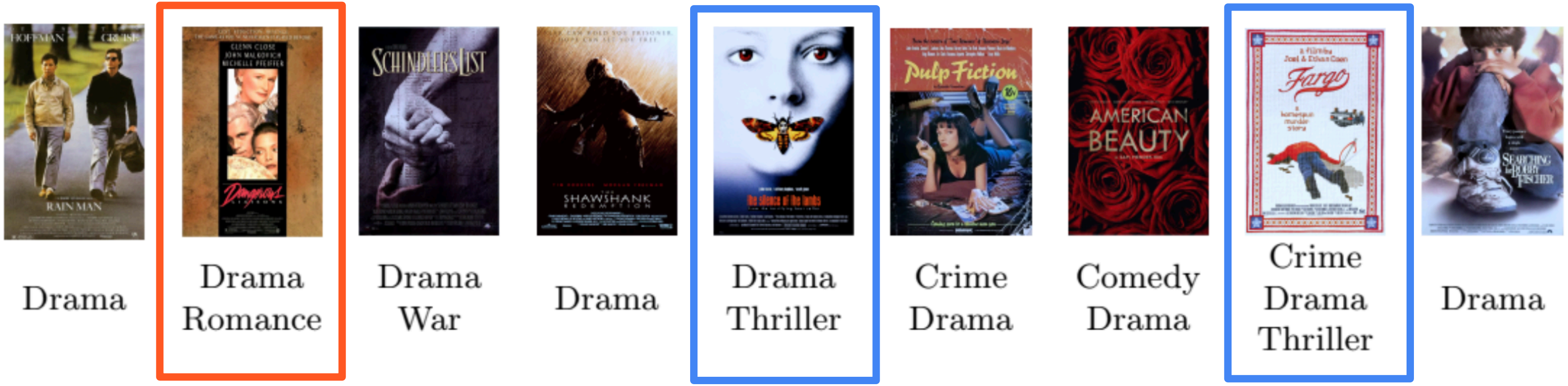
Towards Accurate and Interpretable Sequential Prediction: A CNN and Attention-Based Feature Extractor

