理解深度学习

许志钦

2023 年 11 月 28 日

目录

Chapter 1

一些有意思的现象

这个章节我们介绍一些深度神经网络中有意思的现象。

Leo Breiman's four problems . Leo Breiman（1928 年出生，2005 年逝世）是加州大学伯克利分校的一位杰出统计学家。他的工作帮助弥合了统计学和计算机科学之间的鸿沟，尤其是在机器学习领域。他最著名的工作之一是随机森林（Random Forests）算法的提出，该算法在机器学习领域广泛应用，并取得了显著的成功。他还提出了其他集成方法，如 Bagging。他对机器学习和统计学领域产生了深远的影响。

1995 年，Breiman 在评审完 NIPS（现在叫 NeurIPS）会议的文章后，发表了一篇反思文章 (Breiman 1995) 《Reflections After Refereeing Papers for NIPS》，并提出四个至今仍然很重要的问题

1. Why don't heavily parameterized neural networks overfit the data?
2. What is the effective number of parameters?
3. Why doesn't backpropagation head for a poor local minima?
4. When should one stop the backpropagation and use the current parameters?

相对于“为什么深度学习可以做得好”这种大而空的问题，Breiman 的这几个问题指出了一些重要的具体因素。在这些问题之后，Breiman 评论了关于如何进行研究的方式: Mathematical theory is not critical to the development of machine learning. But scientific inquiry is.

随着人工神经网络的规模越来越大，其复杂也越来越大，单纯地从数学理论出发，可能很难高效地研究实际使用的人工神经网络的基本原理。基于实际问题中的现象，总结更多的规律和研究这些规律背后的理论会是一种更高效的方式。

过参数化不过拟合谜团 ．1995 年, Leo Breiman 提出的问题包括“为什么高度参数化的神经网络不会过拟合数据”。2016 年，一项经验性研究 (Zhang et al. 2017) 再次引起了对这个过参数化之谜的广泛关注，该研究在现代深度神经网络架构和数据集上进行了系统演示。这项研究 (Zhang et al. 2017) 表明，无论是否使用正则化技巧，如权重衰减，一个参数远多于样本的神经网络都能够拟合带有随机标签的数据，并且在真实图像数据集中能够良好地泛化。

这一过参数化不过拟合的谜团与传统的泛化理论和传统建模智慧相矛盾，传统理论表明，参数过多的模型很容易过拟合数据。冯 • 诺伊曼（von Neumann）的著名引言“用四个参数我可以拟合一只大象”(Dyson 2004) 正是一个例子。因此，建立对这一过参数化谜团的良好理论理解已经变得越来越关键，因为现代深度神经网络架构包含了越来越多的参数。为了解决这个谜团，一个显著的工作线路，从传统基于复杂性的泛化理论出发，试图提出适用于深度神经网络的新型基于范数的复杂性度量。然而，一项经验性研究表明，许多基于范数的复杂性度量不仅性能不佳，而且在优化过程中引入一些随机性时与泛化呈负相关 (Jiang et al. 2019)。因此，过参数化仍然是人工神经网络中的一个重要现象谜团。

频率原则/谱偏差 。频率原则/谱偏好 (Xu et al. 2019, 2020, Rahaman et al. 2019, Zhang et al. 2021) 表明，在训练过程中，深度神经网络（DNNs）通常会逐渐拟合目标函数的低到高频分量。在实际研究神经网络学习过程时所面临的主要难题之一是神经网络参数空间和目标函数的极高维度。以 MNIST 数据集为例，这是一个用于测试 DNN 的已知简单（如果不是太简单）的基准数据集。然而，这一函数已经是一个非常高维（784 维）的映射，几乎不可能准确可视化和精确分析。理解训练过程的一个重要步骤是仔细设计足够简单以进行深入分析和可视化 DNN 学习过程的拟合问题，但又足够复杂以重现感兴趣的现象。例如，如图1.1，考虑使用 DNN 来拟合一个具有一维输入和一维输出的函数，如 $\sin(x) + \sin(5x)$。在经过一定的训练轮次后，DNN 的输出会非常靠近 $\sin(x)$，然后才逐渐准确拟合所有的数据。幸运的是，在各种实验中观察到了这一稳定的现象，包括不同的网络结构、训练算法等等，即 DNN 首先捕捉目标函数“平坦”的特征，然后逐渐捕捉更多的振荡细节。这一现象启发了对训练过程进行傅里叶分析，以及高维度任务。频率原则显示，DNNs 擅长学习低频函数，但难以学习高频函数。因此，频率原则有助于解释一系列现象，例如 DNN 难以学习高频分量主导的奇偶函数 (Xu et al. 2020)；DNN 难以重构图像中的高频信息，例如深度图像先验 (Ulyanov et al. 2018, Chakrabarty 2019)；提前停止训练可以避免学习高频噪声 (Xu et al. 2019)。为了克服学习高频分量的难题，许多研究工作提出了多种结构，例如相位偏移 DNN(Cai et al. 2019)、多尺度 DNN(Liu et al. 2020)、自适应激活函数 (Jagtap & Karniadakis 2019)、傅里叶特征网络 (Tancik et al. 2020) 以及神经辐射场中的傅里叶特征输入层 (Mildenhall et al. 2020)。

