Making Neural Networks Interpretable with Attribution: Application to Implicit Signals Prediction

Recsys '20

Darius Afchar, Romain Hennequin

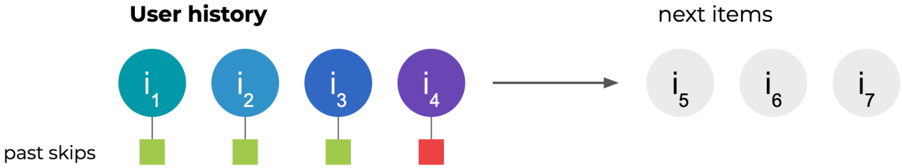
Deezer Research

September 22–26, 2020

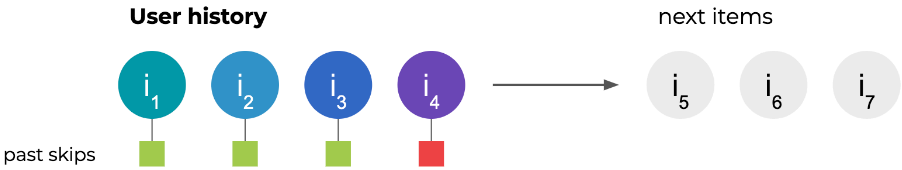
Interpretation for Recommender Systems

- Explain recommended items [4, 9, 11]
- Inspect recommender system models
  - *model performances, fairness, ...*
  - Interpret **implicit data** [5]
    interaction, skips, user churn, ...

e.g. listening session



e.g. listening session



What will be predicted as a skip?

- a user disliking a music? *musical features*
- a user exploring the music catalog? *interaction features*
- something else?

Attribution

Supervised task: *r.v.* $X \in \mathbb{R}^n$, $Y \in \{0, 1\}$

$$X \xrightarrow{f_\theta} Y$$

Attribution

Supervised task: *r.v.* $X \in \mathbb{R}^n$, $Y \in \{0, 1\}$

$$X \xrightarrow{f_\theta} Y$$

Attribution: *r.v.* $S \subset [n]$ be a random set of indices

- *completeness:*

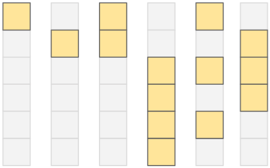
$$X_{|S} \xrightarrow{f_{\theta|S}} Y$$

- *interpretability:* $\mathbb{E}[\mathrm{Card}(S)]$ is as small as possible

Another way to see attribution

With infinite computing capacities,

1. 2^n possibilities for $S: s_1, \dots s_{2^n}$;
2. Train all restricted models $f_1, \dots f_{2^n}$;
3. Select model k^* with **small domain** and **low error**.



$\rightarrow$ *NP-hard, overlapping subsets, ...*

... for real applications

In real-data applications,

- *gradient-based* proxy [1, 8, 10] or *approximations* [7, 2] ;
- *intrinsically attributable* models

$$e.g.\ f(x) = \sum_{i=1}^{n} f_i(x_i)\ [3]$$

$$e.g.\ f(x) = \sum_{i=1}^{n} f_i(x_i) + \sum_{i<j} \tilde{f}_{i,j}(x_i, x_j)\ [6]$$

Our method: $\boxed{f(x) = \sum_{s\in\mathcal{S}} \alpha_s(x_{|s}) f_s(x_{|s})}$

Contributions

1. We propose a novel approach for attribution, and show that it is **applicable to a large class of deep neural networks** to turn them **intrinsically interpretable** ;

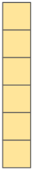
Contributions

1. We propose a novel approach for attribution, and show that it is **applicable to a large class of deep neural networks** to turn them **intrinsically interpretable** ;
2. we derive a fast algorithm to train our networks ;

Contributions

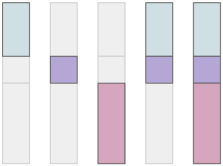
1. We propose a novel approach for attribution, and show that it is **applicable to a large class of deep neural networks** to turn them **intrinsically interpretable** ;
2. we derive a fast algorithm to train our networks ;
3. we demonstrate the effectiveness of our method for **prediction** and **interpretation** on synthetic and real-data tasks (*e.g. sequential skip prediction*).

Mask space reduction



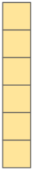
X

→ Reduce the 2^n possible values for S.



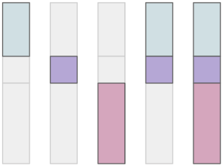
Considered candidates

Mask space reduction



X

→ Reduce the 2^n possible values for S.



Considered candidates

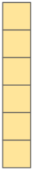
<u>Group-sparsity</u>: we partition $[n]$

$$\mathcal{X} = \{X_1, \ldots X_N\}$$

Codomain of S now has a size

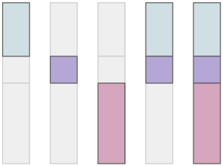
$$2^N \ll 2^n$$

Mask space reduction



X

→ Reduce the 2^n possible values for S.



