

- ① Background
- ② A Two-Dimensional Example
- ③ 监督学习

① Background

② A Two-Dimensional Example

③ 监督学习

Spectral Theorem

- The spectral theorem:

$$M = U\Lambda U^* (\text{unitary, symplectic})$$

$$M = U\Lambda U^T (\text{orthogonal})$$

M can be any kinds of matrices.

- Three ensembles:¹
 - **Gaussian Orthogonal Ensemble**: the set of $N \times N$ random **real symmetric** matrices
 - **Gaussian Unitary Ensemble**: the set of $N \times N$ random **complex Hermitian** matrices
 - **Gaussian Symplectic Ensemble**: the set of $N \times N$ random **quaternion self-dual Hermitian** matrices

¹Yanqing Yin and Zhidong Bai. "Convergence rates of the spectral distributions of large random quaternion self-dual Hermitian matrices". In: *Journal of Statistical Physics* 157 (2014), pp. 1207–1224.

Spectral Theorem

- Every **real symmetric** matrix Q with **real** eigenvalues in Λ and orthonormal eigenvectors in the columns of Q (**orthogonal**) can be diagonalized:

$$S = Q\Lambda Q^{-1} = Q\Lambda Q^T \quad \text{with} \quad Q^{-1} = Q^T$$

- Every **Hermitian** matrix S with **real** eigenvalues in Λ and orthonormal eigenvectors in the columns of U (**unitary**) can be diagonalized:

$$S = U\Lambda U^{-1} = U\Lambda U^* \quad \text{with} \quad U^{-1} = U^*$$

Gaussian Orthogonal Ensemble(GOE)

- **Gaussian Orthogonal Ensemble**: the set of $N \times N$ random **real symmetric** matrices

Gaussian Unitary Ensemble(GUE)

- **Gaussian Unitary Ensemble**: the set of $N \times N$ random **complex Hermitian** matrices

Gaussian Symplectic Ensemble(GSE)

- **Gaussian Symplectic Ensemble**: the set of $N \times N$ random **quaternion self-dual Hermitian** matrices

Spectral Theorem

- 谱定理 (spectral theorem) 描述了厄密特矩阵 (Hermitian Matrix)
- 谱分解常见有两种形式:

$$M = U\Lambda U^* (\text{unitary, symplectic})$$

$$M = U\Lambda U^T (\text{orthogonal})$$

- 这可以被视为一种坐标系变化: $M \mapsto (\Lambda, U)$
- 此时可以考虑计算雅可比行列式 (Jacobian): $\left| \det \frac{\partial M}{\partial (\Lambda, U)} \right|$

Spectral Decomposition

- 对于厄密特矩阵 (Hermitian Matrix) 的谱分解来说, 其形式为: $M = U\Lambda U^*$, 其中 M 为厄密特矩阵, U 为酉矩阵 (unitary matrix), $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$
- 对于辛矩阵 (Symplectic Matrix) 的谱分解来说,

① Background

② A Two-Dimensional Example

③ 监督学习

① Background

② A Two-Dimensional Example

③ 监督学习

监督学习


- 监督学习，作为机器学习的核心范式之一，主要依赖于充分标记的数据集来训练算法。
- 常见的技术：支持向量机、决策树、随机森林、朴素贝叶斯、K 近邻算法、神经网络、梯度提升、线性回归、逻辑回归
- 简述四种下游任务：
 - 文本分类
 - 图像识别
 - 机器翻译
 - 异常检测

文本分类

- 文本分类是一项自动将文本文档分类到预定义类别的任务²
- 支持向量机：SVMs 构建在文档类别之间的最优分隔超平面上，以分类新的未标记示例³
- k-最近邻：kNN 算法根据新文档与训练集中标记示例的相似性/距离来预测类别⁴

²Femi Joseph and N Ramakrishnan. "Text categorization using improved K nearest neighbor algorithm". In: *Int J Trends Eng Technol* 4 (2015), pp. 65–68, Krystian Horecki and Jacek Mazurkiewicz. "Natural language processing methods used for automatic prediction mechanism of related phenomenon". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer. 2015, pp. 13–24.

³Khushbu Khamar. "Short text classification using kNN based on distance function". In: *International Journal of advanced research in computer and communication engineering* 2.4 (2013), pp. 1916–1919.

⁴Kwok James Tin-Yau. "Automated text categorization using support vector machine". In: *Proc. the International Conference on Neural Information Processing*. 1999, pp. 347–351, Jason DM Rennie and Ryan Rifkin. "Improving multiclass text classification with the support vector machine". In: (2001). 