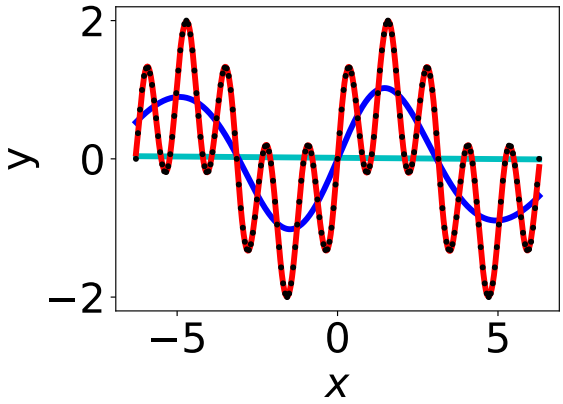

图 1.1: 深度神经网络（DNN）训练过程的示意图。黑色点表示从目标函数 $\sin(x)+\sin(5x)$ 中采样的训练数据。青色、蓝色和红色曲线分别表示在训练轮次 $t=0,2000,17000$ 时 DNN 的输出。

凝聚现象 。文献 (Luo et al. 2021, Zhou, Zhou, Luo, Zhang & Xu 2022, Zhou, Zhou, Jin, Luo, Zhang & Xu 2022) 在神经网络线性区域（Neural Tangent Kernel）之外的非线性训练过程中观察到了一种凝聚现象。例如，在一个两层宽度较大的 ReLU 神经网络中，经过训练后，隐藏神经元的输入权重会集中在孤立的方向上。如1.2所示，该网络的隐藏层聚成两类神经元，因此，其等效成具有两个隐藏神经元的网络。每个等效的隐藏神经元的输入权重和该类里的神经元的输入权重一样，但其输出权重是该类中所有神经元的输出权重的求和。参考文献 (Luo et al. 2021) 证明，在两层 ReLU 网络（宽度无穷大）的相图非线性区域，凝聚现象在合成数据集和真实数据集中都是普遍存在的。类似的观察也适用于三层 ReLU 神经网络 (Zhou, Zhou, Jin, Luo, Zhang & Xu 2022) 以及使用不同激活函数的网络 (Zhou, Zhou, Luo, Zhang & Xu 2022)。此外，已经证明，通过使用较大的学习率 (Andriushchenko et al. 2022) 和丢弃技术 (Zhang & Xu 2022)，可以促进中等大小网络在常见初始化下的凝聚现象。关于凝聚的理论研究已经提供了关于在训练的初始阶段发生凝聚的见解，考虑到具有不同激活函数的网络 (Maennel et al. 2018, Pellegrini & Biroli 2020, Zhou, Zhou, Luo, Zhang & Xu 2022)。凝聚将一个过参数化的网络转化为只有少数有效神经元的网络，从而导致输出函数的复杂度降低。因此，凝聚提供了对 Leo Breiman 提出的一个问题的理解，即关于大型网络的有效神经元数量 (Breiman 1995)。

Lottery ticket hypothesis 。大规模参数数量使得神经网络的训练计算成本高昂且占用大量内存。为了解决这个问题，人们开发了神经网络剪枝技术，以减少参数数量，从而提高计

