error):

$$\mathcal{E}_{\text{train}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i, \theta) - t_i)^2$$

In practice, what we really care about is the *true* error:

$$\mathcal{E}_{\text{true}}(\theta) = \int (f(\boldsymbol{x}, \theta) - t)^2 \, p(\boldsymbol{x}, t) \, d\boldsymbol{x}dt$$

where $p(\boldsymbol{x}, t)$ is the *true* probability distribution from which the data is coming at test time. The true error is much harder to minimize, because we don't know $p(\boldsymbol{x}, t)$.



Characterizing Datasets

Factors that makes the available dataset $\mathcal{D}$ and the true distribution $p(\boldsymbol{x}, t)$ diverge:

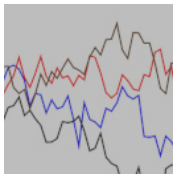
- The fact the dataset is composed of few data points drawn randomly from the underlying data distribution (**finite data**).
- The fact that the dataset may overrepresent certain parts of the underlying distribution, e.g. people of a certain age group (**dataset bias**).
- The fact that the dataset may have been generated from an underlying distribution $p_{\text{old}}(\boldsymbol{x}, t)$ that is now obsolete (**distribution shift**).

图1：数据集 D 和真实分布 $p(x,t)$ 分歧的因素：

1. 数据集是从底层数据分布中随机抽取的少量数据点组成的（有限数据）。
2. 数据集可能过度代表了底层分布的某些部分，例如某个年龄组的人群（数据集偏差）。
3. 数据集可能来源于现已过时的底层分布 $p_{\text{old}}(x,t)$（分布漂移）。

Practical Examples

图像/文本数据	每类可能有数千张图片，由多个来源聚合而成。一些图像组成可能被过度代表（数据集偏差/虚假相关性）。
传感器数据	数据集可能非常大，但传感器可能会随着时间推移失去校准（分布漂移）。
游戏（如围棋、象棋）	通过计算机模拟可以产生无限数量的游戏状态，但大师级的对局生成成本较高，简单对局可能被过度代表（数据集偏差）。

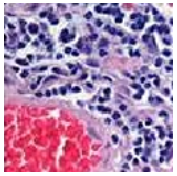
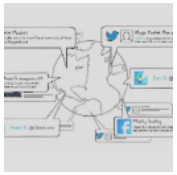
	Data types	Properties
	Image/text data	thousands of images per class, aggregated from many sources. Some image compositions may be overrepresented (**dataset bias / spurious correlations**).
	Sensor data	Potentially very large datasets, but sensors may become decalibrated over time (**distribution shift**).
	Games (GO, chess, etc.)	Infinite number of games states can be produced through computer-computer plays, but master-level plays being more expensive to generate, simple games may be overrepresented (**dataset bias**).

Overrepresented（过度代表）是指在数据集中，某一类别、特征或群体出现的频率显著高于其在真实分布中的频率。这种现象可能导致模型对这一部分数据的过度偏好，从而影响模型的泛化能力和对整体数据的公平性。

举例：如果在一个用户行为分析的数据集中，20-30岁的人群占了80%，而实际人口分布中这个年龄段的人只占30%，这就表示该年龄段在数据集中被过度代表了。

Practical Examples (cont.)

模拟数据（如物理、驾驶）	理论上无限，但由于运行模拟的成本，实际产生的数据有限（有限数据）。从模拟环境转移到现实世界可能有困难（分布漂移）。
医疗数据	由于某种疾病的稀有性或法规约束，患者数量有限（有限数据，数据集偏差）。数据采集设备可能会随着时间演变（分布漂移）。
社交数据	数据量大，但只有最近的数据是相关的，存在未捕捉到最新趋势的风险（分布漂移）

	Data types	Properties
	Simulated data (e.g. physics, car driving)	Theoretically infinite, but practically limited do the cost of running simulations. In practice, we only generate few instances (**finite data**). Problem of transferring from a simulated to a real-world environment (**distribution shift**).
	Medical data	limited number of patients due to rarity of a particular disease, or regulatory constraints (**finite data**, **dataset bias**). Aquisition devices may evolve over time (**distribution shift**).
	Social data	Large amount of data, but only recent data is relevant. Risk of not capturing the most recent trends (**distribution shift**).

Outline

The Problem of Finite Data

- The problem of overfitting
- Mitigating overfitting

Dataset Bias 1: Imbalance Between Subgroups

- Data from Multiple Domains
- Building a 'Domain' Invariant Classifier

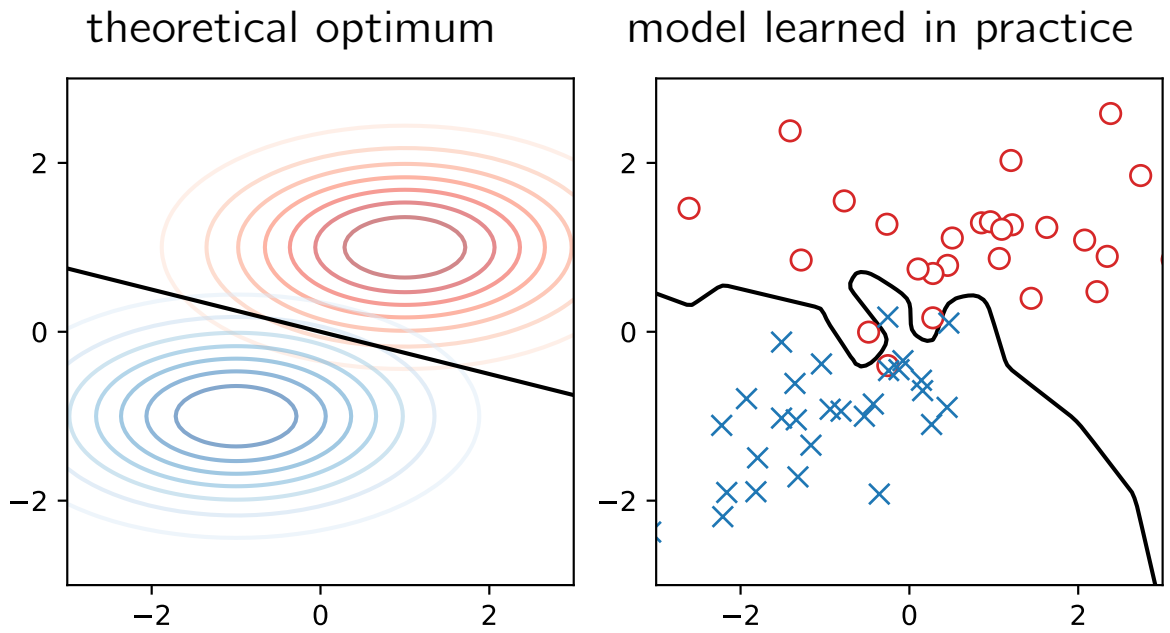
Dataset Bias 2: Spurious Correlations

- Examples of Spurious Correlations
- Detecting and Mitigating Spurious Correlations