Considered candidates

Group-sparsity: we partition $[n]$

$$\mathcal{X} = \{X_1, \dots X_N\}$$

Codomain of S now has a size

$$2^N \ll 2^n$$

Structured sparsity: we restrict solutions to follow defined patterns

$$\mathcal{S} \subset \mathcal{P}(\mathcal{X})$$

Codomain of S now has a size

$$H = |\mathcal{S}| < 2^N \ll 2^n$$

Mixture of experts

$f^1, \ldots f^H$ are **restricted** expert models

Mixture of experts

$f^1, \ldots f^H$ are **restricted** expert models

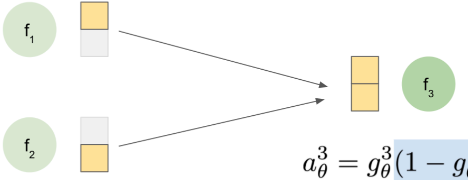
$$f_\theta(x) = \frac{\sum_{s \in \mathcal{S}} \alpha_\theta^s(x_{|s}) f_\theta^s(x_{|s})}{\sum_{s \in \mathcal{S}} \alpha_\theta^s(x_{|s})} \tag{1}$$

Boosting

Example: $X \in \mathbb{R}^2$, S is $\{1\}$, $\{2\}$ or $\{1, 2\}$:

$$a_\theta^1 = g_\theta^1$$

$$a_\theta^2 = g_\theta^2$$



$$a_\theta^3 = g_\theta^3 (1 - g_\theta^1)(1 - g_\theta^2)$$

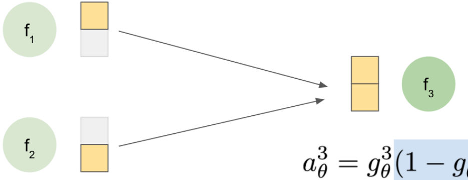
Residuality

Boosting

Example: $X \in \mathbb{R}^2$, S is $\{1\}$, $\{2\}$ or $\{1,2\}$:

$a_\theta^1 = g_\theta^1$

$a_\theta^2 = g_\theta^2$



$a_\theta^3 = g_\theta^3(1 - g_\theta^1)(1 - g_\theta^2)$

Residuality

General case: $\mathrm{child}(s) = \{t | t \subsetneq s\}$

$$\boxed{\alpha_\theta^s(x) = g_\theta^s(x_{|s}) \prod_{t \in \mathrm{child}(s)} (1 - g_\theta^t(x_{|t}))} \quad (2)$$

with $g_\theta^s : x_{|s} \mapsto [0,1]$

Selection modelisation

Using a deep neural network $F^s_\theta : x_{|s} \mapsto [-1, 1]$:

Predictions: binary setting

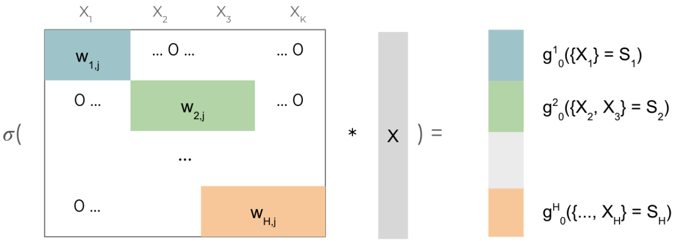
$$f^s_\theta(x) = (F^s_\theta(x) + 1)/2$$

Selections:

$$g^s_\theta(x) = |F^s_\theta(x)|$$

Generalisation

- Can be combined in a single neural network ;
  spoiler: *by masking the weight matrices*



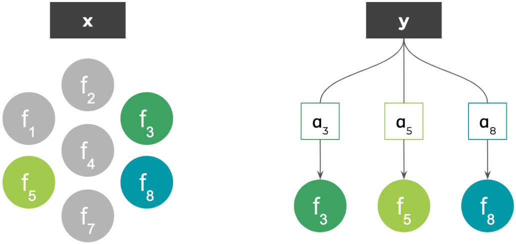
- Time dimension added ;
- RNN, **Transformer**, ... can be easily made interpretable.

The section is fully detailed in the paper!

See also: `github.com/deezer/interpretable_nn_attribution`

Training

- Naive approach: train sequentially with residuals
- Instead → *Generalised Expectation-Maximisation*



