ICML | 2022

Thirty-ninth International Conference on Machine Learning

Flowformer: Linearizing Transformers with Conservation Flows

Haixu Wu [1] Jialong Wu [1] Jiehui Xu [1] Jianmin Wang [1] Mingsheng Long [1]

Haixu Wu

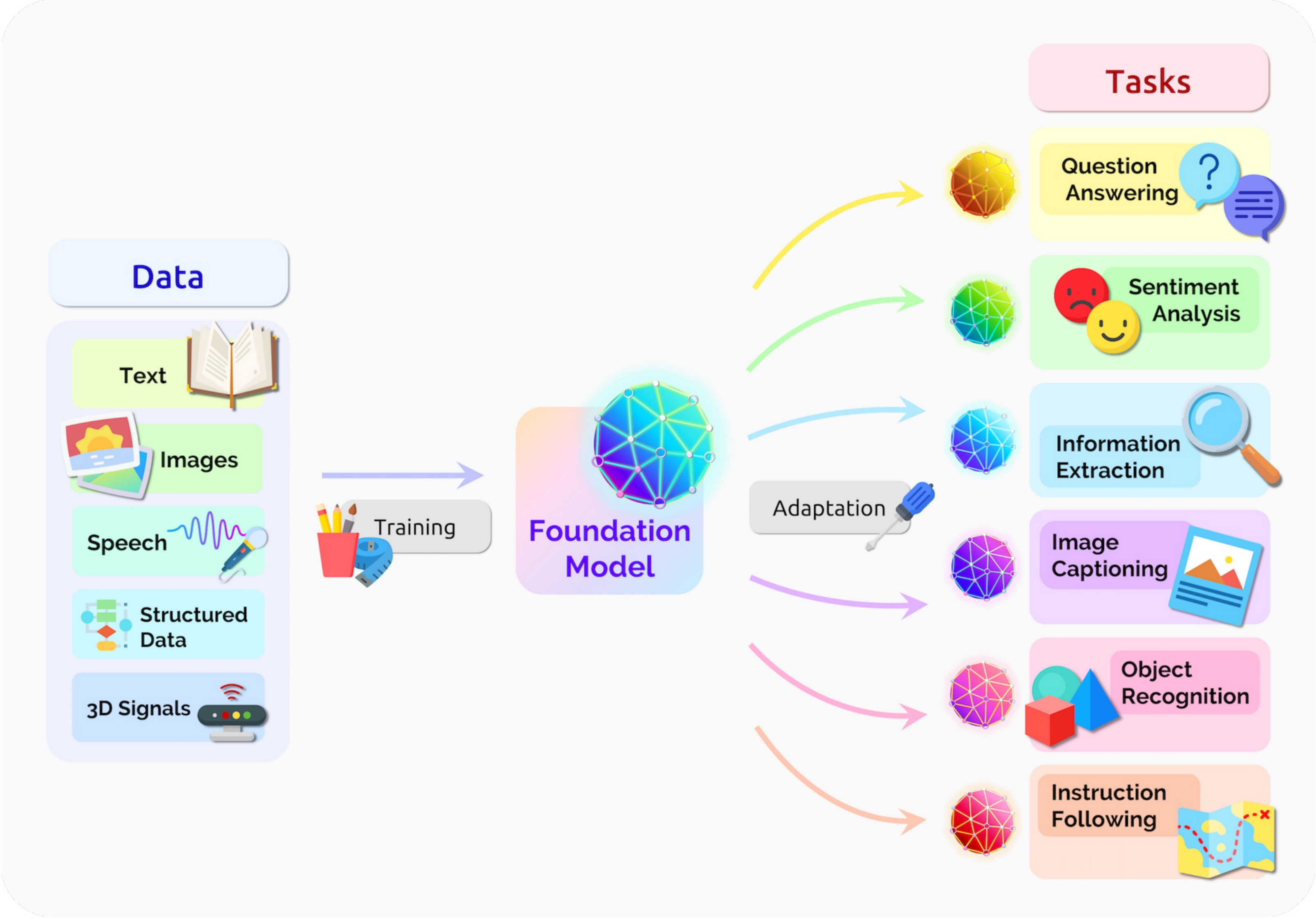
Jialong Wu

Jiehui Xu

Jianmin Wang

Mingsheng Long

Foundation Models



[Data Universal]

Learn from various modalities

[Task Universal]

Adapt to a wide range of downstream tasks

Bommasani et al. On the Opportunities and Risks of Foundation Models. *Arxiv 2021.*