Part 1

The Problem of Limited Data

Finite Data and Overfitting



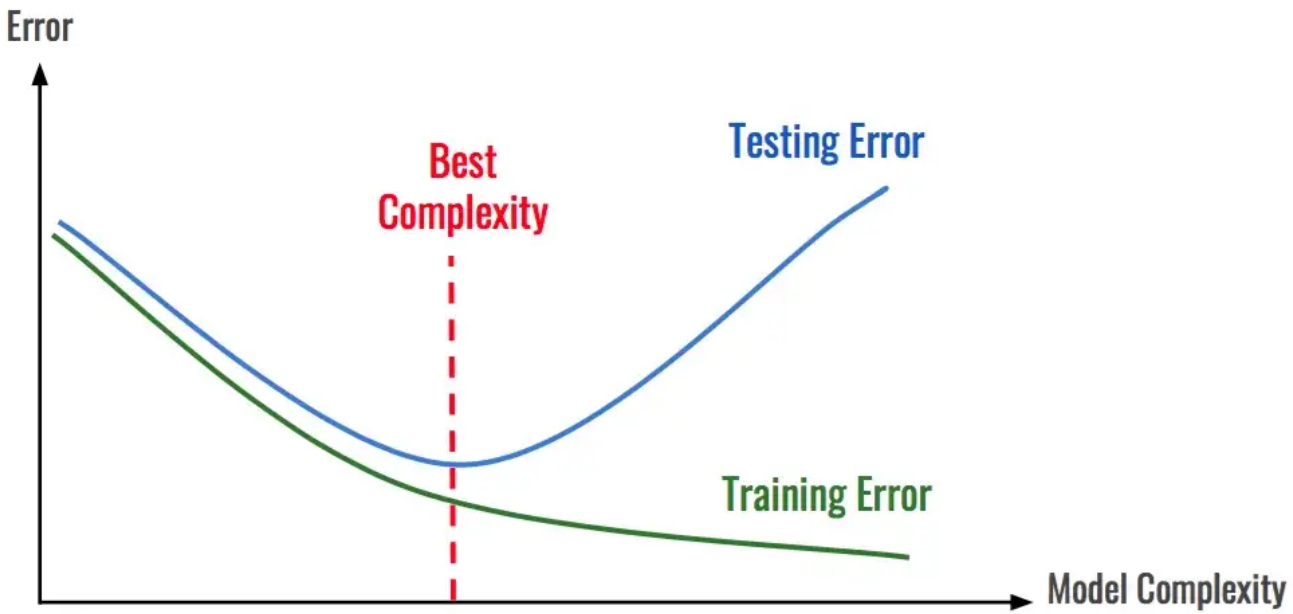
- Assume each data point $\boldsymbol{x} \in \mathbb{R}^d$ and its label $y \in \{0, 1\}$ is generated iid. from two Gaussian distributions.
- With limited data, one class or target value may be locally predominant 'by chance'. Learning these spurious variations is known as overfitting.
- An overfitted model predicts the training data perfectly but works poorly on new data.

- 假设每个数据点 $x \in \mathbb{R}^d$ 及其标签 $y \in \{0, 1\}$ 是从两个高斯分布中独立同分布生成的。
- 由于数据有限，某一类别或目标值可能会在局部区域占优，这种随机变化被学习到的情况称为**过拟合**（overfitting）。
- **过拟合的模型可以完美拟合训练数据，但在新数据上的表现很差。**

Model Error and Model Complexity

William of Ockham (1287-1347)
Linked model complexity to how suitable the model is for explaning phenomena. “Entia non sunt multiplicanda praeter necessitatem”

Vladimir Vapnik
Showed a formal relation between model complexity (measured as the VC-dimension) and the error of a classifier.



Complexity and Generalization Error

this slide has been updated

Generalization Bound [Vapnik]

Let h denote the VC-dimension of $\mathcal{F}$. The difference between the true error $R[f]$ and the traning error $R_{\text{emp}}[f]$ is upper-bounded as:

$$\mathcal{E}_{\text{true}}(\theta) - \mathcal{E}_{\text{train}}(\theta) \leq \sqrt{\frac{h(\log \frac{2N}{h} + 1) - \log(\eta/4)}{N}}$$

The VC-dimension h defines the complexity (or flexibility) of the class of considered models.

Factors that reduce the gap between test error $\mathcal{E}_{\text{true}}(\theta)$ and training error $\mathcal{E}_{\text{train}}(\theta)$:

- Lowering the VC-dimension h.
- Increasing the number of data points N.

Generalization Bound [Vapnik]

- Assume the training data is drawn iid. from the true data distribution.
- Let h denote the VC-dimension, a measure of complexity (or flexibility) of the considered class of models (e.g. one-layer networks).
- Let θ be the parameter of the learned model.
- The difference between the true error $\mathcal{E}_{\text{true}}(\theta)$ and the training error $\mathcal{E}_{\text{train}}(\theta)$ is upper-bounded as:

$$\mathcal{E}_{\text{true}}(\theta) - \mathcal{E}_{\text{train}}(\theta) \leq \sqrt{\frac{h(\log \frac{2N}{h} + 1) - \log(\eta/4)}{N}}$$

Factors that reduce the gap between training and true error:

- Lowering the VC-dimension h (i.e. choosing a less flexible class of models).
- Increasing the number of training points N.

泛化界限（Vapnik）

- 假设训练数据是从真实数据分布中独立同分布 (i.i.d.) 抽取的。
- 设 h 表示 VC 维度（模型类别复杂度或灵活性的度量，例如单层神经网络）。
- 设 θ 为学习到的模型参数。
- 真实误差 $\mathcal{E}\text{true}(\theta)$ 与训练误差 $\mathcal{E}\text{train}(\theta)$ 之间的差异被如下公式上界：

$$\mathcal{E}\text{true}(\theta) - \mathcal{E}\text{train}(\theta) \leq \sqrt{\frac{h\left(\log \frac{2N}{h} + 1\right) - \log(\eta/4)}{N}}$$

减少训练误差与真实误差之间差距的因素：

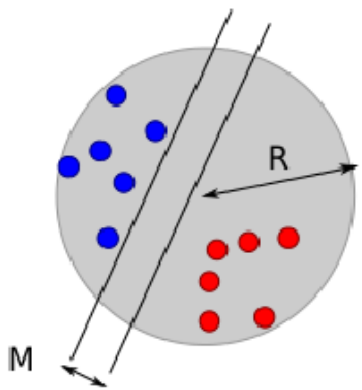
- **降低 VC 维度** h（即选择一个较低复杂度的模型类别）。
- **增加训练样本数量** N。

在这个公式中，η 通常表示**置信水平 (confidence level) 或容忍概率 (tolerance probability)**。

具体来说，η 与不等式中的概率界限相关联，描述了在给定置信区间内真实误差和训练误差之间差异的可能性。通常，η 定义为一个接近于 0 的小值，例如 0.05 或 0.01，对应高置信度 ($1-\eta$，比如 95% 或 99%)。