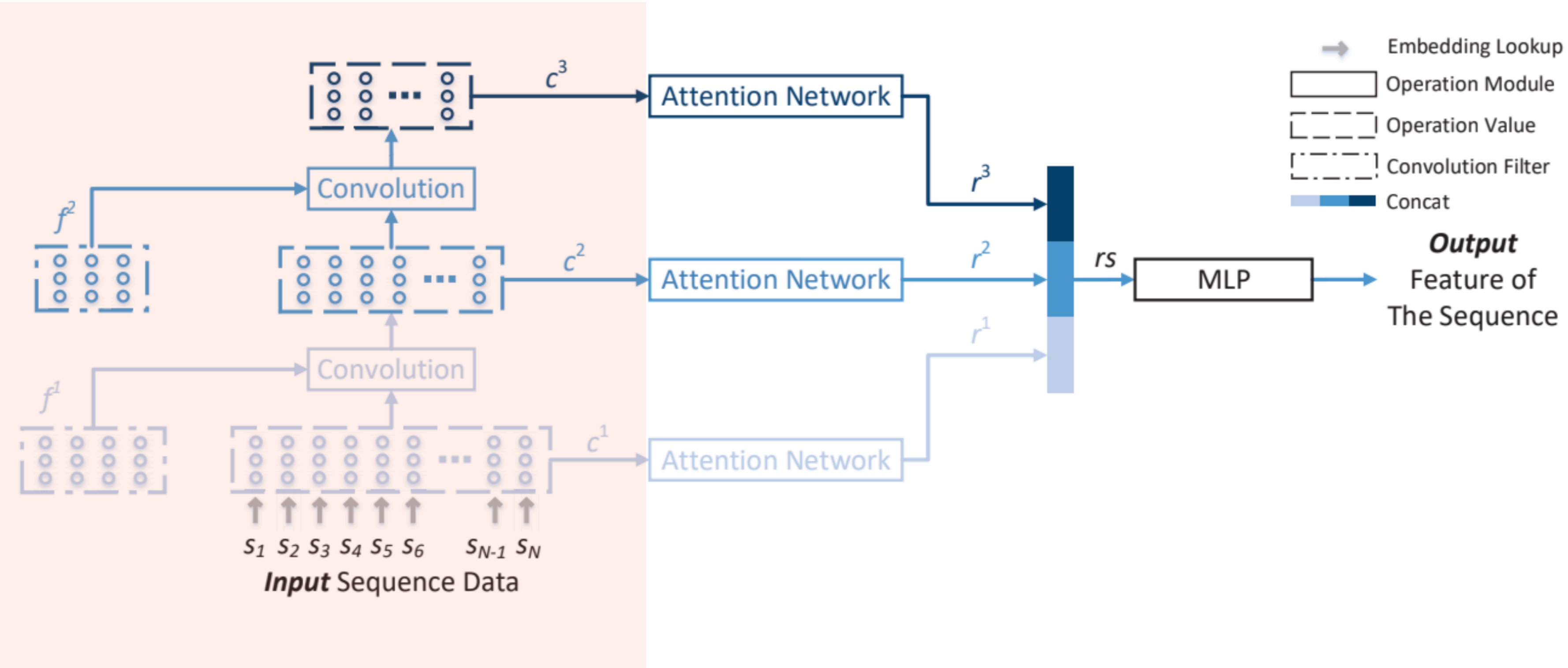
ADVISOR: JIA-LING, KOH
SPEAKER: WEI, LAI
SOURCE: CIKM’ 19
DATE: 2021/1/12

Introduction Sequential Prediction



- Few models could capture the possible sequential features during time periods with different fixed length.

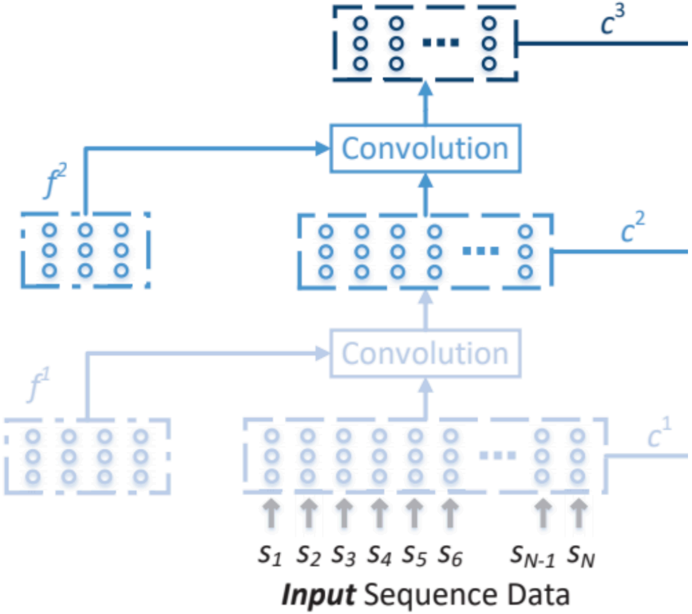
Method

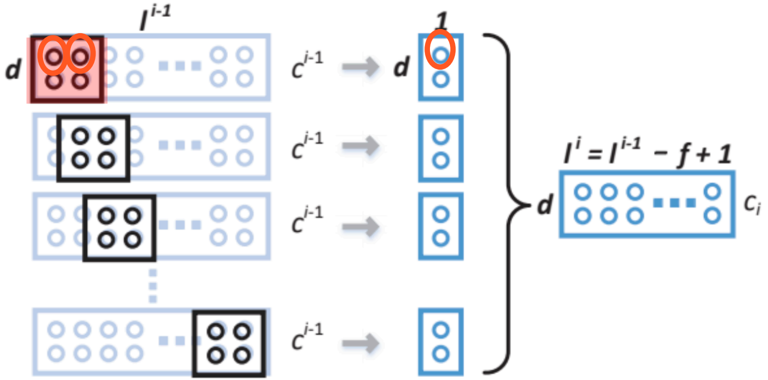


Embedding Lookup
Operation Module
Operation Value
Convolution Filter
Concat
Attention Network
Attention Network
Attention Network
Convolution
Convolution
c^3
c^2
c^1
f^2
f^1
r^3
r^2
r^1
rs
MLP
Output Feature of The Sequence
S_1 S_2 S_3 S_4 S_5 S_6 S_{N-1} S_N
Input Sequence Data

$$c_{j,k}^{i} = \sum_{l=k}^{k+f_w-1} c_{j,l}^{i-1} * f_{j,l}^{i-1}$$

Convolution Layer

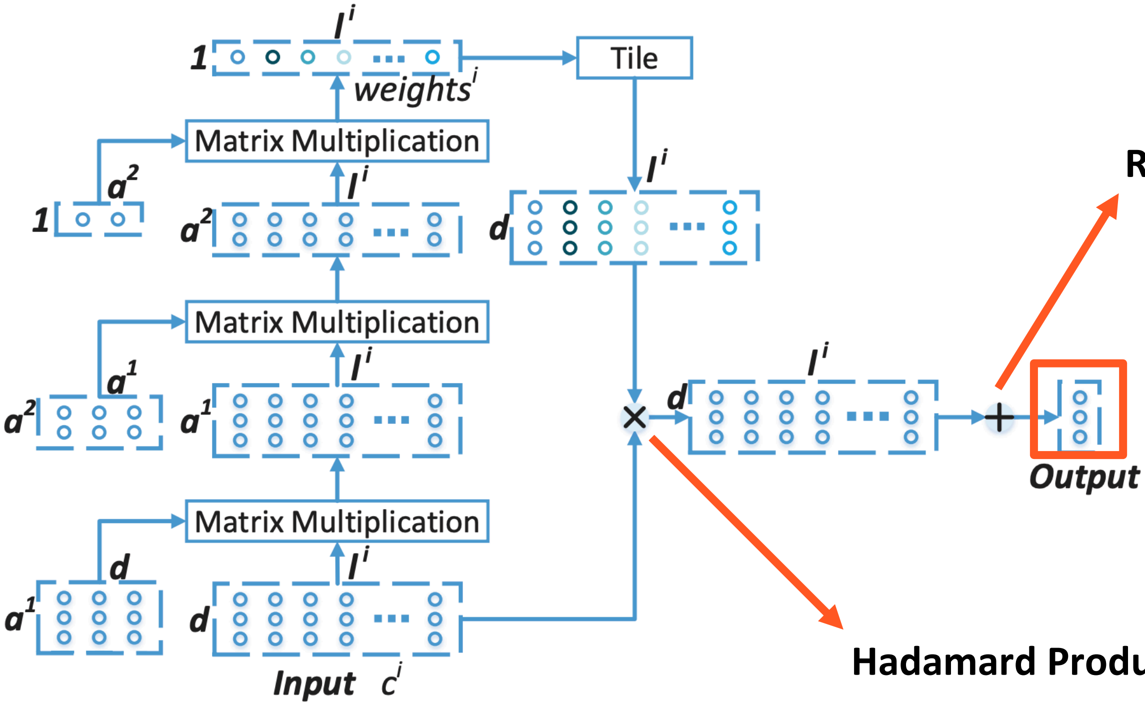




Item embedding ($e_t \in R^d$, 1≤ $t \le N$)

Item

Attention Layer



Row Sum $B_i = \sum_{j=1}^{n} A_{i,j}$

$\boldsymbol{r^i \in d * 1}$

Hadamard Product $(A * B)_{i,j} = A_{i,j} * B_{i,j}$

Loss Function

$$z = y \times D^T. \quad \Rightarrow \quad \hat{p}_i = \frac{e^{z_i}}{\sum_{j=1}^{|\boldsymbol{I}|} e^{z_j}}.$$

$$Loss(\hat{p}) = -\sum_{i=1}^{|\boldsymbol{I}|} p_i * \log(\hat{p}_i)$$

Dataset

1000 users

6000 users on 4000 movies

	Music						MovieLens				
DataSet	m5	l5	l10	l20	l50	l100	n5	n10	n20	ca	cb
Sequences	0.616M	2.101M	1.050M	0.525M	0.209M	0.104M	0.113M	0.055M	0.026M	0.946M	0.946M

Evaluation Metrics

Test set
(1, 2, 6) ➔ 1
(2, 4, 5) ➔ 4
(1, 3, 5) ➔ 2

Prediction
(1, 2, 6) ➔ 1, 2, 3, 4, 5, 6
(2, 4, 5) ➔ 2, 4, 1, 3, 6, 5
(1, 3, 5) ➔ 3, 4, 1, 5, 6, 2