A Universal Architecture for General Proposes

Image

Language

Time Series

Agent Trajectory

Universal Architecture

A Universal Architecture for General Proposes

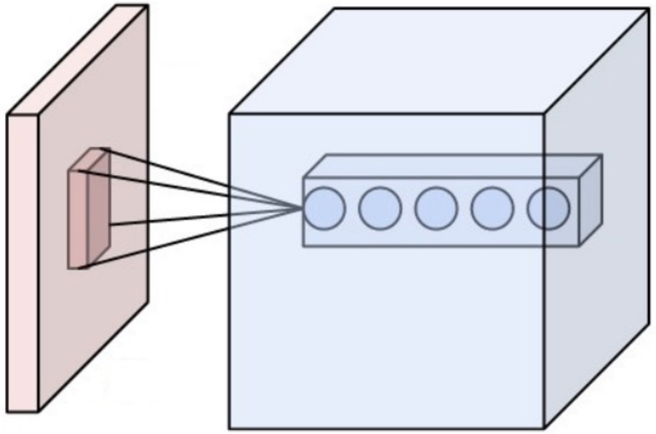
Image

Language

Time Series

Agent Trajectory

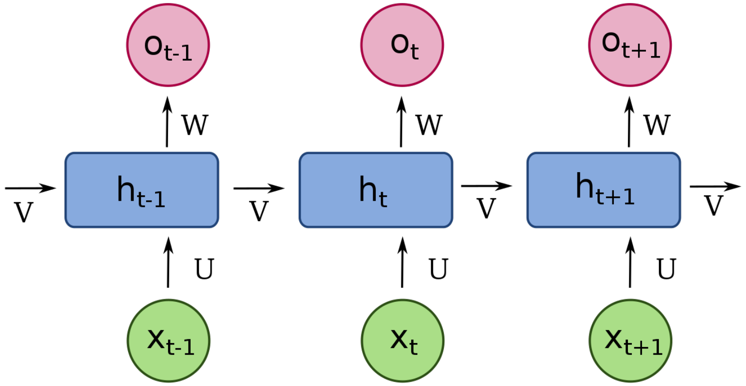
Universal Architecture



CNNs?

Locality

Shift Invariance ☹



RNNs?

Markov ☹

A Universal Architecture for General Proposes

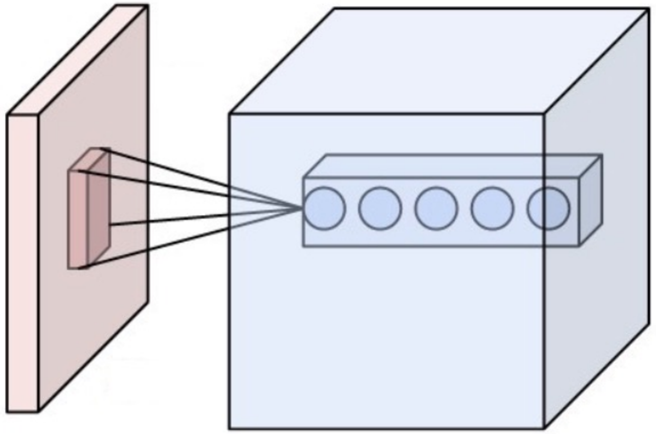
Image

Language

Time Series

Agent Trajectory

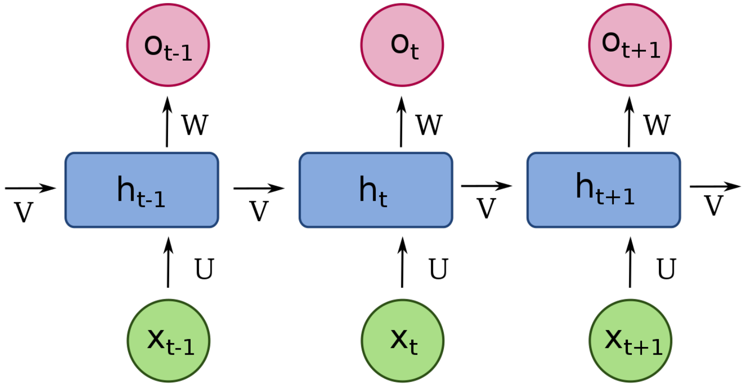
Universal Architecture



CNNs?