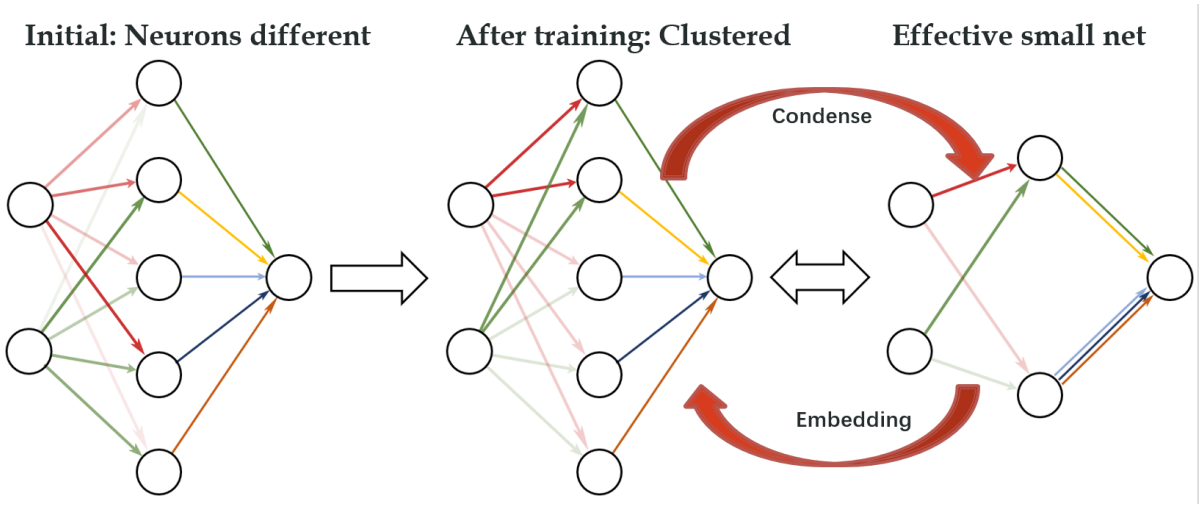

图 1.2: 理想的凝聚现象。

算效率而几乎不损失准确性。(Frankle & Carbin 2018) 观察到，随机初始化的非稀疏前馈网络包含子网络，这些子网络在单独训练时（使用和全网络完全一样的初始化）可以在相似的迭代次数内达到与原始网络相媲美的测试准确性，这被称为“彩票票据假设”(Lottery ticket hypothesis)。(You et al. 2020) 发现这些中奖票据可以在非常早期的训练阶段被识别出，并将其称为“早鸟票据”，通过低成本的训练方案（例如，提前停止和低精度训练）以及较大的学习率。

Edge of stability. 许多实验观察到一种现象，称为 Edge of Stability（EoS）(Wu et al. 2018, Cohen et al. 2021, Arora et al. 2022)，在神经网络（NN）训练过程中，损失函数的 Hessian 矩阵的最大特征值，$\lambda_{\max}$，逐渐增加，直到达到 $2/\eta$（其中 η 是学习率），然后 $\lambda_{\max}$ 会稳定在大约 $2/\eta$ 附近。注意，$2/\eta$ 是二次损失函数上梯度流达到稳定的临界锐度。在 EoS 阶段，损失将持续下降，有时会出现轻微的振荡。使用更大的学习率会导致 $\lambda_{\max}$ 较小的解。

最小值点的平坦度. 在随机梯度下降的训练中，许多实验观察到，当使用较大的批量时，模型的泛化能力会下降。(Keskar et al. 2016) 通过大量的实验为此提供了证据，并且他们发现，当使用大批量的随机梯度下降时，模型会倾向于收敛到锐度更大的极小值。直观想象一下，如果损失景观在该极小值点附近非常陡峭，那由于训练集和测试集在采样时总是有一定差异，这一点差异将使模型在测试集上有很大的误差。这种对损失景观的最小值附近的平坦度的研究也启发了一系列的后续工作。例如 (Zhu et al. 2018, Feng & Tu 2021) 经验发现随机梯度下降的噪

声的协方差与损失函数的 Hessian 具有显著性对齐，或者说在平坦的地方，训练的噪音的方差比较小，而在陡峭的地方，噪音比较大，也称为方差-平坦度逆关系。大的噪音直观上有助于模型跳出陡峭的极小值点。

懒惰训练或神经正切线核体制. (Jacot et al. 2018, Chizat et al. 2019) 发现在适当的参数初始化尺度（通常较大尺度）下，理论和实验的结果都表明，训练后的参数与它们的初始值非常接近。在这种情况下，神经网络在训练期间可以通过一阶泰勒展开很好地近似，即神经网络模型等效于一个核方法。核方法中模型的输出关于所有可训练参数都是线性依赖的，因此这种初始化区域也被称为线性区域。对于两层宽度较大的 ReLU 神经网络，(Luo et al. 2021) 在理论上确定了完整的线性区域。理论上有许多用于分析核方法的工具，因此线性区域也提供了一个窗口来理解神经网络。然而，值得注意的是，神经网络主要是在非线性区域下体现出非常强大的学习能力，但研究非线性模式要困难得多。

