Interpretable Online Banking Fraud Detection based on Hierarchical Attention Mechanism

Idan Achituve[†], *Sarit Kraus*[†] and *Jacob Goldberger*[‡]

[†] Dept. of Computer Science, Bar-Ilan University, Israel
[‡] Faculty of Engineering, Bar-Ilan University, Israel

Main Contribution

- An attention based network that efficiently integrates cues from a sequence of transactions into a global fraud decision yielding improved detection results.
- Interpretability - The decision made by our system can be explained in comprehensible to users terms.

Online Banking Fraud Detection

- Real time detection - allow/deny/authenticate transactions.
- Interpretability of the classifier decision is required - banks need to explain it to their customers.
- High imbalance class distribution - 0.03% of transactions are labeled as fraudulent.
- Fraudster transactions may be interleaved within normal user transactions.
- labels are provided by a bank's analyst.
- There is a natural order in the data - a sequence based classifier is needed.
- Previous works either make a Markovian assumption or use complex modeling with RNNs.
- **Goal**: Real time F/G decision for each transaction based on sequence of previous user's transactions while being able to explain the classifier decisions.

Table 1: An example of a fraudulent sequence

Time	Type	OS	Browser	...	Label
2017-06-01 15:32:00	Login	Windows	Chrome	...	G
2017-06-01 15:34:50	Payment	Windows	Chrome	...	G
2017-06-03 15:14:22	Login	Windows	Firefox	...	F
2017-06-03 15:16:10	Change Phone	Windows	Firefox	...	F
2017-06-05 15:00:39	Login	Windows	Chrome	...	G
2017-06-06 15:42:25	Login	Windows	Chrome	...	F
2017-06-06 15:43:51	Payment	Windows	Chrome	...	F

Transaction Level Processing

- **Input**: A transaction sequence $S = (r_1, ..., r_m)$ where each transaction is made of k categorical features: $r_t = (f_{t1}, ..., f_{tk})$.
- **Feature embedding**:

$$e_{ti} = M_i f_{ti}, \quad i = 1, .., k, \quad t = 1, .., m$$

- **Feature level attention**:

$$\alpha_{ti} = \frac{\exp(w^\top \cdot g(e_{ti}))}{\sum_{j=1}^{k} \exp(w^\top \cdot g(e_{tj}))}$$

$$x_t = \sum_{i=1}^{k} \alpha_{ti} \cdot e_{ti}$$

Sequence Level Decision

- **Transaction level decisions**:

$$p(y = F|x_t) = \sigma(h(x_t)), \quad t = 1, ..., m$$

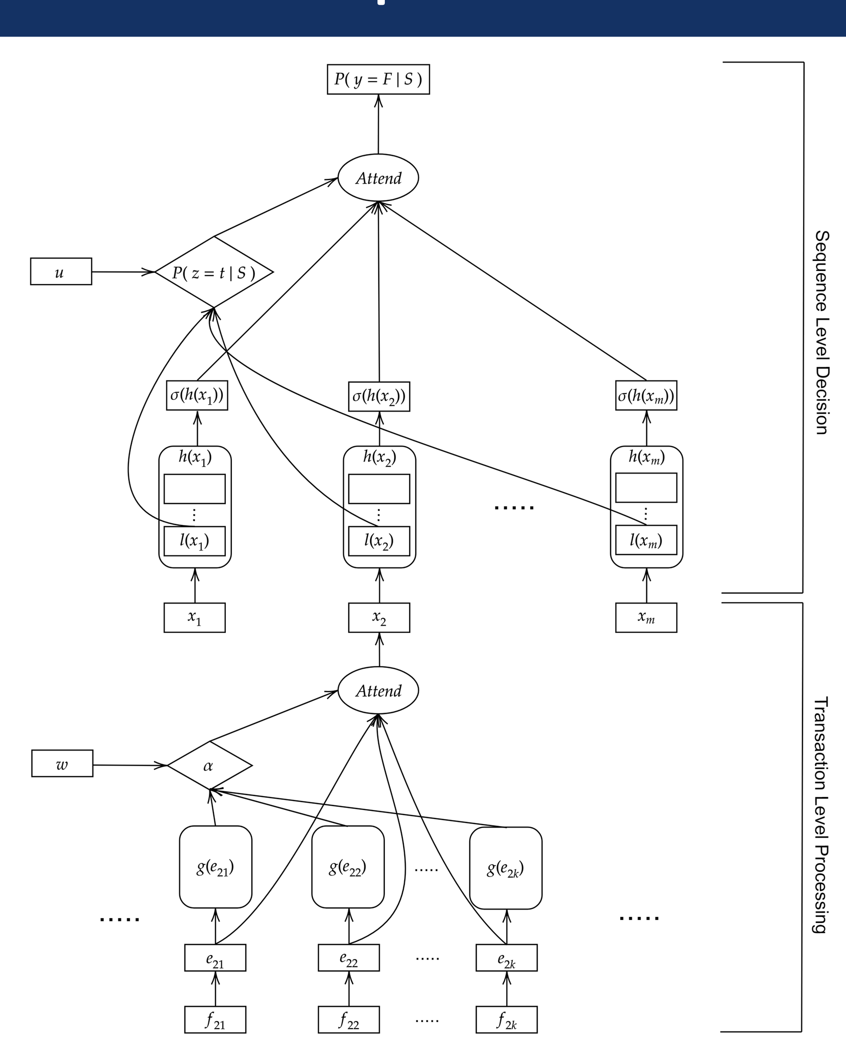
- **Sequence level attention**:

$$p(z = t|S) = \frac{\exp(u^\top \cdot l(x_t))}{\sum_{j=1}^{m} \exp(u^\top \cdot l(x_j))}$$

- **Weighted averaging of local decisions**:

$$p(y = F|S) = \sum_{t=1}^{m} p(z = t|S) p(y = F|x_t)$$

Features and Sequence Attention Model



Compared Classifiers

- **Last Transaction (Last-T)** Attention over the features and using a FC network.
- **Decaying Weight (DW)** Attention over the features and a fixed weighted averaging of the sequence items with decaying parameter of 1.5.
- **Features Attention (F-Attn)** Attention over the features and unweighted averaging of the items in the sequence.
- **LSTM** Attention over the features and using LSTM to process sequences.
- **Sequence Attention (S-Attn)** Attention over the transactions in the sequence and unweighted feature averaging.
- **Features and Sequence Attention (FS-Attn)** Our proposed method of applying attention over both the features and the transactions.

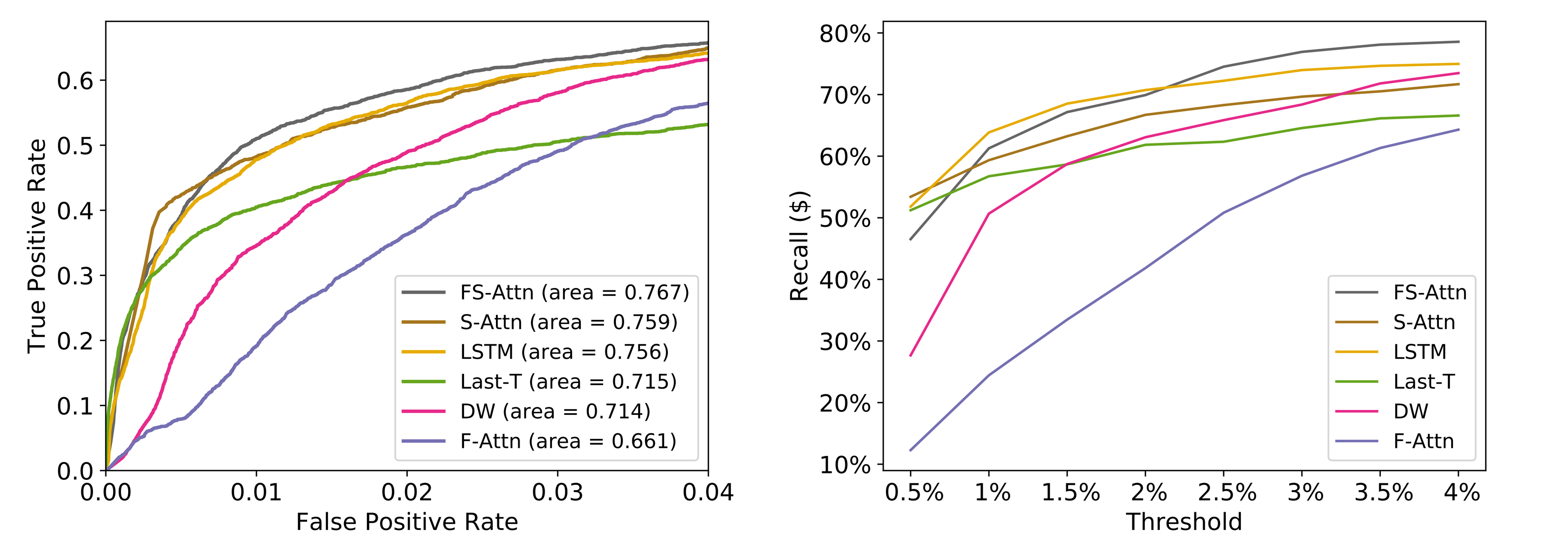


Figure 1: Area under the top 4% FPR of the ROC curve (left) and recall in US dollars presented in percentages on 8 thresholds (right).

Attention Mechanism Analysis

A-transaction - the transaction with the highest attention weight in the sequence. Given a sequence $S = (x_1, ..., x_m)$, the index of the A-transaction is:

$$\text{index} = \arg\max_t p(z = t|S) = \arg\max_t (u^\top l(x_t))$$

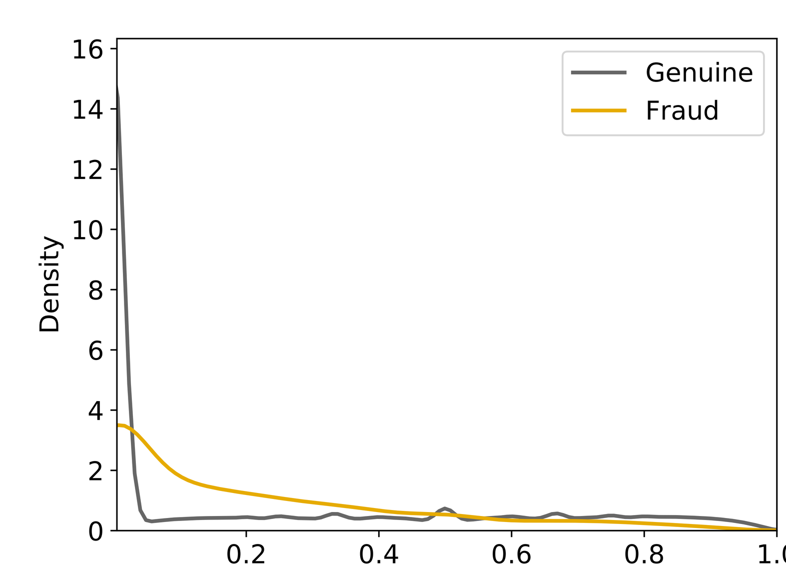


Figure 2: Distribution of the location of the A-transaction for each class.

Table 2: Use cases

Seq. Class	Highest Weighted Transactions	Highest Weighted Features	Details
F	6,8,9	Time features, transaction type, amount, beneficiary	**True positive**, The last transactions in the sequence were all payments attempts committed by the fraudster. All the transactions were executed in a short period of time, the amount at each transaction was slightly different, and the beneficiary was the same in most of them.
F	8	Change information, device details	**True positive**, The fraudster tried to fool the system by changing some of the user's personal information in the 8^{th} transaction. In addition, in both the 8^{th} transaction and the last transaction the fraudster connected from a device that differed from previous devices of the user.
F	10	Location, device details	**False Negative**, The sequence extended across several days and only the last transaction was a fraud attempt. The information in the last transaction was similar to information in previous transactions (e.g., device elements, location) and there wasn't anything unusual in the sequence or in the last transaction. Therefore we speculate that the system considered the sequence as genuine.

Conclusions

- Real time fraud detection using an attention based classifier.
- Improved performance compared to standard techniques.
- Interpretable model by identifying transactions and features that contributed the most to the final decision.