Submodel selection

Gradient-step

Results - completeness

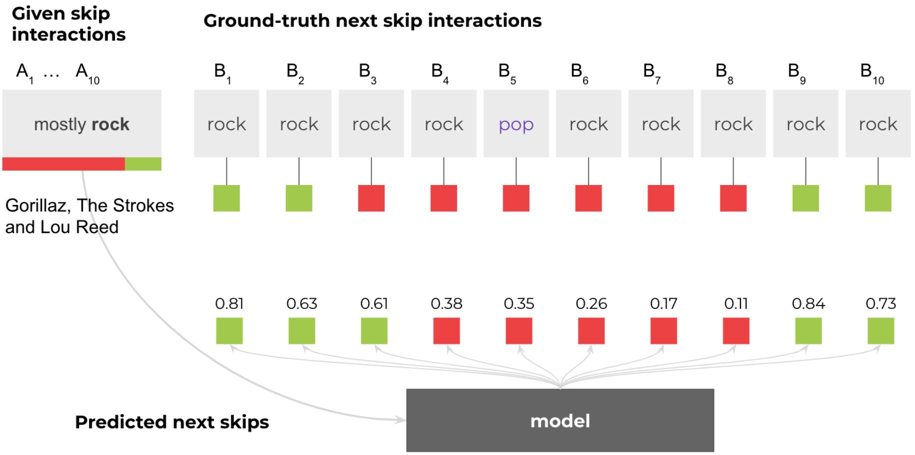
Model	Acc (%)	Acc@1 (%)	MAA (%)
Baseline	70.1	73.3	60.9
Transformer	78.9 ± 0.1	83.4 ± 0.1	70.2 ± 0.1
Interpretable Transformer	77.7 ± 0.1	82.4 ± 0.1	68.8 ± 0.1

Table: Deezer sequential skip prediction test results

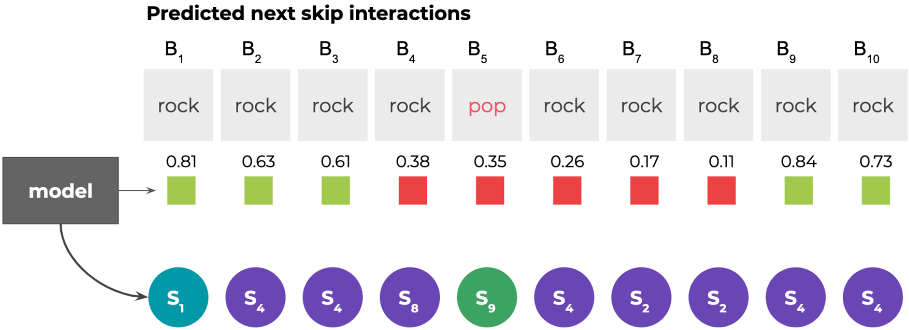
Model	Acc (%)	Acc@1 (%)	MAA (%)
Baseline	63.0	74.2	54.3
Transformer	72.2 ± 0.2	80.0 ± 0.2	62.8 ± 0.2
Interpretable Transformer	70.9 ± 0.2	78.8 ± 0.2	61.1 ± 0.2

Table: Spotify sequential skip prediction test results

Results - interpretation I



Results - interpretation II



Persistence	**0.89**	0.14	0.01	0	0	0	0	0	0	0
Interaction	0.07	**0.73**	**0.92**	**0.53**	0.47	**0.70**	**0.68**	**0.67**	**0.69**	**0.63**
Musical	0.04	0.13	0.06	0.30	**0.52**	0.30	0.32	0.32	0.30	0.35

Thank you!

Repository: github.com/deezer/interpretable_nn_attribution

Future directions:

- Learnable space of candidates $\mathcal{S}$
- Attribution solutions geometry

References I

BACH, S., BINDER, A., MONTAVON, G., KLAUSCHEN, F., MÜLLER, K.-R., AND SAMEK, W.
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
PloS one 10, 7 (2015).

CHEN, J., SONG, L., WAINWRIGHT, M., AND JORDAN, M.
Learning to explain: An information-theoretic perspective on model interpretation.
In *International Conference on Machine Learning* (2018), pp. 883–892.

HASTIE, T. J., AND TIBSHIRANI, R. J.
Generalized additive models, vol. 43.
CRC press, 1990.

HERLOCKER, J. L., KONSTAN, J. A., AND RIEDL, J.
Explaining collaborative filtering recommendations.
In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (2000), pp. 241–250.

References II

HU, Y., KOREN, Y., AND VOLINSKY, C.
Collaborative filtering for implicit feedback datasets.
In *2008 Eighth IEEE International Conference on Data Mining* (2008), Ieee, pp. 263–272.

LOU, Y., CARUANA, R., GEHRKE, J., AND HOOKER, G.
Accurate intelligible models with pairwise interactions.
In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), pp. 623–631.

SCHULZ, K., SIXT, L., TOMBARI, F., AND LANDGRAF, T.
Restricting the flow: Information bottlenecks for attribution.
In *International Conference on Learning Representations* (2019).

SHRIKUMAR, A., GREENSIDE, P., AND KUNDAJE, A.
Learning important features through propagating activation differences.
In *International Conference on Machine Learning* (2017), pp. 3145–3153.

References III

SINHA, R., AND SWEARINGEN, K.
The role of transparency in recommender systems.
In *CHI'02 extended abstracts on Human factors in computing systems* (2002), pp. 830–831.

SMILKOV, D., THORAT, N., KIM, B., VIÉGAS, F., AND WATTENBERG, M.
Smoothgrad: removing noise by adding noise.
Workshop on Visualization for Deep Learning, ICML (2017).

TINTAREV, N., AND MASTHOFF, J.
A survey of explanations in recommender systems.
In *2007 IEEE 23rd international conference on data engineering workshop* (2007), IEEE, pp. 801–810.