Locality

Shift Invariance ☹

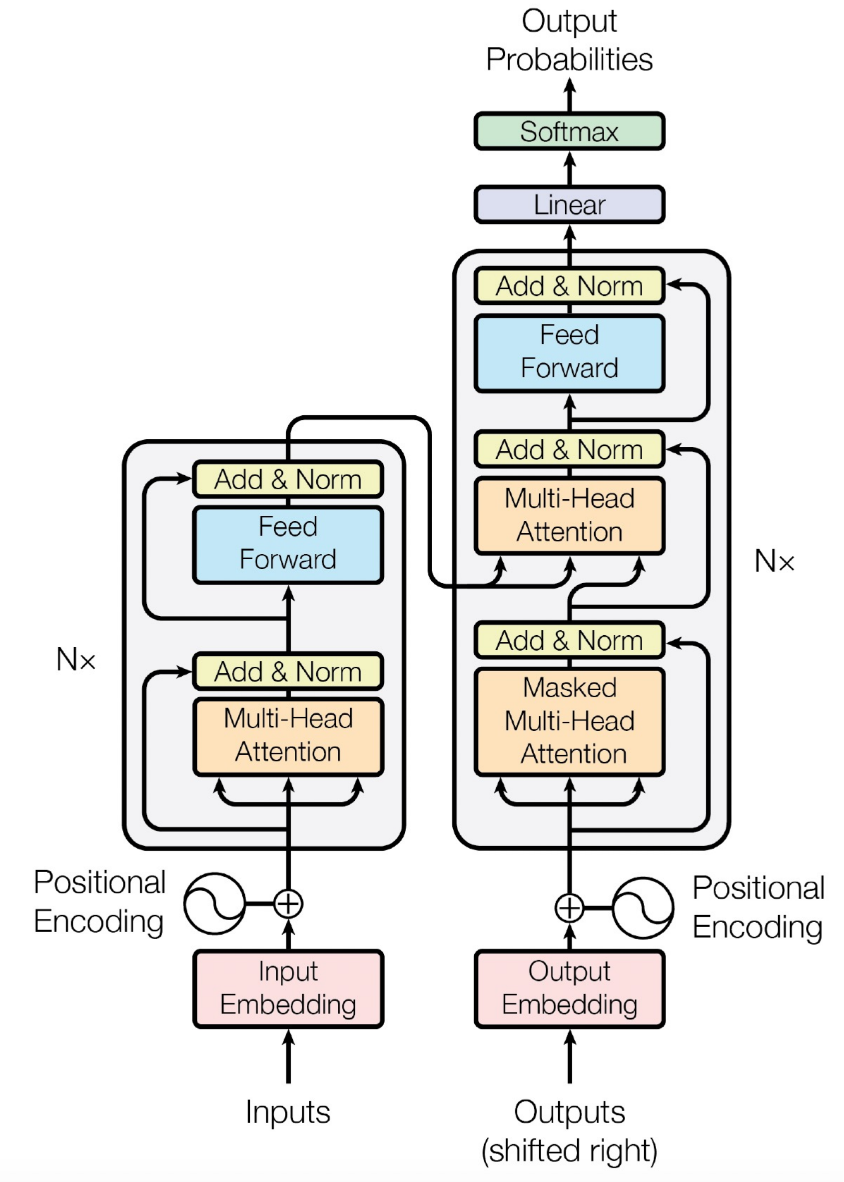


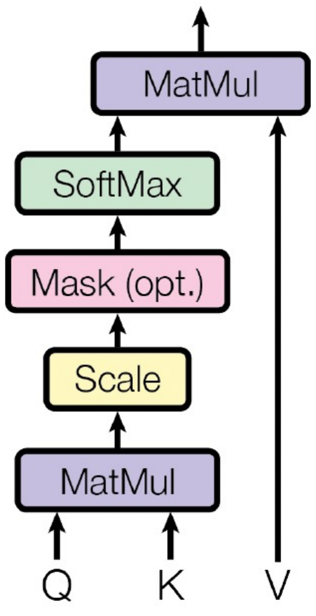
RNNs?

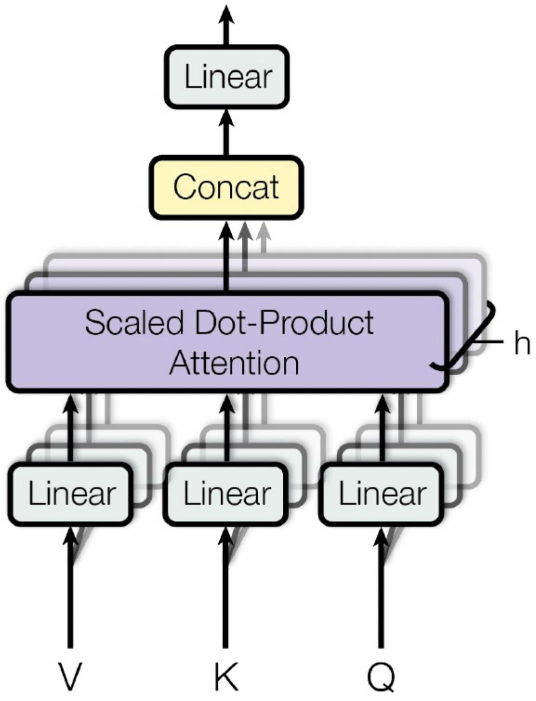
Markov ☹

Specific Inductive Biases Limit the Model Universality

Transformers







Self-Attention Multi-head **Self-Attention**

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Dot-product Similarity & Without Specific Inductive Biases

Vaswani et al. Attention is All you Need. *NeurIPS 2017.*

General Relation Modeling

Image

Relation among **Image Patches**

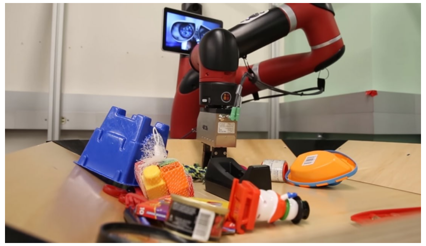
Language

Relation among **Words**

[SOS] Flowformer is a task-universal linear Transformer. [EOS]

Time Series

Relation among **Time Points**

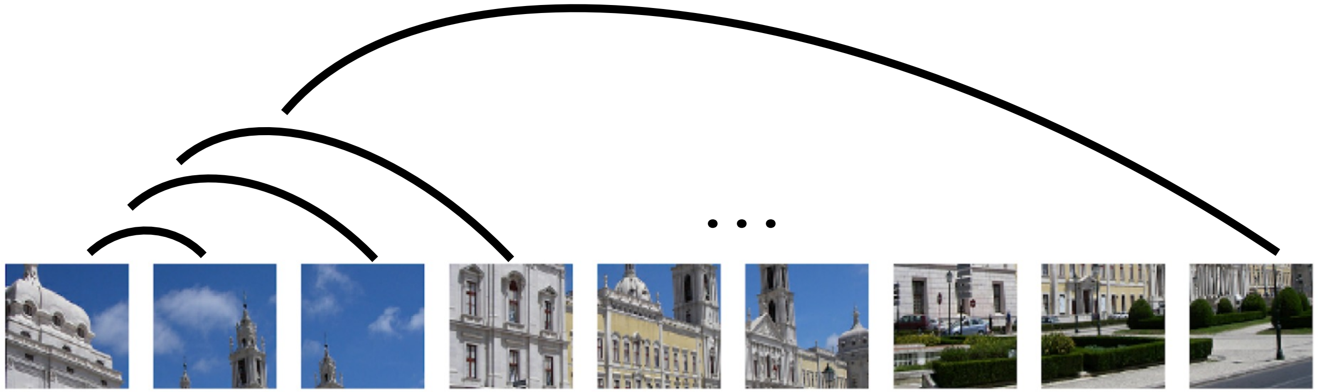


Agent Trajectory

Relation among **Agent-Environment Interactions**

… [S] [A] [R] …

Quadratic Complexity in Self-Attention

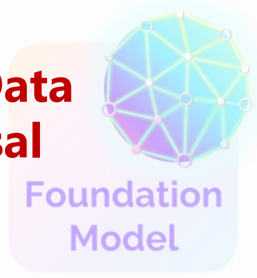


Pair-wise Relation Modeling: $\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

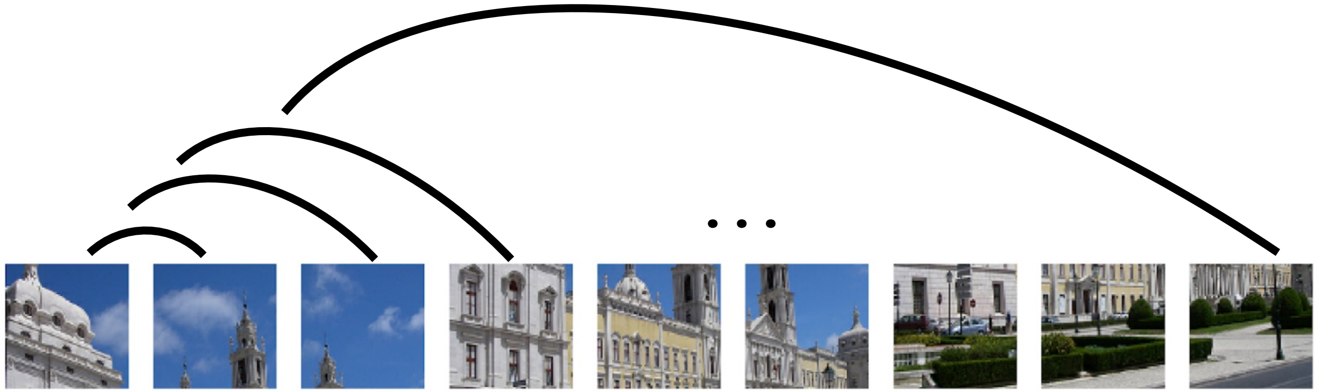
⇩

General Relation Modeling → **[Quadratic Complexity]** → **Task & Data Universal**

Long Sequence
Model Efficiency
Big Model ☹



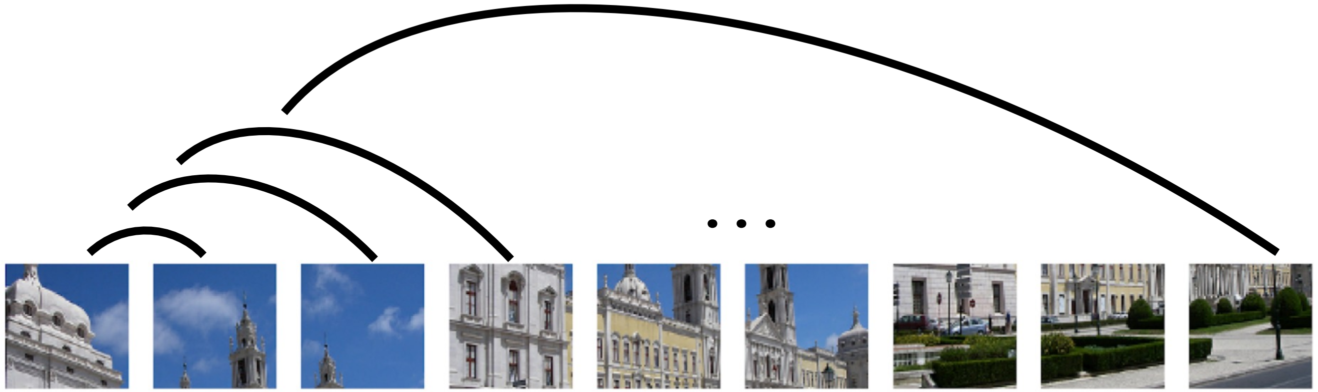
Quadratic Complexity in Self-Attention



Pair-wise Relation Modeling: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

$\mathcal{O}(n^2 d)$

Quadratic Complexity in Self-Attention



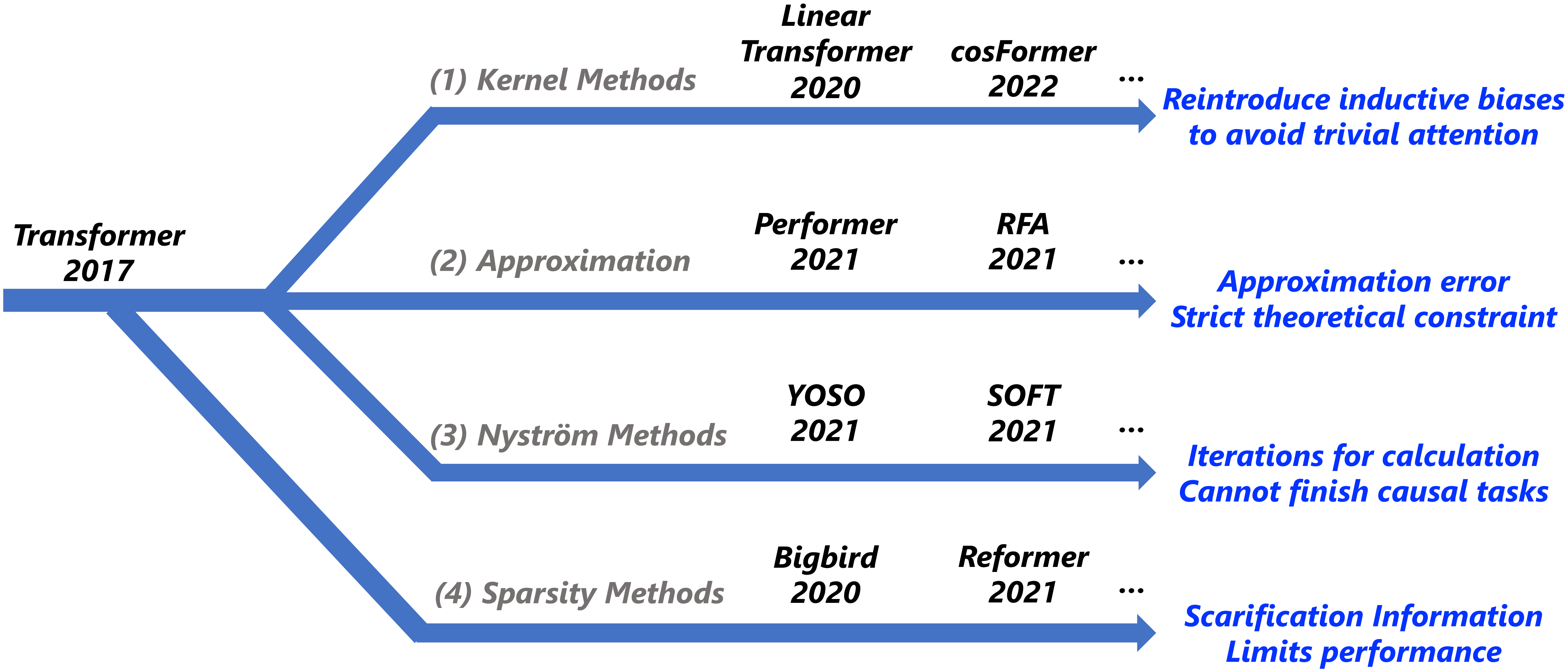
Pair-wise Relation Modeling: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

$\mathcal{O}(n^2 d)$

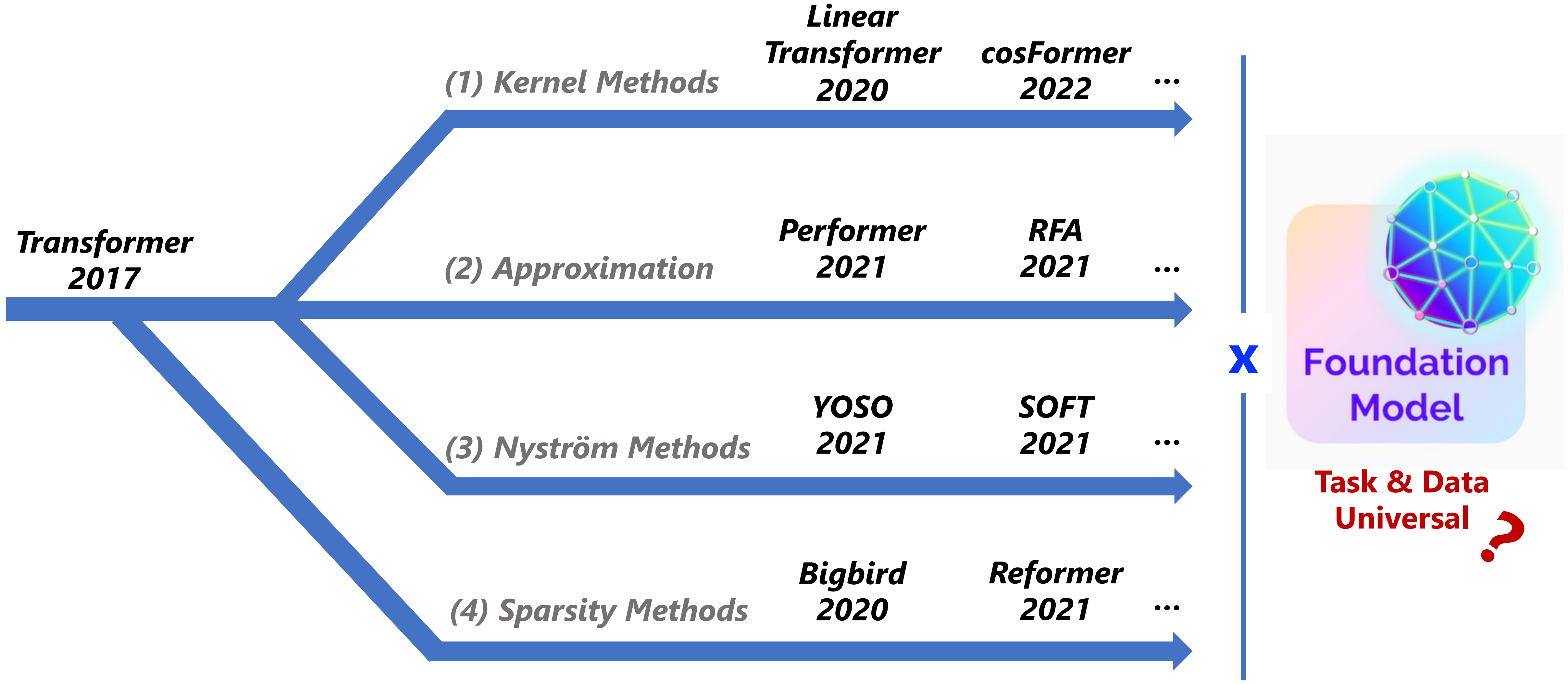
Can we remove Softmax function?

$(QK^T)V = Q(K^TV) \Rightarrow \mathcal{O}(n^2 d) \rightarrow \mathcal{O}(nd^2)$

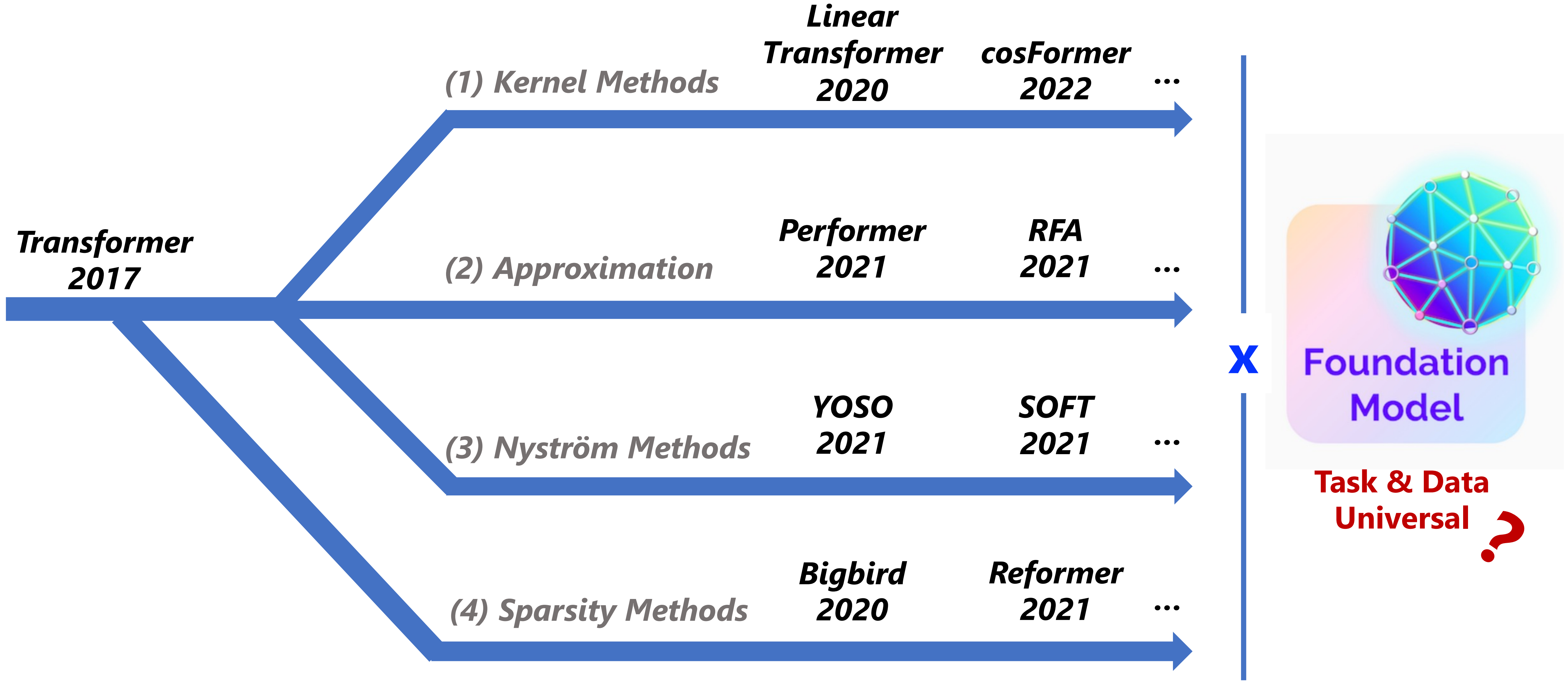
Linear Transformers



Linear Transformers



Linear Transformers



How to achieve the linear complexity and maintain the universality simultaneously?

Recap: Softmax Function

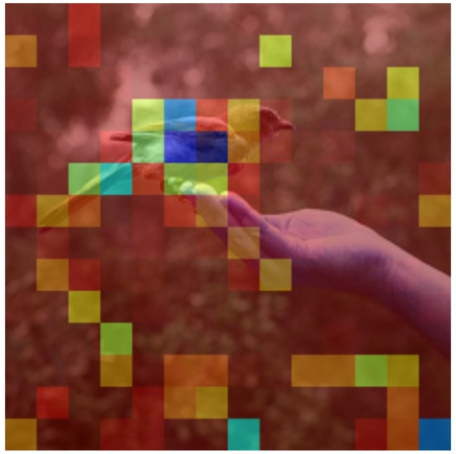
Softmax function is proposed as a differentiable generalization of the ***"winner-take-all"*** picking maximum operation.

⇓

Competition Mechanism

⇒

The key to avoid trivial attention



Bridle et al. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *NeurIPS 1989.*

Recap: Softmax Function

Softmax function is proposed as a differentiable generalization of the ***"winner-take-all"*** picking maximum operation.

$$\phi(Q)(\phi(K)^T V) + \textbf{Competition Mechanism} \iff \textbf{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Bridle et al. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *NeurIPS 1989.*

Recap: Softmax Function

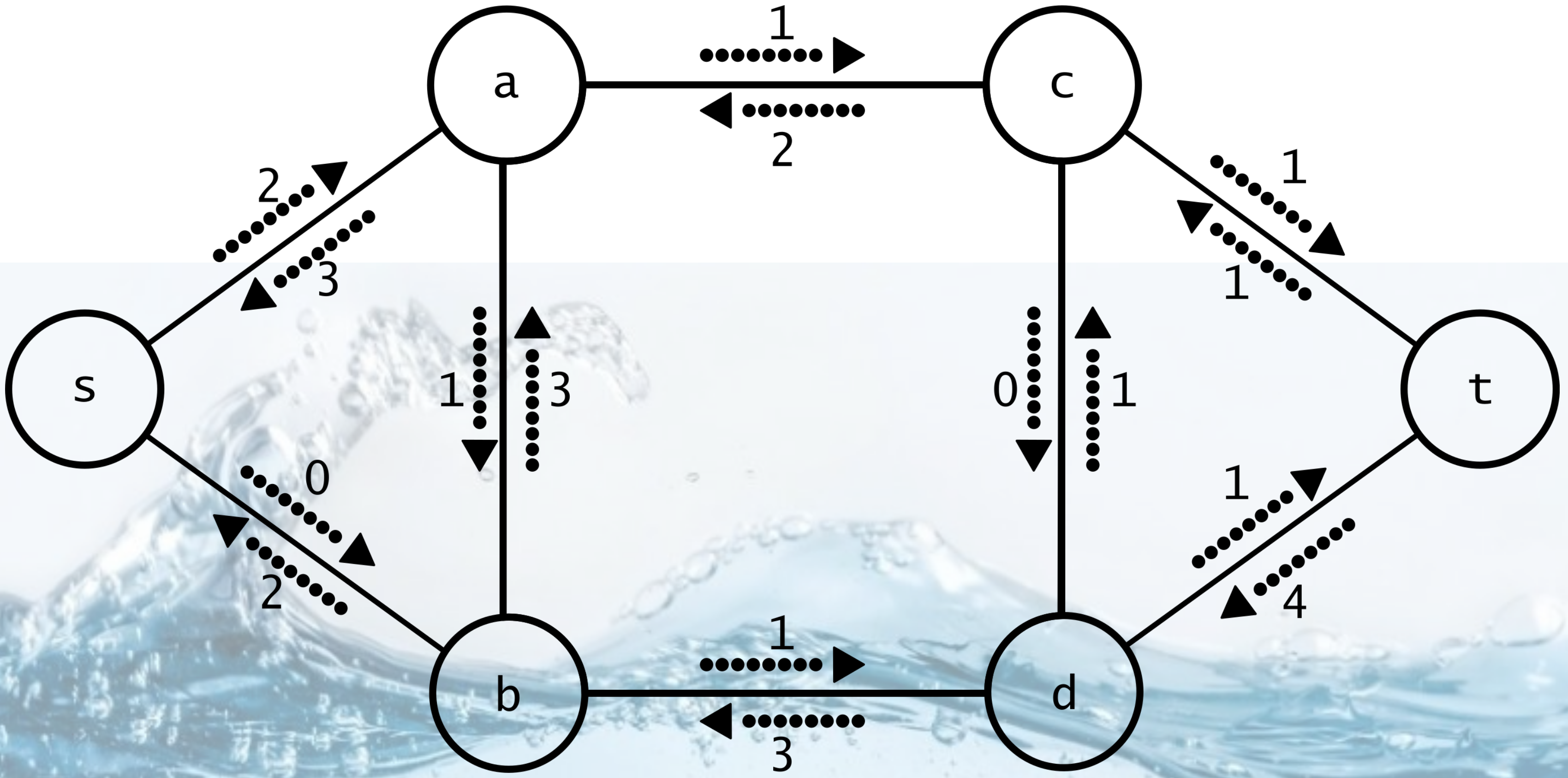
Softmax function is proposed as a differentiable generalization of the ***"winner-take-all"*** picking maximum operation.

$$\phi(Q)(\phi(K)^T V) + \textbf{Competition Mechanism} \Longleftrightarrow \mathbf{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

"fixed resource will cause competition"