Characterizing Complexity (One-Layer Networks)

$$f(\boldsymbol{x}, (\boldsymbol{w}, b)) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$



$$h = \min\left\{d + 1, 4\frac{R^2}{M^2} + 1\right\}$$

dimensions margin =m

在机器学习（尤其是支持向量机和分类问题中），margin（边界） 是指决策边界与最近的样本点之间的最小距离

R 表示数据的半径或数据分布的范围，即样本点到原点的最远距离。换句话说，所有样本点都被包含在以原点为中心、半径为 R 的球体内。

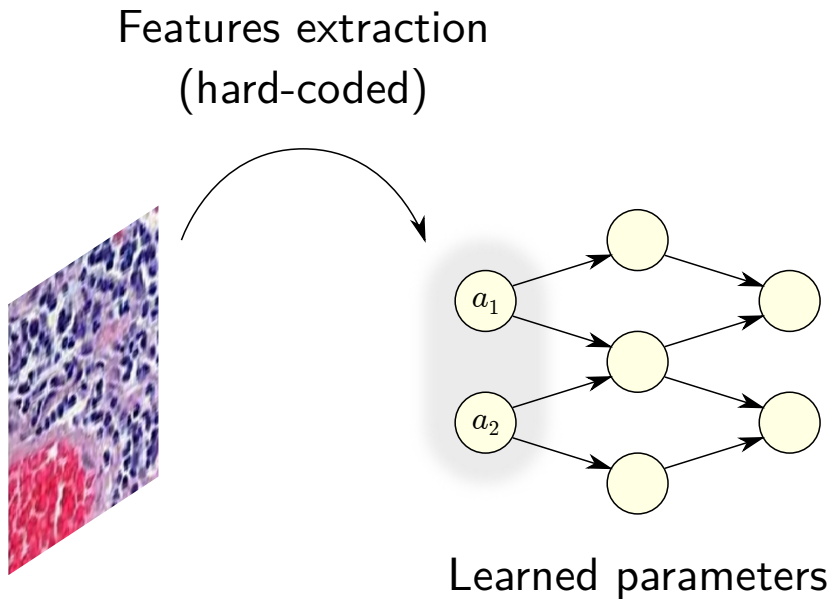
Interpretation:

- Model complexity can be restrained if the input data is low-dimensional **or** if the model builds a large margin (i.e. has low sensitivity).

Question:

- Can we build similar concepts for deep neural networks?

Reducing Complexity via Low Dimensionality



Approach:

- First, generate a low-dimensional representation by extracting a few features from the high-dimensional input data (either hand-designed, or automatically generated using methods such as PCA).
- Then, learn a neural network on the resulting low-dimensional data.

Reducing Complexity via Low Dimensionality

Observations:

- Building low-dimensional representations can be useful when predicting noisy high-dimensional data such as gene expression in biology ($d > 20000$)
- On other tasks such as *image recognition*, low dimensional representation can also delete class-relevant information (e.g. edges).

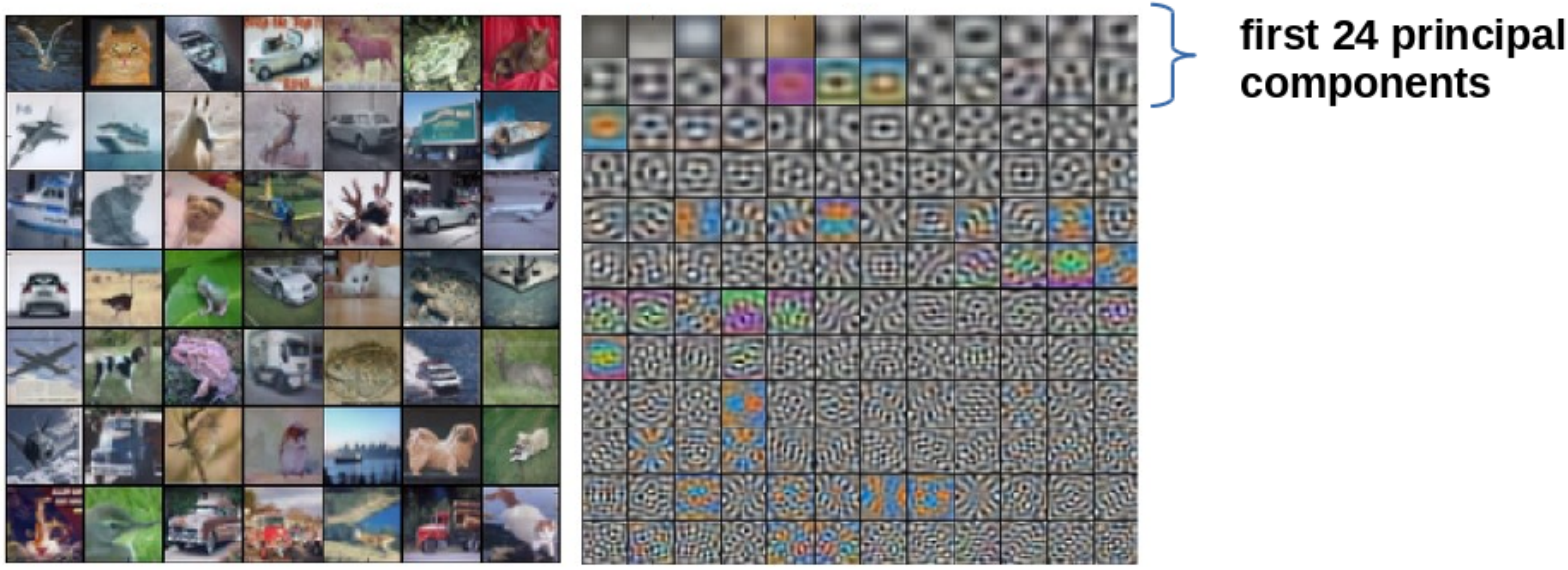
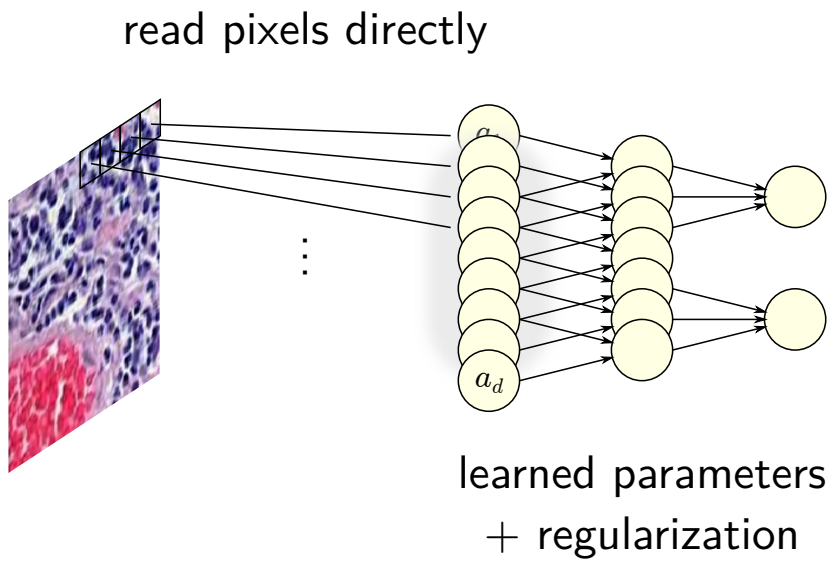


Image from Stanford CS class CS231n: Convolutional Neural Networks for Visual Recognition.

Reducing Complexity by Reducing Sensitivity



Weight Decay [4]:

- 在目标函数中加入一个项，使得那些对预测任务不必要的权重趋向于 0。

- Include in the objective a term that makes the weights tend to zero if they are not necessary for the prediction task.

$$\mathcal{E}(\theta) = \sum_{i=1}^{N} (f(x_i, \theta) - t_i)^2 + \lambda \|\theta\|^2$$

- The higher the parameter λ, the more the exposure of the model to variations in the input domain is being reduced.