- Recall @ 5 = (1+1+0)/3
- Mrr @ 5 = 1/1 + 1/2 + 1/6
- NDCG @5 of (1, 2, 6) test pair

$$= \left(\frac{2^1-1}{log(1+1)} + \frac{2^0-1}{log(2+1)} + \frac{2^0-1}{log(3+1)} + \frac{2^0-1}{log(4+1)} + \frac{2^0-1}{log(5+1)}\right)$$
$$/ \left(\frac{2^1-1}{log(1+1)} + \frac{2^0-1}{log(2+1)} + \frac{2^0-1}{log(3+1)} + \frac{2^0-1}{log(4+1)} + \frac{2^0-1}{log(5+1)}\right)$$

- NDCG @5 of (2, 4, 5) test pair

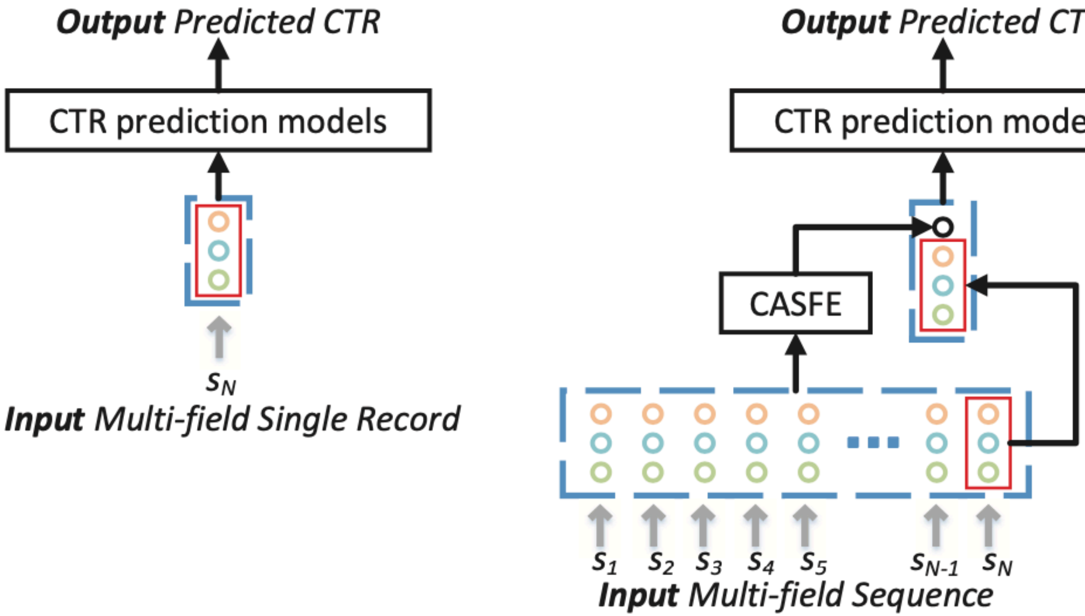
$$= \left(\frac{2^0-1}{log(1+1)} + \frac{2^1-1}{log(2+1)} + \frac{2^0-1}{log(3+1)} + \frac{2^0-1}{log(4+1)} + \frac{2^0-1}{log(5+1)}\right)$$
$$/ \left(\frac{2^1-1}{log(1+1)} + \frac{2^0-1}{log(2+1)} + \frac{2^0-1}{log(3+1)} + \frac{2^0-1}{log(4+1)} + \frac{2^0-1}{log(5+1)}\right)$$