图像识别

- 经典的特征提取与分类器方法
- 卷积神经网络方法 (CNN)

经典的特征提取与分类器方法

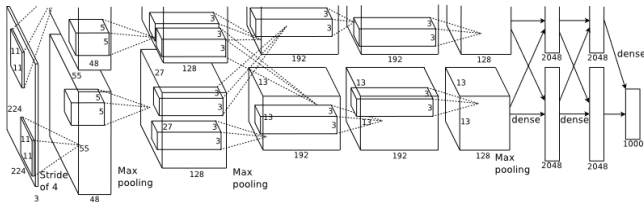
- 在 20 世纪 90 年代之前传统的机器学习时代，主要使用允许特征工程和传统机器学习算法来进行图像识别。这一时期的研究者和工程师们需要大量依赖专业知识和领域经验，去手工设计识别各类目标所需的特征提取算法。这类特征通常针对不同图像识别任务定制，主要有 SIFT⁵, HOG⁶等。

⁵David G Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

⁶N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

卷积神经网络方法 (CNN)

- AlexNet 在深度学习模型中应用了比较深的 8 卷积层的神经网络, 包含了卷积层, 池化层, 全连接层等模块, 如图所示。这种网络结构的设计成为后续深度网络的范式。之后便发展出了更多使用多层感知机 (MLP), 卷积神经网络 (CNN) 等的模型来进行端到端的图像特征学习和分类⁷。

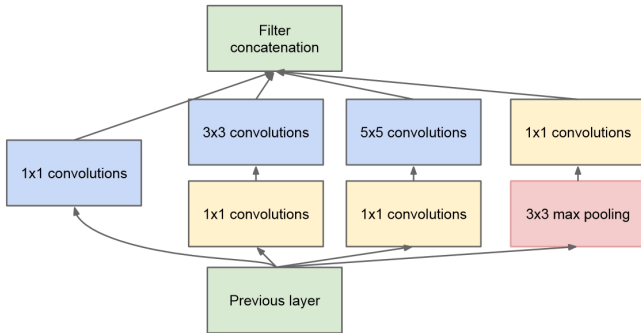


- AlexNet 的成功启示了后续深度学习模型对于层次化特征学习的重要性, 并为图像识别任务提供了一种强大的架构范例。

⁷Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: <https://doi.org/10.1145/3065386>.

卷积神经网络方法 (CNN)

- GoogLeNet 在图像识别中采用了独特的 Inception 模块以及一系列创新性的设计，其处理过程充分利用了多尺度特征的丰富性。总体而言，GoogLeNet 是一种注重了模型的深度、宽度和计算效率的模型⁸。



⁸Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

机器翻译

- 基于规则的机器翻译 (RBMT)
- 统计机器翻译 (SMT)
- 神经机器翻译 (NMT)
- 深度强化学习机器翻译 (DRLMT)

基于规则的机器翻译 (RBMT)

- RBMT 是早期机器翻译方法之一，它的发展可以追溯到 20 世纪 60 年代和 70 年代。如 Fakhrahmad 等人⁹于 2012 年针对词义消歧 (WSD) 这一机器翻译过程中最具挑战性的任务，提出各种监督和无监督学习方法来解决这一问题。近年最新的研究包括 Chauhan 等人¹⁰提出的基于规则的模糊计算方法在印地语机器翻译中的自监督情感极性分类和词义消歧方面的应用

⁹Seyed Mostafa Fakhrahmad et al. "A new fuzzy rule-based classification system for word sense disambiguation". In: *Intelligent Data Analysis* 16.4 (2012), pp. 633–648.

¹⁰Shweta Chauhan et al. "Rule Based Fuzzy Computing Approach on Self-Supervised Sentiment Polarity Classification with Word Sense Disambiguation in Machine Translation for Hindi Language". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.5 (2023), pp. 1–21. 

统计机器翻译 (SMT)

- 基于监督学习的统计机器翻译 (SMT) 是一种传统的机器翻译方法，它使用大量的双语平行语料来训练模型，包括源语言和目标语言之间的句子对。
- González-Rubio 等人¹¹提出了一个用于计算机辅助翻译的成本敏感主动学习框架，其目标是使翻译过程尽可能轻松。与传统的主动学习场景不同，该论文所提出的主动学习框架的设计不仅是为了最小化用户必须监督的翻译数量，还包括最小化每个翻译的监督难度。

¹¹Jesús González-Rubio and Francisco Casacuberta. "Cost-sensitive active learning for computer-assisted translation". In: *Pattern Recognition Letters* 37 (2014), pp. 124–134.