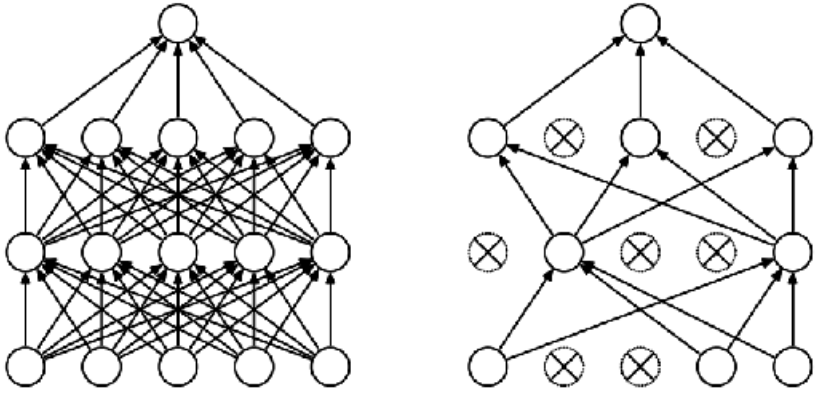
- 第一项：模型的误差（例如均方误差）。
- 第二项：正则化项，控制权重的大小。
- 参数 λ 越大，模型对输入数据域变化的敏感性就越低。

Reducing Complexity by Reducing Sensitivity

- 作为权重衰减的一种替代方法，通过对输入层和中间神经元添加人工乘性噪声来训练模型，使其适应这种噪声。
- 方法：
  - 在神经网络中插入“dropout”层，随机失活一些神经元。具体实现方式是将每个输入（或激活值）乘以一个随机变量 $b_j \sim \text{Bernoulli}(p)$，其中 p 是失活的概率。

Dropout [7]:

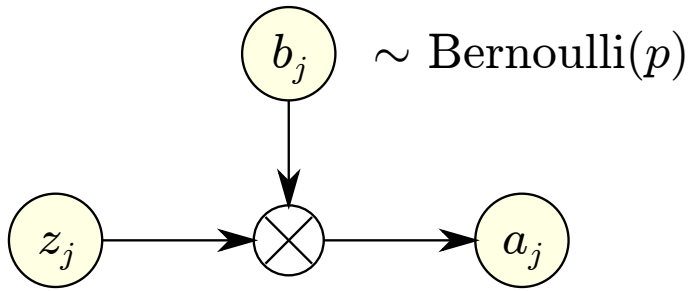
- Alternative to weight decay, which consists of adding artificial multiplicative noise to the input and intermediate neurons, and training the model subject to that noise.



(a) Standard Neural Net (b) After applying dropout.

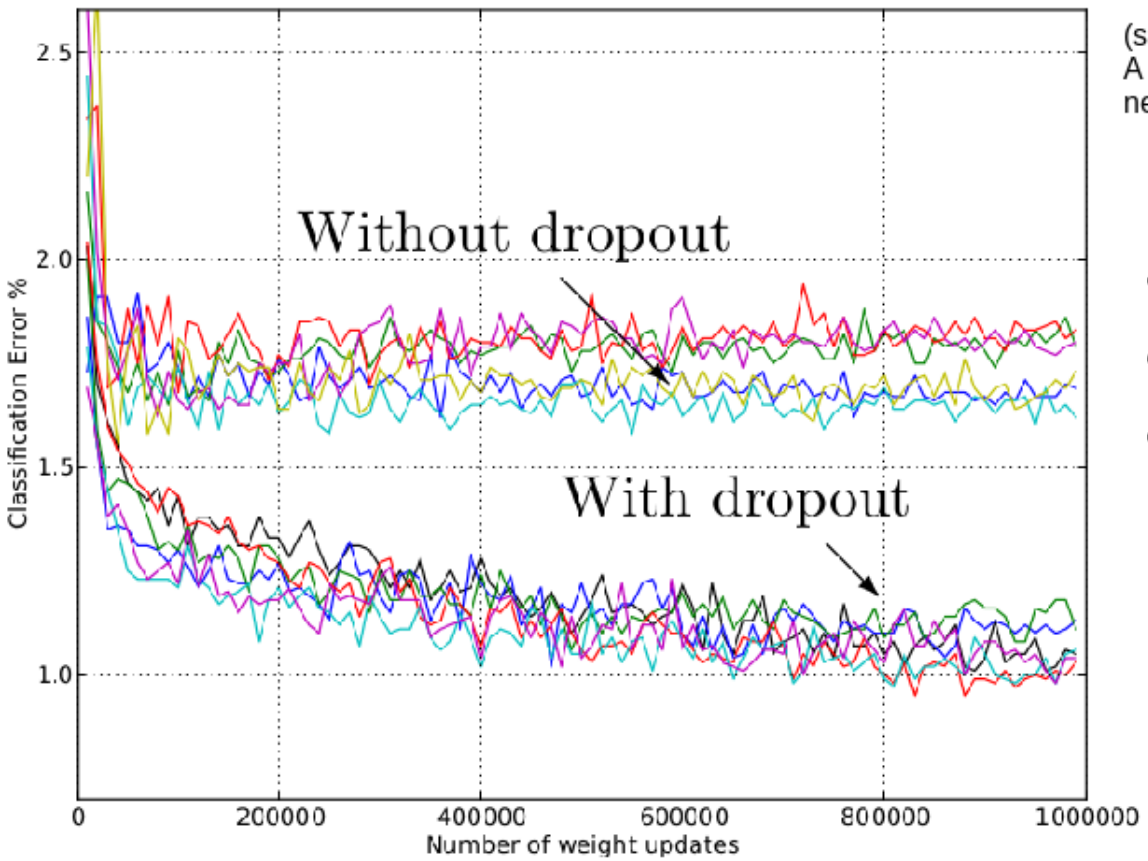
(Source: Srivastava et al. 2014: Dropout: A simple way to prevent neural networks from overfitting).

- This is achieved by inserting a “dropout” layer in the neural network, which multiplies each input (or activation) by a random variable $b_j \sim \text{Bernoulli}(p)$:

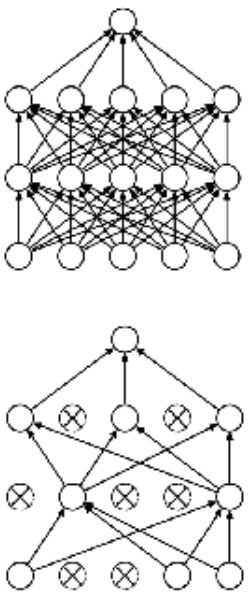


Reducing Complexity by Reducing Sensitivity

Effect of dropout on performance on the MNIST dataset



(source: Srivastava et al. 2014: Dropout: A simple way to prevent neural networks from overfitting).



Note:

- On neural networks for image data, dropout tends to yield superior performance compared to simple weight decay.