语境学习（In-context learning）. 一些研究通过实验证明，语言模型可以通过上下文学习执行各种下游自然语言处理（NLP）任务，在这种情况下，模型在执行未见过的任务之前会在提示的一部分中获得输入-标签对的示例 (Brown et al. 2020)。越来越多的研究表明，良好的性能主要受语义先验和其他因素的驱动，例如提问的格式 (Min et al. 2022, Wang et al. 2022)。也有部分研究工作致力于从贝叶斯角度和数据组成角度理解为什么上下文学习有效 (Xie et al. 2021, Razeghi et al. 2022)。

涌现（Emergence）. 复杂系统的新兴特性一直以来都在交叉学科领域中进行研究，从物理学到生物学再到数学。最近，由于观察到大型语言模型（LLMs）表现出所谓的“涌现能力”（emergent abilities）(Brown et al. 2020, Wei et al. 2022)，即在较小规模的模型中不存在但在较大规模的模型中存在的能力，涌现的概念在机器学习中引起了大量关注。这一涌现现象可以通过神经缩放定律（neural scaling laws）现象定量描述：经验观察表明，深度神经网络的测试损失与训练数据集大小、参数数量或计算量之间呈现幂律尺度关系 (Kaplan et al. 2020, Gordon et al. 2021)。

双下降（Double Descent）. 双下降现象 (Nakkiran et al. 2021) 是指当考虑模型的泛化能力以下面四个参数之一为自变量时，泛化能力会首先变好，接着变差，最后再变好：数据量、数据维度、模型中的参数数量和训练轮次数量。这与通常的观念相矛盾，以参数量为例，通常认为过度参数化会导致过拟合。一个重要的问题是要理解双下降现象与传统的偏差和方差权衡观念之间的关系。现有的主要结果是分析解析可处理的模型，这些模型表现出双下降现象 (Adlam & Pennington 2020, Rocks & Mehta 2022, Mei & Montanari 2022)。

参考文献

Adlam, B. & Pennington, J. (2020), 'Understanding double descent requires a fine-grained bias-variance decomposition', *Advances in neural information processing systems* **33**, 11022–11032.

Andriushchenko, M., Varre, A., Pillaud-Vivien, L. & Flammarion, N. (2022), 'Sgd with large step sizes learns sparse features', *arXiv preprint arXiv:2210.05337* .

Arora, S., Li, Z. & Panigrahi, A. (2022), Understanding gradient descent on the edge of stability in deep learning, *in* 'International Conference on Machine Learning', PMLR, pp. 948–1024.

Breiman, L. (1995), 'Reflections after refereeing papers for nips', *The Mathematics of Generalization* **XX**, 11–15.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* **33**, 1877–1901.

Cai, W., Li, X. & Liu, L. (2019), 'A phase shift deep neural network for high frequency wave equations in inhomogeneous media', *Arxiv preprint, arXiv:1909.11759* .

Chakrabarty, P. (2019), The spectral bias of the deep image prior, *in* 'Bayesian Deep Learning Workshop and Advances in Neural Information Processing Systems (NeurIPS)'.

Chizat, L., Oyallon, E. & Bach, F. (2019), 'On lazy training in differentiable programming', *Advances in neural information processing systems* **32**.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z. & Talwalkar, A. (2021), 'Gradient descent on neural networks typically occurs at the edge of stability', *arXiv preprint arXiv:2103.00065* .

Dyson, F. (2004), 'A meeting with Enrico Fermi', *Nature* **427**(6972), 297–297.

Feng, Y. & Tu, Y. (2021), 'The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima', *Proceedings of the National Academy of Sciences* **118**(9).

Frankle, J. & Carbin, M. (2018), 'The lottery ticket hypothesis: Finding sparse, trainable neural networks', *arXiv preprint arXiv:1803.03635* .

Gordon, M. A., Duh, K. & Kaplan, J. (2021), Data and parameter scaling laws for neural machine translation, *in* 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', pp. 5915–5922.

Jacot, A., Gabriel, F. & Hongler, C. (2018), Neural tangent kernel: Convergence and generalization in neural networks, *in* 'Advances in neural information processing systems', pp. 8571–8580.

Jagtap, A. D. & Karniadakis, G. E. (2019), 'Adaptive activation functions accelerate convergence in deep and physics-informed neural networks', *arXiv preprint arXiv:1906.01170* .

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. & Bengio, S. (2019), 'Fantastic generalization measures and where to find them', *arXiv preprint arXiv:1912.02178* .

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. & Amodei, D. (2020), 'Scaling laws for neural language models', *arXiv preprint arXiv:2001.08361* .

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. (2016), 'On large-batch training for deep learning: Generalization gap and sharp minima', *arXiv preprint arXiv:1609.04836* .

Liu, Z., Cai, W. & Xu, Z.-Q. J. (2020), 'Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains', *Communications in Computational Physics* **28**(5), 1970–2001.