神经机器翻译 (NMT)

- 开发低资源语言翻译技术至关重要，已经成为神经机器翻译中的一个热门研究领域。shi 等人¹²在 2022 发表的一篇综述类论文中对低资源 NMT 中现有的深度学习技术进行了全面的回顾，展示了研究现状及一些广泛使用的低资源数据集，并将这些方法分解为七个类别以总结不同方法之间的共同特点
- 神经机器翻译在过去几年中取得了显著的进展，已成为机器翻译领域的主流方法，但近年更具突破性模型是监督学习下的 SMT 与 NMT 结合，Razaq 等人的研究¹³展示了这一成果，其研究背景为在短语生成 (PG) 中，自然语言中的句子被转换成一个具有不同句法结构但具有相同语义的新句子。

¹²Shumin Shi et al. "Low-resource neural machine translation: Methods and trends". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 21.5 (2022), pp. 1–22.

¹³Abdur Razaq et al. "Improving paraphrase generation using supervised neural-based statistical machine translation framework". In: *Neural Computing and Applications* (2023), pp. 1–15.

异常检测

- 传统监督学习方法
- 深度学习方法
- 应用领域

传统监督学习方法

- 早在 2009 年, Babenko 等人提出了一种基于“多实例学习 (Multiple Instance Learning, MIL)”的方法, 该方法具有潜力应用于异常检测任务¹⁴
- Babenko 等人的研究提出了一种具有实时性能的多实例学习算法, 用于目标跟踪¹⁵。

¹⁴Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. “Visual tracking with online Multiple Instance Learning”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. June 2009, pp. 983–990. DOI: 10.1109/CVPR.2009.5206737.

¹⁵Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. “Visual tracking with online Multiple Instance Learning”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. June 2009, pp. 983–990. DOI: 10.1109/CVPR.2009.5206737.

深度学习方法

- 一项名为"Deep Learning Anomaly Detection Method in Textual Data" 的研究, 由 Amir Jafari 于 2022 年提出, 探讨了如何结合深度学习和传统机器学习算法来检测和识别文本中的异常¹⁶。该研究利用深度学习模型和 Transformer 架构, 将文本数据转化为数值表示, 提供了关于文本数据的关键上下文信息。

¹⁶Amir Jafari. A Deep Learning Anomaly Detection Method in Textual Data. 2022. [arXiv: 2211.13905 \[cs.CL\]](#). ↻ 🔍

应用领域

- 在 2017 年举行的 KDD Workshop on Anomaly Detection in Finance 中，研究人员和从业者聚集在一起，讨论了这些新方法和解决方案¹⁷。
- 一篇题为"Deep Learning based pipeline for anomaly detection and quality enhancement in industrial binder jetting processes" 的研究，由 Alexander Zeiser 等人于 2022 年提出，探讨了在工业制造 使用深度学习进行异常检测和质量提升的方法¹⁸。
- 一篇题为"Cybersecurity Vital Signs: The Role of Anomaly Detection on Insider Threat Triage" 的研究，由 Karla Clarke 和 Yair Levy 于 2019 年提出，探讨了异常检测在内部威胁检 中的重要作用¹⁹。
- 一项名为"Medical Healthcare System Based on Wireless Body Area Networks: The Importance of Anomaly Detection" 的研究，由 Hayder Hassaballah、Rashid Fayadh 和 Bushra AlHayali 于 2020 年提出 旨在探讨在医疗 WBSN 系统中 用异常检测的现代框架²⁰。

¹⁷ Archana Anandakrishnan et al. "Anomaly Detection in Finance: Editors' Introduction". In: *Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance, ADF@KDD 2017, Halifax, Nova Scotia, Canada, August 14, 2017*. Ed. by Archana Anandakrishnan et al. Vol. 71. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1–7. URL: <http://proceedings.mlr.press/v71/anandakrishnan18a.html>.

¹⁸ Alexander Zeiser, Bas van Stein, and Thomas Bäck. *Deep Learning based pipeline for anomaly detection and quality enhancement in industrial binder jetting processes*. 2022. arXiv: 2209.10178 [cs.LG].

¹⁹ Karla Clarke and Yair Levy. "Cybersecurity Vital Signs : The Role of Anomaly Detection on Insider Threat Triage". In: 2019. URL: <https://api.semanticscholar.org/CorpusID:197634071>.

²⁰ Hayder Hassaballah, Rashid Fayadh, and Bushra AlHayali. "Medical Healthcare System Based on Wireless Body Area Networks: The Importance of Anomaly Detection". In: *International Journal of Engineering and Advanced Technology* 9 (Feb. 2020), pp. 3039–3046. doi:10.35940/ijeat.G5945.0220320.

Thanks!