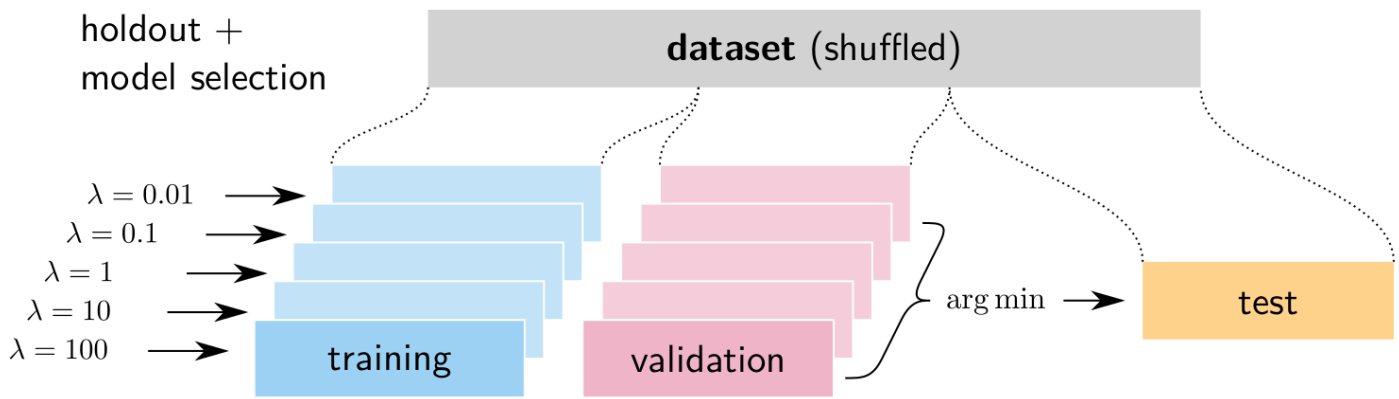
Choosing a Model with Appropriate Complexity

Holdout Validation:

留出验证法（Holdout Validation）

- 训练多个具有不同正则化参数（如 λ）的神经网络，并保留在独立验证集上表现最优的模型。

- Train multiple neural network models with different regularization parameters (e.g. λ), and retain the one that performs the best on some validation set disjoint from the training data.



Problem:

- Training a model for each parameter λ can be costly. One would potentially benefit from training a bigger model only once.

- 针对每个 λ 训练一个模型的计算成本较高，因此训练一个更大的模型可能更具优势。

Accelerating Model Selection

Early Stopping Technique [6]:

- View the iterative procedure for training a neural network as generating a sequence of increasingly complex models $\theta_1, \ldots, \theta_T$.
- Monitor the validation error throughout training and keep a snapshot of the model when it had lowest validation error.

Early stopping:

$\theta^\star = \texttt{None}$
$\mathcal{E}^\star = \infty$
for $t = 1 \ldots T$ **do**
  Run a few SGD steps, and collect the current parameter θ_t
  if $\mathcal{E}_{\text{val}}(\theta_t) < \mathcal{E}^\star$ **then**
    $\theta^\star \leftarrow \theta_t$
    $\mathcal{E}^\star \leftarrow \mathcal{E}_{\text{val}}(\theta)$
  end if
end for

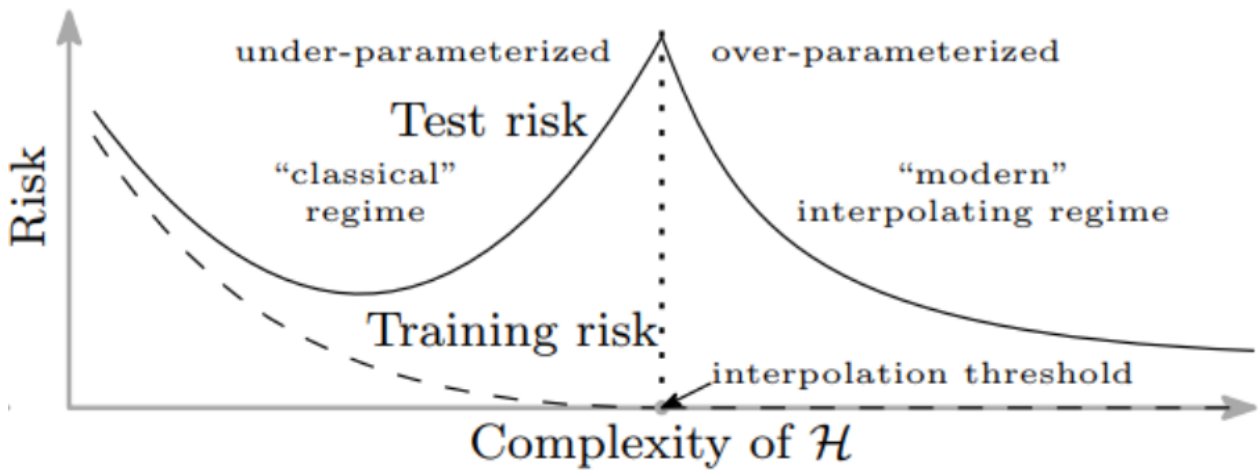
Advantage:

- No need to train several models (e.g. with diffent λ's). Only one training run is needed!

Very Large Mod

- 当模型变得非常大时，在神经网络中会出现一种有趣的**“双降现象 (Double Descent)”**。随着模型复杂度的增加，泛化误差 (generalization error) 在增加后会再次下降。
- 这一现象可以理解为模型中多个组成部分之间的一种**隐式平均** (interpolating regime) 的效果。

- When the model becomes very large there is an interesting '*double descent*' [2] phenomenon that occurs in the context of neural networks, where the generalization error starts to go down again as model complexity increases.
- This can be interpreted as some implicit averaging between the many components of the model (interpolating regime).



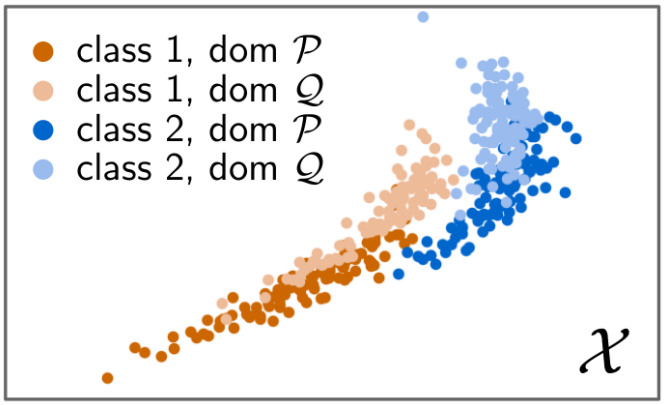
- Increasing model size to a great extent may contribute, without further regularization techniques to achieve lower test set error.
  - 通过大幅增加模型大小，可以在没有进一步正则化技术的情况下，降低测试集误差，从而获得更好的泛化性能。

Part 2

Imbalances between Subgroups

Data from Multiple Domains

- The data might come from different domains ($\mathcal{P}$, $\mathcal{Q}$).
- Domains may e.g. correspond to different acquisition devices, or different ways they are configured/calibrated.
- One of the domains may be overrepresented in the available data, or the ML model may learn better on a given domain at the expense of another domain.



- 数据可能来源于不同的域 (例如, $\mathcal{P}, \mathcal{Q}$)。
- 域可能对应于不同的采集设备，或它们配置/校准方式的不同。
- 数据集中某个域可能被**过度代表**，或者机器学习模型可能对某个域学习得更好，而牺牲了对另一个域的性能。

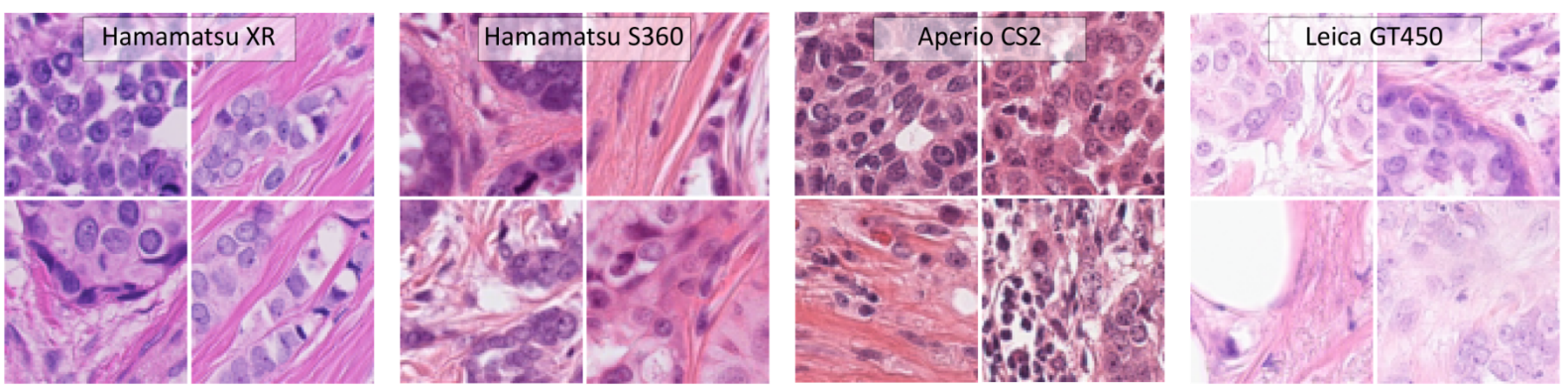


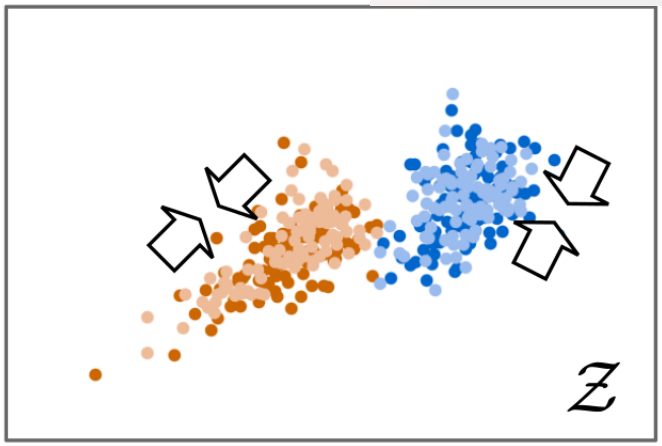
Image source: Aubreville et al. Quantifying the Scanner-Induced Domain Gap in Mitosis Detection. CoRR abs/2103.16515 (2021)

Addressing Multiple Domains

- 这种方法称为**矩匹配 (moment matching)**，它可以扩展为包含更高阶的矩 (如方差等)。
- 更强大的工具 (如 Wasserstein 距离) 可以进一步对分布进行更精细的约束。

右侧图示：

- 箭头表示通过正则化缩小不同域 (橙色和蓝色点) 的分布差异，使其表现更加一致。



Simple Approach (one-layer networks):

- Denoting by $\mathcal{P}$ and $\mathcal{Q}$ the two domains, regularize the ML model $(\boldsymbol{w}^\top \boldsymbol{x})$ so that both domains generate the same responses on average at the output:

$$\min_{\boldsymbol{w}} \quad \mathcal{E}(\boldsymbol{w}) + \lambda \cdot (\mathrm{E}_{\mathcal{P}}[\boldsymbol{w}^\top \boldsymbol{x}] - \mathrm{E}_{\mathcal{Q}}[\boldsymbol{w}^\top \boldsymbol{x}])^2$$

(aka. moment matching). The approach can be enhanced to include higher-order moments such as variance, etc.

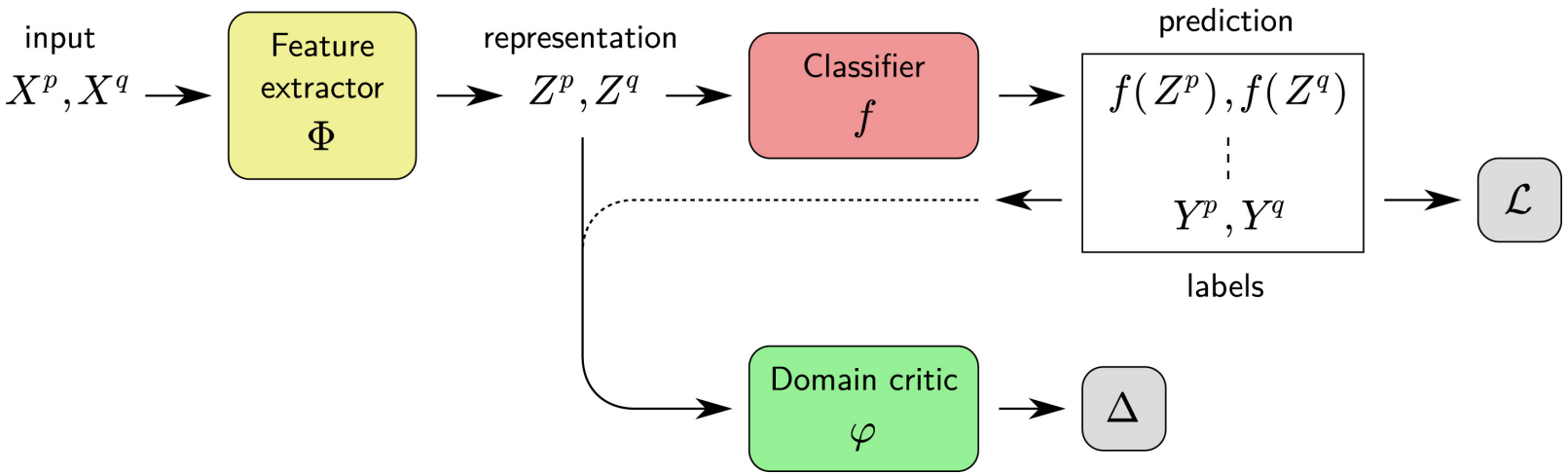
- In practice, more powerful tools exist to constrain distributions more finely in representation space, such as the Wasserstein distance.

Addressing Multiple Domains

- 学习一个辅助神经网络（称为域判别器 **domain critic** φ），其任务是尝试区分来自两个域的数据。
- 学习特征提取器的参数，使得域判别器 φ 无法区分两个域的数据。

More Advanced Approach [1]:

- Learn a auxiliary neural network (domain critic φ) that tries to classify the two domains. Learn the parameters of the feature extractor in a way that the domain critic φ is no longer able to distinguish between the two domains.



Addressing Multiple Domains

- 数据集 Office-Caltech 中某一类别的图片，来自不同域 (如 Amazon、DSLR、Webcam、Caltech)。

实验结果：

- 表格展示了不同方法在多个域上的性能比较：
  - **No critic (无域判别器)：** 性能最低。
  - **Marginal critic (边缘判别器)：** 表现有所改善。
  - **Joint critic (联合判别器)：** 在最差情况下取得最佳准确率，同时平均表现也最优。

Example:

- Example of one particular class of the Office-Caltech dataset and the different domains from which the data is taken.

- Models equiped with a domain critic, although loosing performance on some domains, achieve better worst-case accuracy:

	Amazon	*Caltech*	*DSLR*	*Webcam*	Avg	Min
No critic	*90.63*	*84.27*	*93.75*	*98.31*	91.74	84.27
Marginal critic	*91.32*	*85.46*	*92.71*	*96.07*	91.39	85.46
Joint critic (Ours)	*91.67*	*86.05*	*93.75*	*97.00*	**92.12**	**86.05**

Part 3

Spurious Correlations

Spurious Correlations

- 图片中可能包含特定背景（如围栏、骑手），这些背景与“马”的类别相关，但它们实际上与分类任务无关。

2. **其他案例：**
   - 某一类别图片中只有特定的版权标签出现。
   - 某一类别的病理图像只能由特定设备采集，因此具有不同的颜色配置文件。

总结：
虚假相关性在实际数据集中非常常见，会导致模型学习到错误的特征，影响泛化能力。

虚假相关性（Spurious Correlations）

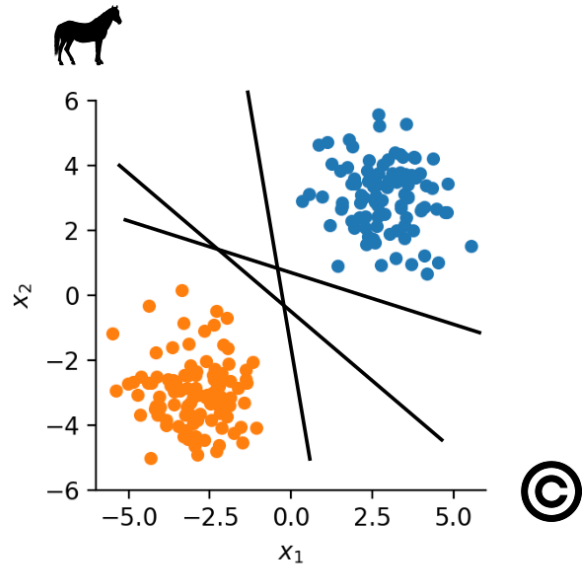
定义：

- 可用数据分布（$\mathcal{P}$）中的伪像，其中一些与任务无关的输入变量与任务相关变量伪相关。

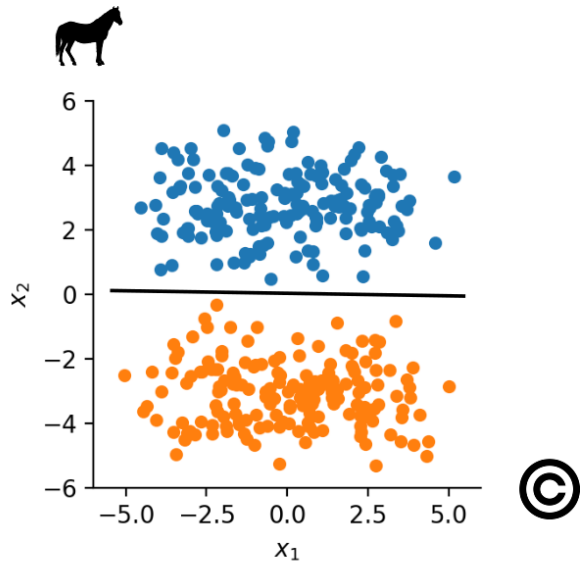
- Artefact of the distribution of available data ($\mathcal{P}$) where one or several task-irrelevant input variables are spuriously correlated to the task-relevant variables.
- Spurious correlations are very common on practical datasets, e.g. a copyright tag occurring only on images of a certain class; histopathological images of a certain class having been acquired with particular device and having as a result a different color profile, etc.

Spurious Correlations and the Clever Hans Effect

Available data ($\mathcal{P}$)



New data ($\mathcal{Q}$)



- A ML classifier is technically able to classify the available data ($\mathcal{P}$) using either the correct features or the spurious ones. The ML model doesn't know a priori which feature (the correct one or the spurious one) to use. A model that bases its decision strategy on the spurious feature is "*right for the wrong reasons*" and is also known as a *Clever Hans classifier*.
- A Clever Hans classifier may fail to function well on the new data ($\mathcal{Q}$) where the spurious correlation no longer exists, e.g. horses without copyright tags, or images of different class with copyright tags.

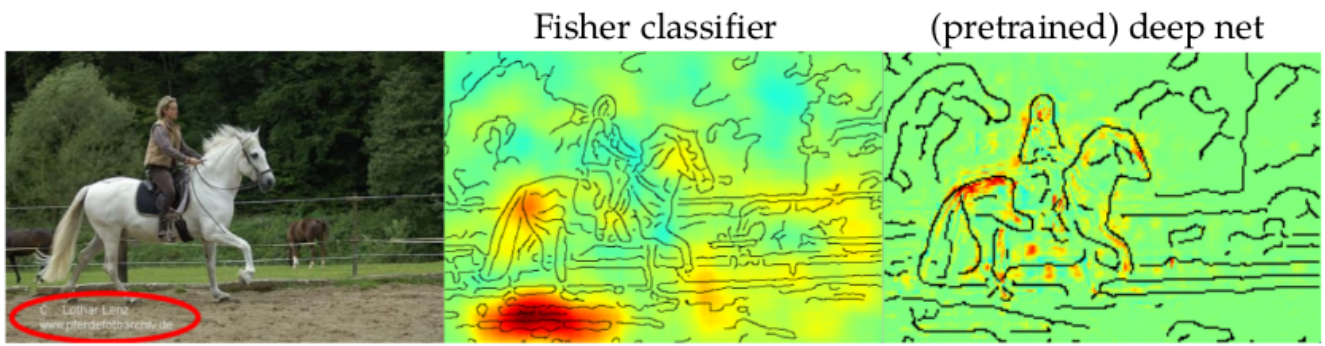
Spurious Correlations and the Clever Hans Effect

- Test set accuracy doesn't give much information on whether the model bases its decision on the correct features or exploits the spurious correlation.

Fisher Vector Classifier vs. DeepNet pretrained on ImageNet

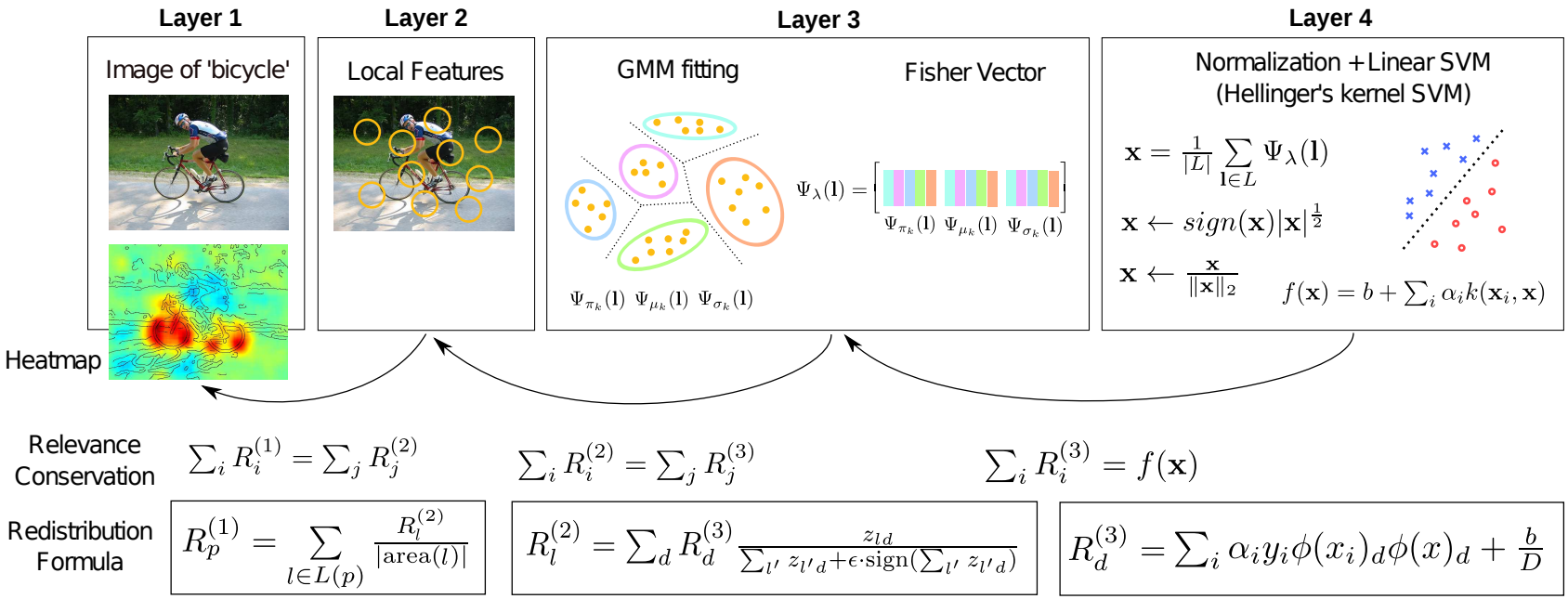
	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	**chair**	**cow**	**diningtable**	**dog**	**horse**	**motorbike**
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

