Exact and Consistent Interpretation for Piecewise Linear Neural Networks: A Closed Form Solution
https://arxiv.org/pdf/1802.06259.pdf
В статье описан метод точной и согласованной (exact and consistent) интерпретации кусочно-линейных нейронных сетей. Модель представляется в виде набора линейных классификаторов (каждый классификатор отвечает за предсказание на некотором выпуклом многограннике в пространстве признаков).
Уже существующие методы не обладают точностью и согласованностью.
Рассмотренный метод обладает этими свойствами, так как близкие объекты классифицируются одним и тем же линейным классификатором.
Рассмотрены методы анализа скрытых слоев, метод подражания модели, метод локальной интерпретации (все они не обладают требуемыми свойствами).
Упомянут метод LIME: интерпретирует предсказания произвольного классификатора путем обучения интерпретируемой модели в окрестности входного объекта. Однако метод несогласованный.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
https://arxiv.org/abs/1602.04938
В работе описан метод LIME.
SP-LIME: метод, позволяющий получить набор объектов, на которых получив интерпретацию классификатора, эксперт сможет доверять модели.
Интерпретация моделей помогает найти проблемы в датасете.
Локальная точность/согласованность(local fidelity/consistency): Помимо интерпретации конкретного предсказания также разумно предоставлять информацию о поведении модели в окрестности исходного объекта.
В идеальном случае необходимо предсказывать любой классификатор и не закладываться на особенности модели.
LIME: находит интерпретируемую модель (а именно линейную модель), работающую с интерпретируемым представлением данных в окрестности исходного объекта.
Алгоритм: случайно перебираем интерпретируемые признаки исходного объекта, восстанавливаем обычные признаки нового объекта. Получаем таким образом выборку, считаем ответы модели на этой выборке. Далее обучаем линейную модель на интерпретируемых признаках и полученных ответах.

Finding Representative Interpretations on Convolutional Neural Networks
https://openaccess.thecvf.com/content/ICCV2021/papers/Lam_Finding_Representative_Interpretations_on_Convolutional_Neural_Networks_ICCV_2021_paper.pdf
Альтернативный подход к интерпретации CNN. Авторы утверждают, что LIME и OpenBox, в отличие от их нового подхода, объясняют предсказание модели локально, а не глобально, и в этом их основная проблема.

On Interpretability of Artificial Neural Networks: A Survey
https://arxiv.org/ftp/arxiv/papers/2001/2001.02522.pdf

Статья содержит обзор методов интерпретации нейросетевых моделей. Определяются и описываются требуемые характеристики моделей.

Основные методы интерпретации нейросетевых моделей:

- Анализ скрытых слоёв (интерпретируются признаки, выученные скрытыми слоями). Показывает структуру сети(как работают скрытые нейроны), однако не показывает поведение сети в общем.
    - **Alexey Dosovitskiy and Thomas Brox. 2016. Inverting visual representations with convolutional networks. In CVPR. 4829–4837**
    - **Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2017. Interpreting Deep Visual Representations via Network Dissection. arXiv:1711.05611 (2017).**
- Метод подражания модели (строится интерпретируемая модель с похожим на исходную поведением). Построенную модель проще интерпретировать, однако полученная модель из-за своей простоты может недостаточно хорошо имитировать исходную.
    - **Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep?. In NIPS. 2654–2662**
    - **Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting Blackbox Models via Model Extraction. arXiv:1705.08504 (2017).**
- Метод локальной интерпретации (анализ поведения модели в окрестности входного объекта). Получаем хорошую интерпретацию для одного объекта, однако для двух похожих объектов интерпретации могут существенно различаться (говорим, что модель inconsistent).
    - **Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In CVPR. 2921–2929.**
    - **Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 (2013).**
    - **R. R Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. 2016. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. arXiv:1610.02391 (2016).**
    - **Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. arXiv:1704.03296 (2017).**
    - **Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. arXiv:1703.04730 (2017).**
    - **A. Shrikumar, P. Greenside, and A. Kundaje. 2017. Learning important features through propagating activation differences. arXiv:1704.02685 (2017).**
    - **D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825 (2017).**
    - **Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 (2017).**

Про CNN **Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In NIPS. 1097–1105**

J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. 2015. Understanding neural networks through deep visualization. arXiv:1506.06579 (2015)

D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Muller. How to explain individual ¨ classification decisions. Journal of Machine Learning Research, 11, 2010

M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. Neural information processing systems (NIPS), pages 24–30, 1996.

A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In Computer Vision and Pattern Recognition (CVPR), 2015

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Computer Vision and Pattern Recognition (CVPR), 2015.

Explaining the black-box model: A survey of local interpretation methods for deep neural networks

Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation