FinalMLP: An Enhanced Two-Stream MLP Model for CTR Prediction

Kelong Mao[1*], Jieming Zhu[2*], Liangcai Su[3], Guohao Cai[2], Yuru Li[2], Zhenhua Dong[2]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Huawei Noah's Ark Lab
[3]Tsinghua University
kyriemkl@gmail.com, jiemingzhu@ieee.org, sulc21@mails.tsinghua.edu.cn

AAAI 2023

Code: MindSpore/models and Fux-iCTR/model zoo

Reported by liang li

Motivation

Details:

- Existing two-stream models often combine two streams via summation or concatenation, which may waste the opportunity to model the high-level (i.e., stream-level) feature interactions.
- In contrast, our empirical study shows that a well-tuned two-stream MLP model that simply combines two MLPs can even achieve surprisingly good performance.

Problem Statement

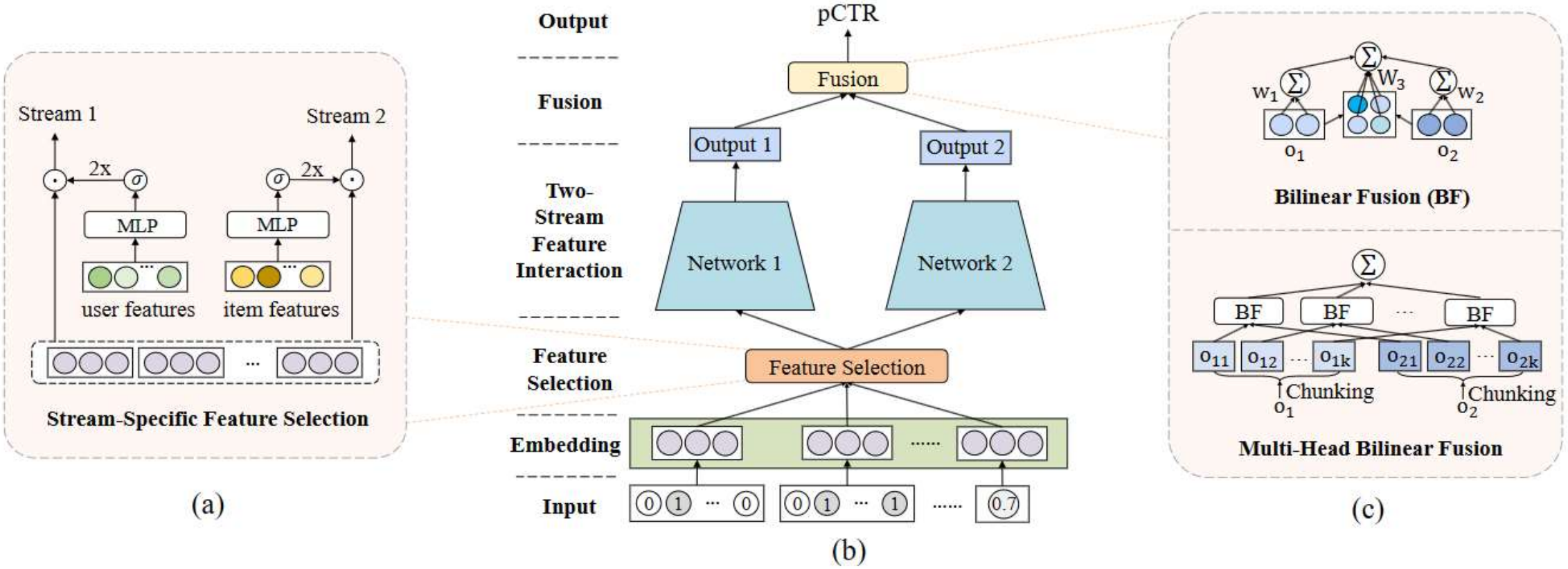


Figure 1: (a) An illustration of stream-specific feature selection. (b) A general framework of two-stream CTR models. (c) The multi-head bilinear fusion.

$$x = \{x_1, ..., x_M\}$$

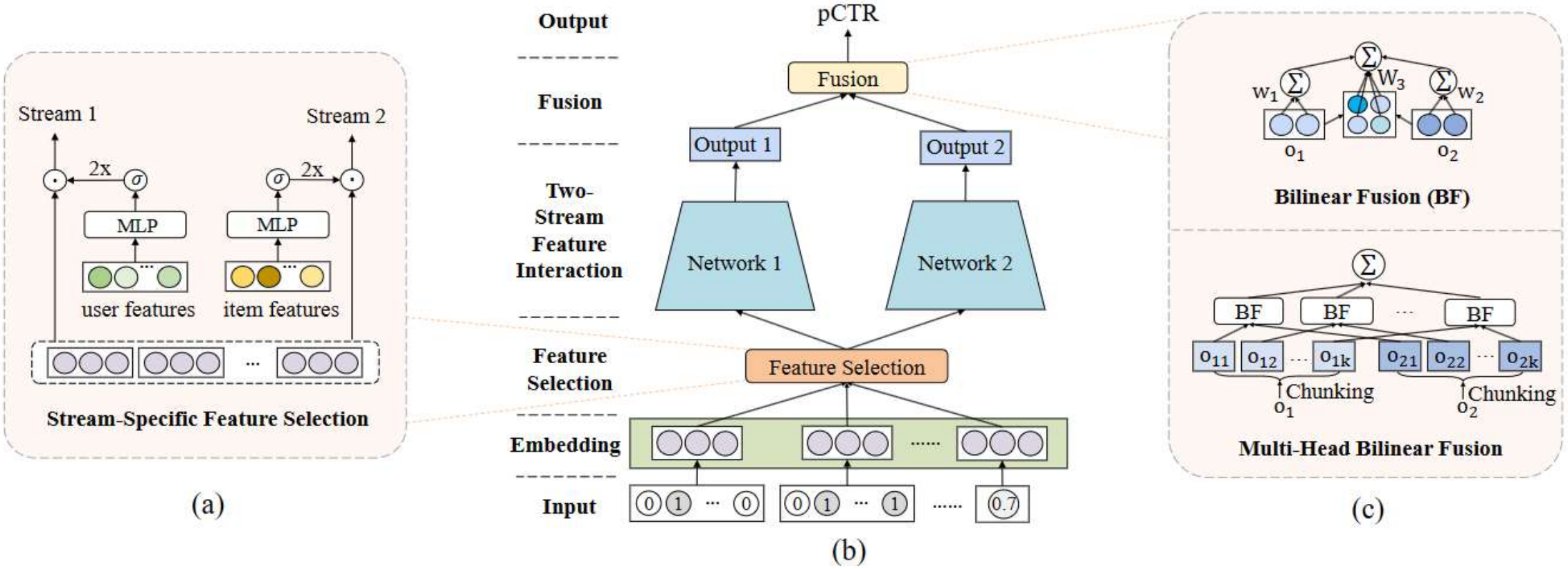


Figure 1: (a) An illustration of stream-specific feature selection. (b) A general framework of two-stream CTR models. (c) The multi-head bilinear fusion.

$$\mathbf{o_1} = MLP_1(\mathbf{h_1}), \quad (1)$$

$$\mathbf{o_2} = MLP_2(\mathbf{h_2}), \quad (2)$$

$$\mathbf{g_1} = Gate_1(\mathbf{x_1}), \quad \mathbf{g_2} = Gate_2(\mathbf{x_2}), \quad (3)$$

$$\mathbf{h_1} = 2\sigma(\mathbf{g_1}) \odot \mathbf{e}, \quad \mathbf{h_2} = 2\sigma(\mathbf{g_2}) \odot \mathbf{e}, \quad (4)$$