- Only an inspection of the decision structure by the user (e.g. using LRP heatmaps) enables the detection of the flaw in the model [5].



Generating LRP heatmaps

1. **Layer 1（第一层）：**
   - 输入：一张“自行车”的图像。
   - 输出：生成初步的热力图，显示图像中与目标相关的区域。
2. **Layer 2（第二层）：**
   - **局部特征（Local Features）：** 提取图像中的局部特征，例如局部边缘或纹理信息。
   - 热力图中每个特征点的相关性 $R_i^{(1)}$ 被分配到下一层的特征点 $R_j^{(2)}$，**遵循相关性守恒规则**。
3. **Layer 3（第三层）：**
   - **高斯混合模型（GMM）拟合：** 将特征点分布拟合到 GMM，并生成 Fisher 向量。
   - 相关性从 $R^{(2)}$ 分配到 Fisher 向量维度 $R^{(3)}$，保持相关性总和一致。
4. **Layer 4（第四层）：**
   - **归一化与线性支持向量机（SVM）：** 使用 Hellinger 核 SVM 进行分类。
   - 最终输出分类结果 $f(x)$。



$$\sum_i R_i^{(1)} = \sum_j R_j^{(2)} \qquad \sum_i R_i^{(2)} = \sum_j R_j^{(3)} \qquad \sum_i R_i^{(3)} = f(\mathbf{x})$$

$$R_p^{(1)} = \sum_{l \in L(p)} \frac{R_l^{(2)}}{|\text{area}(l)|}$$

$$R_l^{(2)} = \sum_d R_d^{(3)} \frac{z_{ld}}{\sum_{l'} z_{l'd} + \epsilon \cdot \text{sign}(\sum_{l'} z_{l'd})}$$

$$R_d^{(3)} = \sum_i \alpha_i y_i \phi(x_i)_d \phi(x)_d + \frac{b}{D}$$

- Explanations are produced using a layer-wise redistribution process from the output of the model to the input features.
- Each layer can have its own redistribution scheme. The redistribution rules are designed in a way that maximizes explanation quality.

1. **解释的生成：**
   - 使用逐层分配的方式，从模型输出到输入特征生成热力图。
2. **分配方案：**
   - 每一层有其独特的分配规则，设计这些规则的目的是最大化解释质量。

Mitigating Reliance on Spurious Correlations

翻译：减轻对虚假相关性的依赖

特征选择 / 反学习 (Unlearning)：

- 在不包含伪像特征的情况下重新训练模型 (例如，通过裁剪图像以去除版权标签)。
- 主动检查模型中响应伪像的单元 (例如，部分神经元)，并将这些单元从模型中移除。

Feature Selection / Unlearning:

- Retrain without the feature containing the artefact (e.g. crop images to avoid copyright tags).
- Actively look in the model for units (e.g. subsets of neurons) that respond to the artifact and remove such units from the model (e.g. [3]).

数据集设计：

- 手动从包含伪像的类别中移除伪像 (例如，移除版权标签)；或者，反过来在每个类别中注入伪像 (使其不再能够作为区分类别的依据)。
- 对数据集进行分层，使伪像特征在所有类别中具有相似的比例。

Dataset Design:

- Manually remove the artifact (e.g. copyright tags) from the classes that contain it, or alternatively, inject the artifact in every class (so that it cannot be used anymore for discriminating between classes).
- Stratify the dataset in a way that the spurious features are present in all classes in similar proportions.

结合解释约束的学习：

- 在目标函数中加入额外的项，惩罚基于不需要特征的决策策略 (这些不需要的特征可以通过解释技术初步揭示出来)。

Learn with Explanation Constraints

- Include an extra term in the objective that penalizes decision strategies that are based on unwanted features (preliminary revealed by an explanation technique).

Summary

Summary

- While deep learning can in principle fit very complex prediction functions, the way they perform in practice is in large part determined by the amount and quality of the data.
- Limited data may subject the ML model to **overfitting** and lead to lower performance on new data. Various methods exist to prevent overfitting (e.g. generating a low-dimensional input vector, or build a model with limited sensitivity to input such as weight decay or dropout).
- Another problem is **dataset bias**, where certain parts of the distribution are over/under-represented, or plagued with **spurious correlations**. Reliance of the model on spurious correlations can lead to low test performance, but this can be detected by Explainable AI approaches. A number of methods exist to reduce reliance on spurious correlations.

References

- 虽然深度学习在理论上可以拟合非常复杂的预测函数，但其实际表现在很大程度上取决于数据的数量和质量。
- 数据有限可能会导致机器学习（ML）模型发生过拟合（overfitting），从而降低其在新数据上的表现。

  目前有多种方法可以防止过拟合，例如：
  - 生成低维输入向量；
  - 构建对输入变化敏感度较低的模型（如使用权重衰减（weight decay）或 Dropout）。
- 另一个问题是数据集偏差（dataset bias），即分布中的某些部分被过度或不足地表示，或者数据存在伪相关（spurious correlations）。
  - 模型依赖伪相关会导致测试性能下降。
  - 可以通过可解释 AI（Explainable AI）方法检测此类问题。
  - 目前已有多种方法可以减少模型对伪相关的依赖。

[1] L. Andéol, Y. Kawakami, Y. Wad
Learning domain invariant represe ı.
Neural Networks, 167:233–243, 2

[2] M. Belkin, D. Hsu, S. Ma, and S. Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
PNAS, 116(32):15849–15854, 2019.

[3] P. Chormai, J. Herrmann, K. Müller, and G. Montavon.
Disentangled explanations of neural network predictions by finding relevant subspaces.
CoRR, abs/2212.14855, 2022.

[4] A. Krogh and J. A. Hertz.
A simple weight decay can improve generalization.
In *NIPS*, pages 950–957, 1991.

[5] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller.
Unmasking clever hans predictors and assessing what machines really learn.
Nature Communications, 10(1), Mar. 2019.

[6] L. Prechelt.
Early stopping-but when?
In *Neural Networks: Tricks of the Trade*, volume 1524 of *LNCS*, pages 55–69. Springer, 1996.

[7] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov.
Dropout: a simple way to prevent neural networks from overfitting.
J. Mach. Learn. Res., 15(1):1929–1958, 2014.