Luo, T., Xu, Z.-Q. J., Ma, Z. & Zhang, Y. (2021), 'Phase diagram for two-layer relu neural networks at infinite-width limit', *Journal of Machine Learning Research* **22**(71), 1–47.

Maennel, H., Bousquet, O. & Gelly, S. (2018), 'Gradient descent quantizes relu network features', *arXiv preprint arXiv:1803.08367* .

Mei, S. & Montanari, A. (2022), 'The generalization error of random features regression: Precise asymptotics and the double descent curve', *Communications on Pure and Applied Mathematics* **75**(4), 667–766.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R. & Ng, R. (2020), Nerf: Representing scenes as neural radiance fields for view synthesis, *in* 'European Conference on Computer Vision', Springer, pp. 405–421.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. & Zettlemoyer, L. (2022), Rethinking the role of demonstrations: What makes in-context learning work?, *in* 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', pp. 11048–11064.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. & Sutskever, I. (2021), 'Deep double descent: Where bigger models and more data hurt', *Journal of Statistical Mechanics: Theory and Experiment* **2021**(12), 124003.

Pellegrini, F. & Biroli, G. (2020), 'An analytic theory of shallow networks dynamics for hinge loss classification', *Advances in Neural Information Processing Systems* **33**.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y. & Courville, A. (2019), On the spectral bias of neural networks, *in* 'International Conference on Machine Learning', pp. 5301–5310.

Razeghi, Y., Logan IV, R. L., Gardner, M. & Singh, S. (2022), 'Impact of pretraining term frequencies on few-shot reasoning', *arXiv preprint arXiv:2202.07206* .

Rocks, J. W. & Mehta, P. (2022), 'Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models', *Physical review research* **4**(1), 013201.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. & Ng, R. (2020), Fourier features let networks learn high frequency functions in low dimensional domains, *in* 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 7537–7547.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. S. (2018), Deep image prior, *in* '2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018', IEEE Computer Society, pp. 9446–9454.
URL: *http://openaccess.thecvf.com/content_cvpr_2018/html/Ulyanov_Deep_Image_Prior_CVPR_2018_pa*

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L. & Sun, H. (2022), 'Towards understanding chain-of-thought prompting: An empirical study of what matters', *arXiv preprint arXiv:2212.10001* .

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. et al. (2022), 'Emergent abilities of large language models', *arXiv preprint arXiv:2206.07682* .

Wu, L., Ma, C. et al. (2018), 'How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective', *Advances in Neural Information Processing Systems* **31**.

Xie, S. M., Raghunathan, A., Liang, P. & Ma, T. (2021), An explanation of in-context learning as implicit bayesian inference, *in* 'International Conference on Learning Representations'.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y. & Ma, Z. (2020), 'Frequency principle: Fourier analysis sheds light on deep neural networks', *Communications in Computational Physics* **28**(5), 1746–1767.

Xu, Z.-Q. J., Zhang, Y. & Xiao, Y. (2019), 'Training behavior of deep neural network in frequency domain', *International Conference on Neural Information Processing* pp. 264–274.

You, H., Li, C., Xu, P., Fu, Y., Wang, Y., Chen, X., Lin, Y., Wang, Z. & Baraniuk, R. G. (2020), 'Drawing early-bird tickets: Towards more efficient training of deep networks', *International Conference on Learning Representations* .

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2017), Understanding deep learning requires rethinking generalization, *in* '5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings', OpenReview.net.
URL: *https://openreview.net/forum?id=Sy8gdB9xx*

Zhang, Y., Luo, T., Ma, Z. & Xu, Z.-Q. J. (2021), 'A linear frequency principle model to understand the absence of overfitting in neural networks', *Chinese Physics Letters* **38**(3), 038701.

Zhang, Z. & Xu, Z.-Q. J. (2022), 'Implicit regularization of dropout', *arXiv preprint arXiv:2207.05952* .

Zhou, H., Zhou, Q., Jin, Z., Luo, T., Zhang, Y. & Xu, Z.-Q. (2022), 'Empirical phase diagram for three-layer neural networks with infinite width', *Advances in Neural Information Processing Systems* **35**, 26021–26033.

Zhou, H., Zhou, Q., Luo, T., Zhang, Y. & Xu, Z.-Q. (2022), 'Towards understanding the condensation of neural networks at initial training', *Advances in Neural Information Processing Systems* **35**, 2184–2196.

Zhu, Z., Wu, J., Yu, B., Wu, L. & Ma, J. (2018), 'The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects', *arXiv preprint arXiv:1803.00195* .