Comparison against Baseline

Recommend popular item

Metric	Model	Music						MovieLens		
		m5	l5	l10	l20	l50	l100	n5	n10	n20
Recall@5	MostPop	0.0078	0.0031	0.0027	0.0031	0.0020	0.0013	0.0281	0.0238	0.0156
	GRURec	0.3269	0.2414	0.2564	0.2689	0.2633	0.2603	0.0820	0.1191	0.0898
	Caser	0.2812	0.2353	0.2623	0.2659	0.2534	0.2400	0.2188	03125	0.1875
	NextItNet	0.4050	0.2250	0.3326	0.3501	0.3477	0.3342	0.2812	0.2500	0.2500
	STAMP	0.4057	0.3223	0.3464	0.3582	0.3643	0.3407	**0.3438**	0.3750	0.2812
	CASFE	**0.4091**	**0.3940**	**0.3572**	**0.3651**	**0.3698**	**0.3744**	0.2812	**0.3750**	**0.2812**
Mrr@5	MostPop	0.0041	0.0006	0.0014	0.0011	0.0009	0.0008	0.0140	0.0104	0.0077
	GRURec	0.2593	0.1910	0.1863	0.1899	0.1851	0.1851	0.0442	0.0635	0.0529
	Caser	0.2354	0.2021	0.2123	0.2121	0.1932	0.1870	0.1719	0.1698	0.1562
	NextItNet	0.3302	0.2690	0.2844	0.2896	0.2988	0.2904	0.2005	0.2188	0.1469
	STAMP	0.3345	0.2675	0.2854	0.2948	0.3017	0.2812	0.2078	**0.2485**	0.1807
	CASFE	**0.3402**	**0.2749**	**0.2974**	**0.3012**	**0.3050**	**0.3085**	**0.2448**	0.2302	**0.2292**
NDCG@5	MostPop	0.0050	0.0012	0.0017	0.0016	0.0012	0.0010	0.0197	0.0137	0.0096
	GRURec	0.2762	0.2036	0.2037	0.2088	0.2046	0.2039	0.0536	0.0772	0.0621
	Caser	0.2465	0.2209	0.2248	0.2255	0.2082	0.2003	0.1841	0.2038	0.1644
	NextItNet	0.3489	0.2361	0.2980	0.3014	0.3106	0.3079	0.2207	0.2269	0.1722
	STAMP	0.3522	0.2812	0.3006	0.3106	0.3174	0.2965	0.2405	**0.2775**	0.2056
	CASFE	**0.3574**	**0.2882**	**0.3123**	**0.3172**	**0.3212**	**0.3250**	**0.2571**	0.2659	**0.2426**

Compatibility of CASFE on CTR Prediction

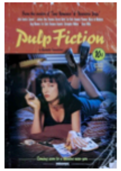
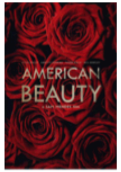


Compatibility of CASFE on CTR Prediction

$$Log\ loss = -\frac{1}{N}\sum y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Model	AUC	Log loss	Improvement
IPNN	0.8010	0.4327	2.2‰
IPNN+CASFE	**0.8028**	**0.4253**	
DNN	0.8018	0.4381	1.4‰
DNN+CASFE	**0.8027**	**0.4301**	
DeepFM	0.8059	0.4191	5.8‰
DeepFM+CASFE	**0.8106**	**0.3837**	

Attention Visualization

<table>
<tr><td>movie id</td><td>1961</td><td>2020</td><td>527</td><td>318</td><td>593</td><td>296</td><td>2858</td><td>608</td><td>529</td><td>3006</td></tr>
<tr><td>movie poster</td><td>
</td><td></td><td>
</td><td>
</td><td></td><td>
</td><td>
</td><td>
</td><td></td><td></td></tr>
<tr><td>movie category</td><td>Drama</td><td>Drama Romance</td><td>Drama War</td><td>Drama</td><td>Drama Thriller</td><td>Crime Drama</td><td>Comedy Drama</td><td>Crime Drama Thriller</td><td>Drama</td><td>Drama</td></tr>
<tr><td>attention weights 1</td><td>0.115</td><td>0.105</td><td>0.056</td><td>0.130</td><td>0.097</td><td>0.099</td><td>0.149</td><td>0.069</td><td>0.180</td><td></td></tr>
<tr><td rowspan="6">attention weights 2</td><td colspan="4">0.210</td><td colspan="6"></td></tr>
<tr><td></td><td colspan="4">0.153</td><td colspan="5"></td></tr>
<tr><td colspan="2"></td><td colspan="4">0.152</td><td colspan="4"></td></tr>
<tr><td colspan="3"></td><td colspan="4">0.198</td><td colspan="3"></td></tr>
<tr><td colspan="4"></td><td colspan="4">0.149</td><td colspan="2"></td></tr>
<tr><td colspan="5"></td><td colspan="4">0.138</td><td></td></tr>
<tr><td rowspan="3">attention weights 3</td><td colspan="7">0.381</td><td colspan="3"></td></tr>
<tr><td></td><td colspan="7">0.469</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td><td colspan="7">0.150</td><td></td></tr>
</table>

Conclusion

- The information of all CNN layer is used in order to capture the periodic features of user behavior. The deep layer corresponds to longer time periods.
- We apply attention mechanism after each CNN layer to focus on the important features.
- CASFE can be applied in CTR prediction.