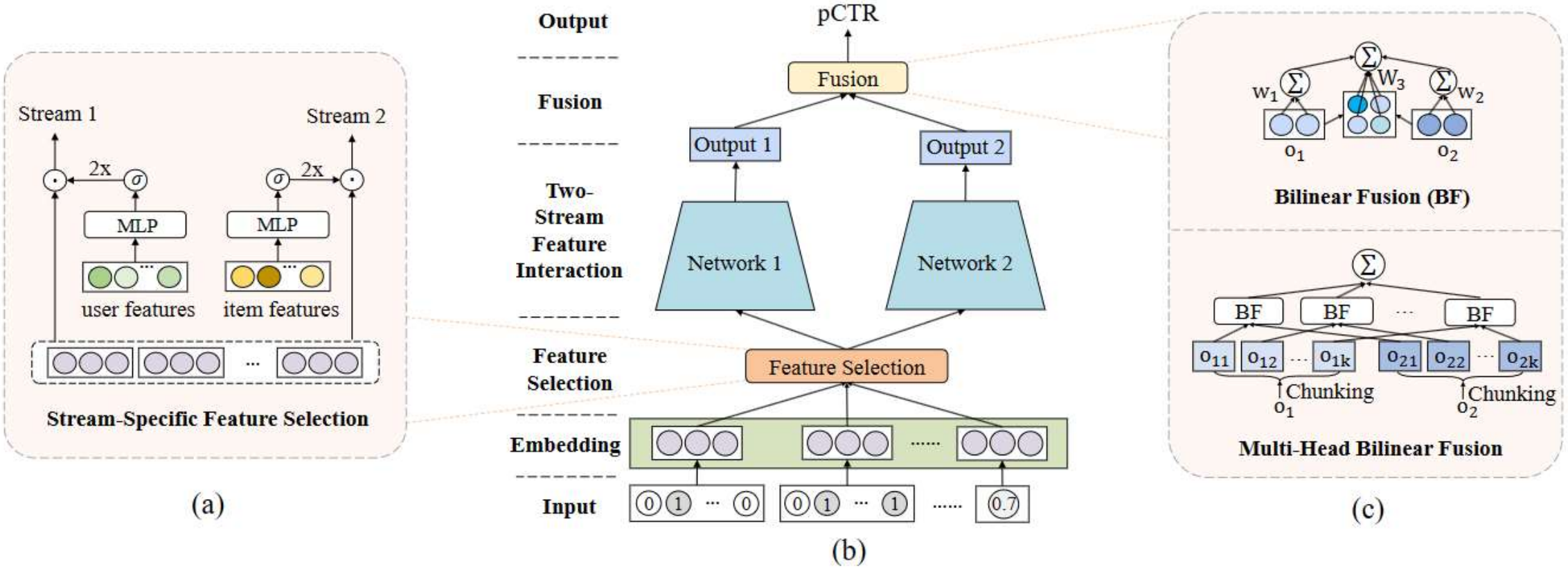


Figure 1: (a) An illustration of stream-specific feature selection. (b) A general framework of two-stream CTR models. (c) The multi-head bilinear fusion.

$$\hat{y} = \sigma(b + \mathbf{w}_1^T \mathbf{o}_1 + \mathbf{w}_2^T \mathbf{o}_2 + \mathbf{o}_1^T \mathbf{W}_3 \mathbf{o}_2), \quad (5)$$

$$\hat{y} = \sigma(b + \mathbf{w}^\top \mathbf{x} + \mathbf{x}^\top \mathrm{upper}(\mathbf{P}\mathbf{P}^\top)\mathbf{x}), \quad (6)$$

$$\mathbf{o_1} = [\mathbf{o_{11}}, \ldots, \mathbf{o_{1k}}], \quad (7)$$

$$\mathbf{o_2} = [\mathbf{o_{21}}, \ldots, \mathbf{o_{2k}}], \quad (8)$$

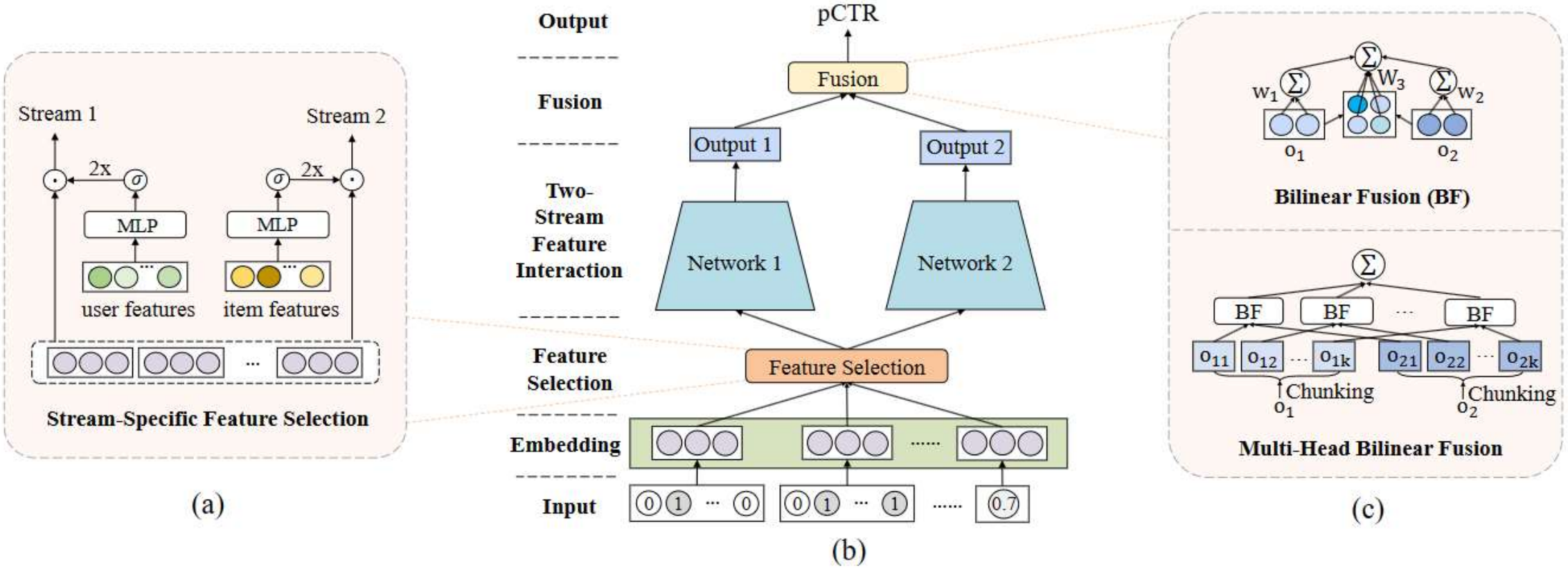


Figure 1: (a) An illustration of stream-specific feature selection. (b) A general framework of two-stream CTR models. (c) The multi-head bilinear fusion.

$$\hat{y} = \sigma(\sum_{j=1}^{k} BF(\mathbf{o_{1j}}, \mathbf{o_{2j}})), \quad (9)$$

$$\mathcal{L} = -\frac{1}{N}\sum \left(y log\hat{y} + (1-y) log(1-\hat{y})\right)$$

Dataset	#Instances	#Fields	#Features
Criteo	45,840,617	39	2,086,936
Avazu	40,428,967	22	1,544,250
MovieLens	2,006,859	3	90,445
Frappe	288,609	10	5,382

Table 1: The statistics of open datasets.

Dataset	Metric	WideDeep	DeepFM	DCN	xDeepFM	AutoInt+	AFN+	DeepIM	MaskNet	DCNv2	EDCN	DualMLP	FinalMLP
Criteo	AUC	81.38	81.38	81.39	81.39	81.39	81.43	81.40	81.39	81.42	81.47	81.42	**81.49**
	Std	5.7e-5	8.0e-5	4.9e-5	9.5e-5	1.4e-4	5.9e-5	5.9e-5	1.3e-4	2.0e-4	6.6e-5	5.6e-4	1.7e-4
Avazu	AUC	76.46	76.48	76.47	76.49	76.45	76.48	76.52	76.49	76.54	76.52	76.57	**76.66**
	Std	5.4e-4	4.4e-4	1.2e-3	4.1e-4	5.2e-4	3.7e-4	9.2e-5	2.6e-3	4.7e-4	3.0e-4	3.5e-4	4.9e-4
MovieLens	AUC	96.80	96.85	96.87	96.97	96.92	96.42	96.93	96.87	96.91	96.71	96.98	**97.20**
	Std	3.2e-4	1.6e-4	5.5e-4	9.0e-4	4.4e-4	5.8e-4	5.8e-4	2.8e-4	3.6e-4	3.4e-4	4.3e-4	1.8e-4
Frappe	AUC	98.41	98.42	98.39	98.45	98.48	98.26	98.44	98.43	98.45	98.50	98.47	**98.61**
	Std	7.9e-4	1.6e-4	3.1e-4	3.7e-4	7.9e-4	1.4e-3	6.3e-4	5.7e-4	4.3e-4	5.1e-4	3.5e-4	1.7e-4

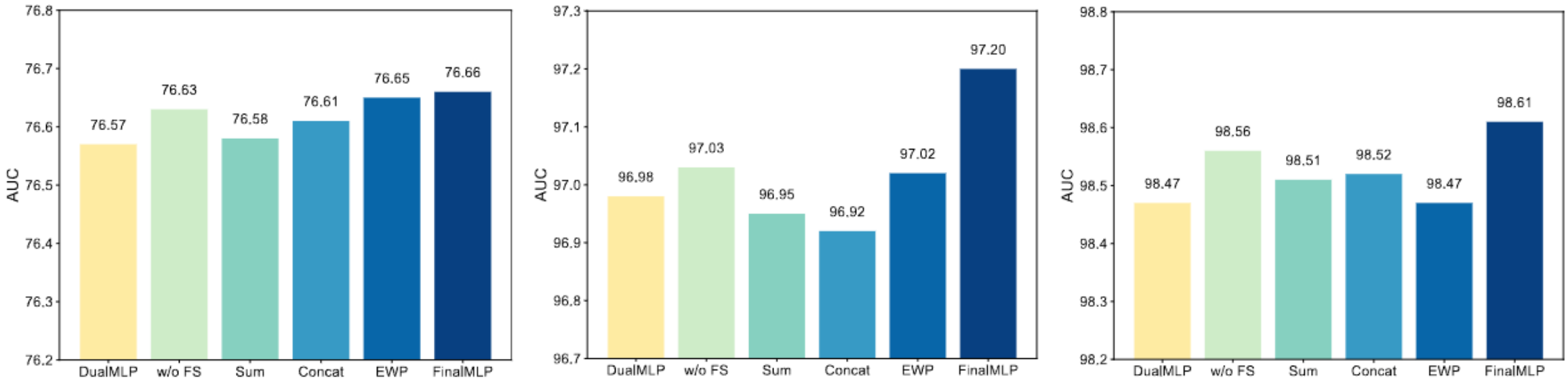
Table 2: Performance comparison of two-stream models for CTR prediction. The best results are in bold and the second-best results are underlined.

Experiments

Class	Model	Criteo	Avazu	MovieLens	Frappe
First-Order	LR	78.86	75.16	93.42	93.56
Second-Order	FM	80.22	76.13	94.34	96.71
	AFM	80.44	75.74	94.72	96.97
	FFM	80.60	76.25	95.22	97.88
	FwFM	80.63	76.02	95.58	97.76
	FmFM	80.56	75.95	94.65	97.49
Third-Order	HOFM(3rd)	80.55	76.01	94.55	97.42
	CrossNet(2L)	79.47	75.45	93.85	94.19
	CrossNetV2(2L)	81.10	76.05	95.83	97.16
	CIN(2L)	80.96	76.26	96.02	97.76
Higher-Order	CrossNet	80.41	75.97	94.40	95.94
	CrossNetV2	81.27	76.25	96.06	97.29
	CIN	81.17	76.24	<u>96.74</u>	97.82
	AutoInt	81.26	76.24	96.63	<u>98.31</u>
	FiGNN	<u>81.34</u>	76.22	95.25	97.61
	AFN	81.07	75.47	96.11	98.11
	SAM	81.31	**76.32**	96.31	98.01
	MLP	**81.37**	<u>76.30</u>	**96.78**	**98.33**

Table 3: Performance comparisons between MLP and explicit feature interaction networks. The best results w.r.t. AUC are in **bold** and the second-best results are <u>underlined</u>.

Experiments



(a) Avazu (b) MovieLens (c) Frappe

Figure 2: The ablation study results of FinalMLP.

#Heads (k)	Criteo	Avazu	MoiveLens	Frappe
1	OOM	0.7649	0.9691	**0.9862**
5	0.8141	0.7661	0.9707	0.9851
10	0.8144	**0.7669**	**0.9724**	0.9849
50	**0.8148**	0.7657	0.9703	0.9841

Table 4: Bilinear fusion with different numbers of heads.

Experiments

	BaseModel	EDCN	FinalMLP	
			#Heads=1	#Heads=8
AUC	71.78	72.22	72.83	72.93
ΔAUC	–	+0.44	+1.05	+1.15
Latency	45ms	–	70ms	47ms

Table 5: Offline results in production settings.

	Day1	Day2	Day3	Day4	Day5	Average
ΔCTR	1.6%	0.6%	1.7%	1.5%	2.4%	1.6%

Table 6: Online results of a five-day online A/B test.

Thanks