Explainability in Recurrent Neural Networks

Sri Kalidindi

Remi Eyraud, Remi Emonet, Amaury Habrard

University Jean Monnet

November 16, 2021

Introduction

Black box Models

Black box Models

Are subclass of machine learning models with complex functions which are hard to explain, understand and interpret.

- What are the features used in a model?
- Which features effect a models decision?
- Does the model consider sensitive features (race, religion, gender)?

Explainability and Interpretability

Interpretability

Interpretability is the degree to which a human can understand the cause of decision in machine or deep learning. [Mol20]

Explainability

Explainability is the degree to which a human can understand the internal mechanics of a machine or deep learning. [Gal19]

Explainability $>$ Interpretability

Variations in Explainability and Interpretability

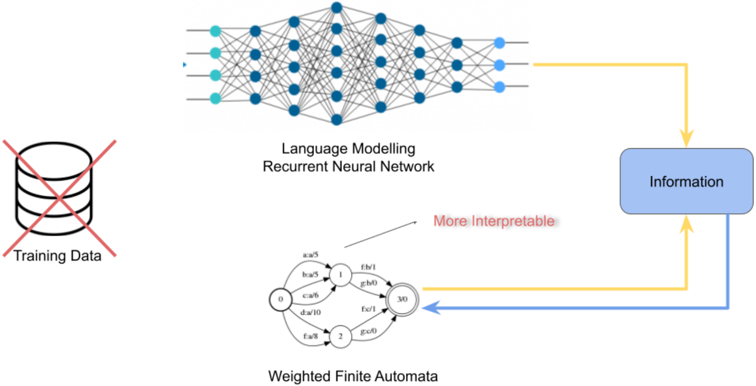
Category of Techniques

- **Global:** A Technique which could explain a model's behaviour for the entire data distribution.
- **Local:** A Technique which could explain a prediction for a particular data-point.
- **Ante-hoc:** A Technique which involves explainability from the learning stage.
- **Post-hoc:** A Technique which can be implemented after the model has finished training.
- **Surrogate:** A Technique which creates a different model approximating the original model function

Distillation of RNN to WFA

In our Approach: Distillation of RNN

- We do not use Training Data.
- We choose Student model (WFA) that is more **Interpretable**.
- We use **Information** from Teacher model.



Language Modelling Recurrent Neural Network (LM-RNN)

Language Modelling Recurrent Neural Network is a recurrent neural network designed to sequential data such as sentences in natural language.

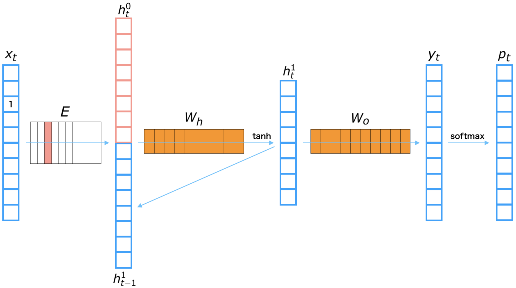


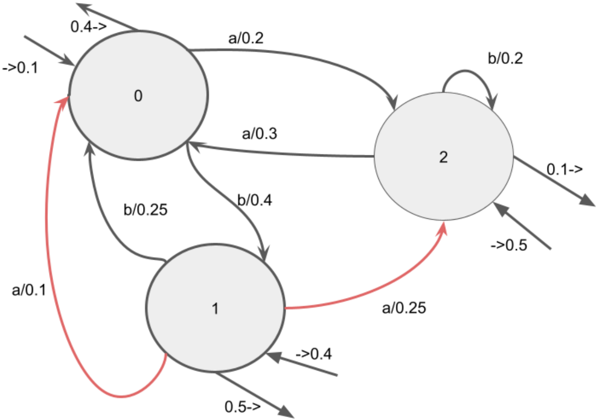
Figure: LM-RNN [SYW16]

- x_t : one-hot vector of t-th letter
- y_t: t-th output.
- $h_t^{(i)}$: t-th hidden vector of i-th layer.
- P_t : $t+1$ word's probability.
- E : Embedding Matrix.
- W_h : Hidden layer Matrix.
- W_o : Output layer Matrix.

Probabilistic Finite Automata

Probabilistic Finite Automata

Probabilistic Finite Automaton (PFA) is a finite automaton whose transitions and states carry probability measure.



Contribution

PFA Distillation

Distillation of LM-RNN to Probabilistic Finite Automata by clustering over hidden state space.

Opening the black box

In almost all the previous approaches the only information we distill from the LM-RNN is conditional probability.

Does LM-RNN has inner representations that could be useful?

Opening the black box

In almost all the previous approaches the only information we distill from the LM-RNN is conditional probability.
Does LM-RNN has inner representations that could be useful?

Lets open the black box!

RNN vs LSTM [Col15]

$$h_t = RNN(h_{t-1}, x_t)$$

Opening the black box

In almost all the previous approaches the only information we distill from the LM-RNN is conditional probability.
Does LM-RNN has inner representations that could be useful?

Lets open the black box!

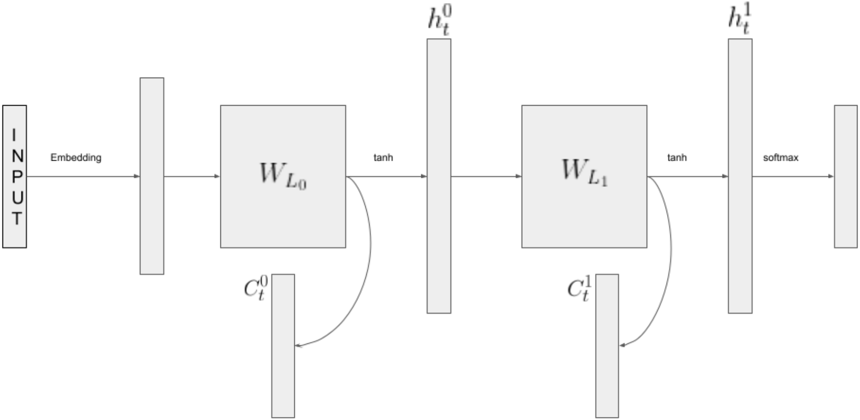
RNN vs LSTM [Col15]

$$h_t = RNN(h_{t-1}, x_t)$$

1. Exploit the information of Hidden states and its space.
2. Does there exist a structure in this Hidden space that correspond to the finite states of an automata?

What is hidden state?

Hidden state at time h_t : $h_t^0 \cdot C_t^0 \cdot h_t^1 \cdot C_t^1$



PFA Distillation

PFA Distillation

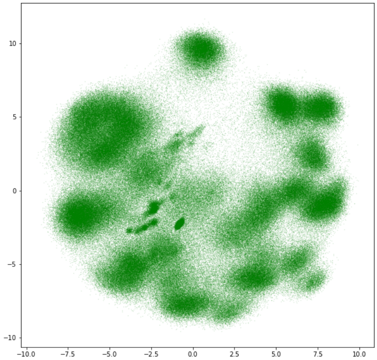


Figure: PCA Plot of Hidden Vectors

1. Sample sequences Z and respective Hidden state vectors H_z.

PFA Distillation

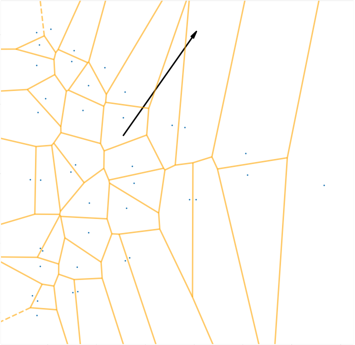


Figure: Vernoi boundaries of K-Means

1. Sample sequences Z and respective Hidden state vectors H_z.
2. Obtain clusters over the vectors sampled.

PFA Distillation

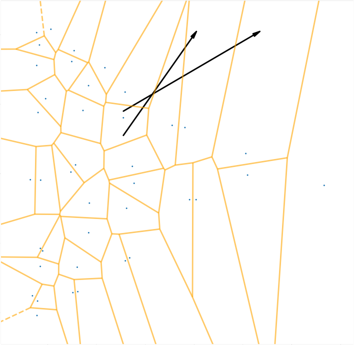


Figure: Vernoi boundaries of K-Means

1. Sample sequences Z and respective Hidden state vectors H_z.
2. Obtain clusters over the vectors sampled.
3. Fill the transitions between clusters by observing all the transitions between hidden vector states.

PFA Distillation

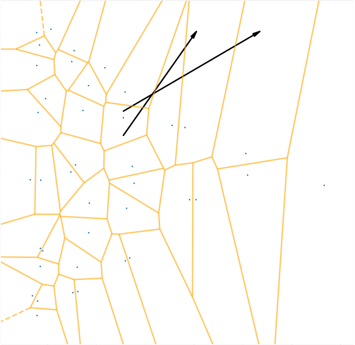


Figure: Vernoi boundaries of K-Means

1. Sample sequences Z and respective Hidden state vectors H_z.
2. Obtain clusters over the vectors sampled.
3. Fill the transitions between clusters by observing all the transitions between hidden vector states.
4. The probabilities are filled for a transition with the fraction of samples that support a transition in a cluster.

Results

Datasets

1. **SPiCe:** 15 Real World sequential datasets from various domains.
2. **PAutomaC:** 48 artificial generated data from HMM, PFA and PDFA.

Metrics - NDCG5

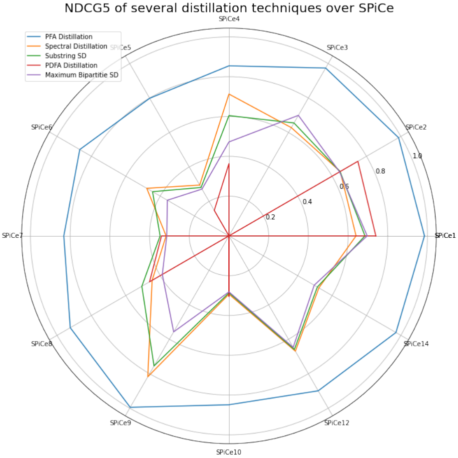
Normalized Discounted Cumulative Gain(NDCG)

NDCG is a popular metric to measure ranking quality. It compares the probabilities of top k candidates between learned and Ideal Model.

$$\mathrm{NDCG}_n(w, \hat{\sigma_1}, ..., \hat{\sigma_n}) = \frac{\sum_{k=0}^{n} \frac{P_{WA}(\hat{\sigma_k}|w)}{log(k+1)}}{\sum_{k=0}^{n} \frac{P_{RNN}(\sigma_k|w)}{log(k+1)}}$$

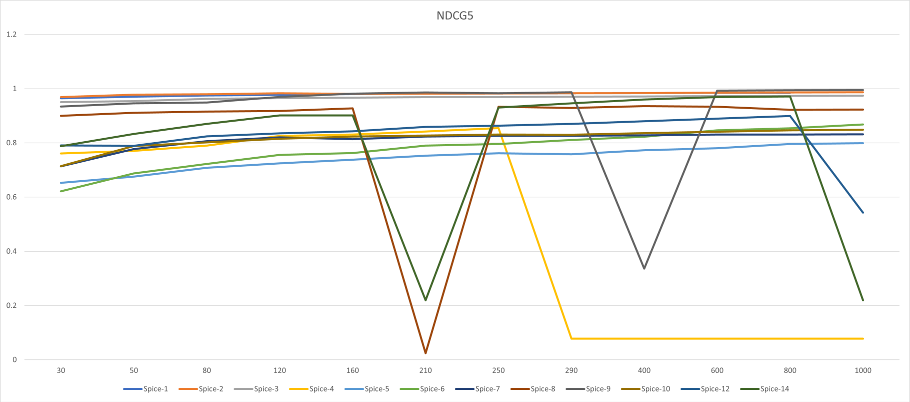
We are comparing the probability distribution between LM-RNN and WFA.

Results- SPiCe



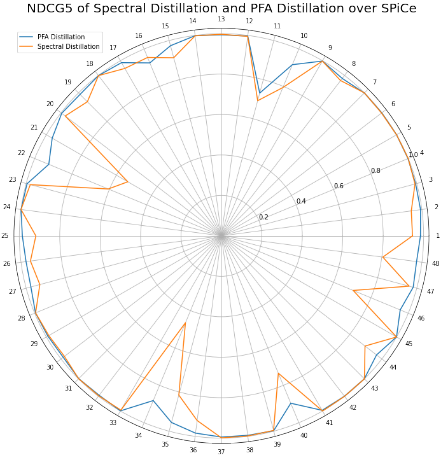
- PFA Distillation shows significant improvements in NDCG Score.

PFA Distillation: Change of NDCG with number of clusters



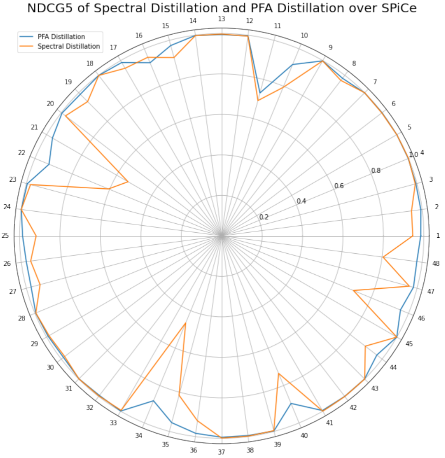
- With the increase in number of cluster NDCG5 keeps increasing but the improvements diminish along the way.

PFA Distillation: NDCG on PAutomaC



- PFA Distillation shows significant improvements in NDCG Score to Spectral Distillation.

PFA Distillation: NDCG on PAutomaC



- PFA Distillation shows significant improvements in NDCG Score to Spectral Distillation.
- The results of PFA Distillation on PAutomaC show Finite States in the hidden state space.

Conclusions

Conclusions

- Hidden states and its Space has information to understand LM-RNN behaviour.

Conclusions

- Hidden states and its Space has information to understand LM-RNN behaviour.
- On Artificial datasets PFA's extracted approximates the LM-RNN almost perfectly.

Conclusions

- Hidden states and its Space has information to understand LM-RNN behaviour.
- On Artificial datasets PFA's extracted approximates the LM-RNN almost perfectly.
- On Real world datasets PFA's extracted very closely approximates the LM-RNN.

Conclusions

- Hidden states and its Space has information to understand LM-RNN behaviour.
- On Artificial datasets PFA's extracted approximates the LM-RNN almost perfectly.
- On Real world datasets PFA's extracted very closely approximates the LM-RNN.
- From entropy analysis, PFA's are fairly deterministic.
- Zhang, Xiyue, et al. "Decision-Guided Weighted Automata Extraction from Recurrent Neural Networks." 2021 [ZDX+21]

Thank you

References I

- Colah, *Understanding lstm networks*, `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015.
- Richard Gall, *Machine learning explainability vs interpretability: Two concepts that could help restore trust in ai*, KDnuggets News **19** (2019), no. 1.
- Christoph Molnar, *Interpretable machine learning*, Lulu. com, 2020.
- H Francis Song, Guangyu R Yang, and Xiao-Jing Wang, *Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework*, PLoS computational biology **12** (2016), no. 2, e1004792.

References II

Xiyue Zhang, Xiaoning Du, Xiaofei Xie, Lei Ma, Yang Liu, and Meng Sun, *Decision-guided weighted automata extraction from recurrent neural networks*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 11699–11707.