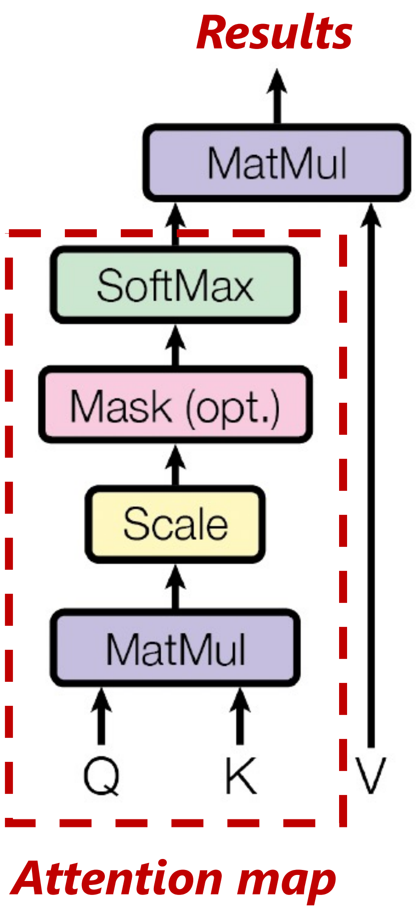
Bridle et al. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *NeurIPS 1989.*

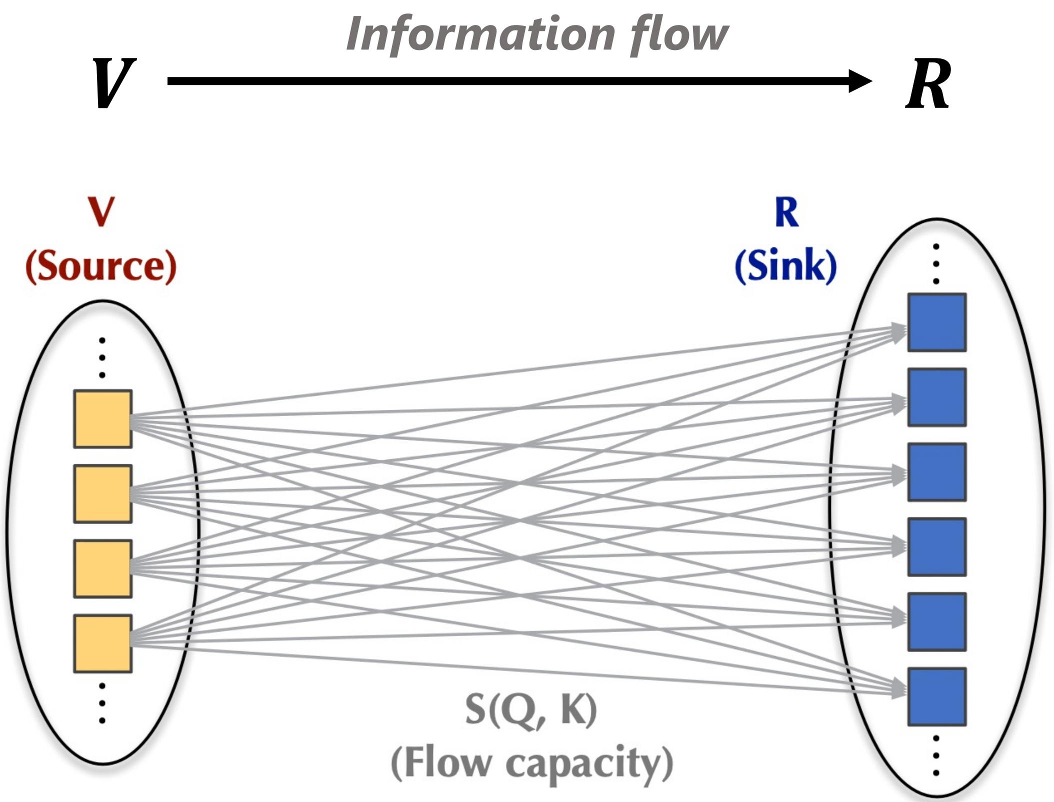
Flow Network Theory



[Conservation Property]: *The incoming flow capacity of each node is equal to the outgoing flow.*

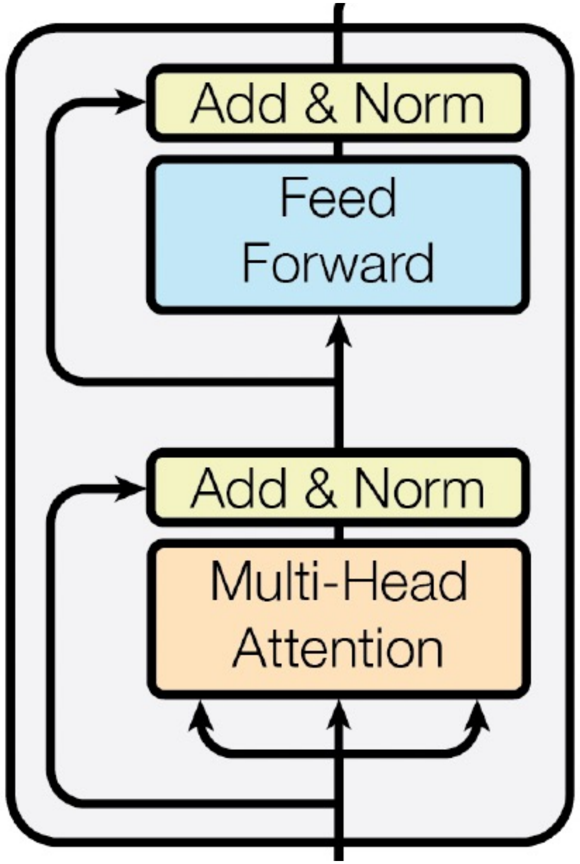
Attention: A Flow Network View

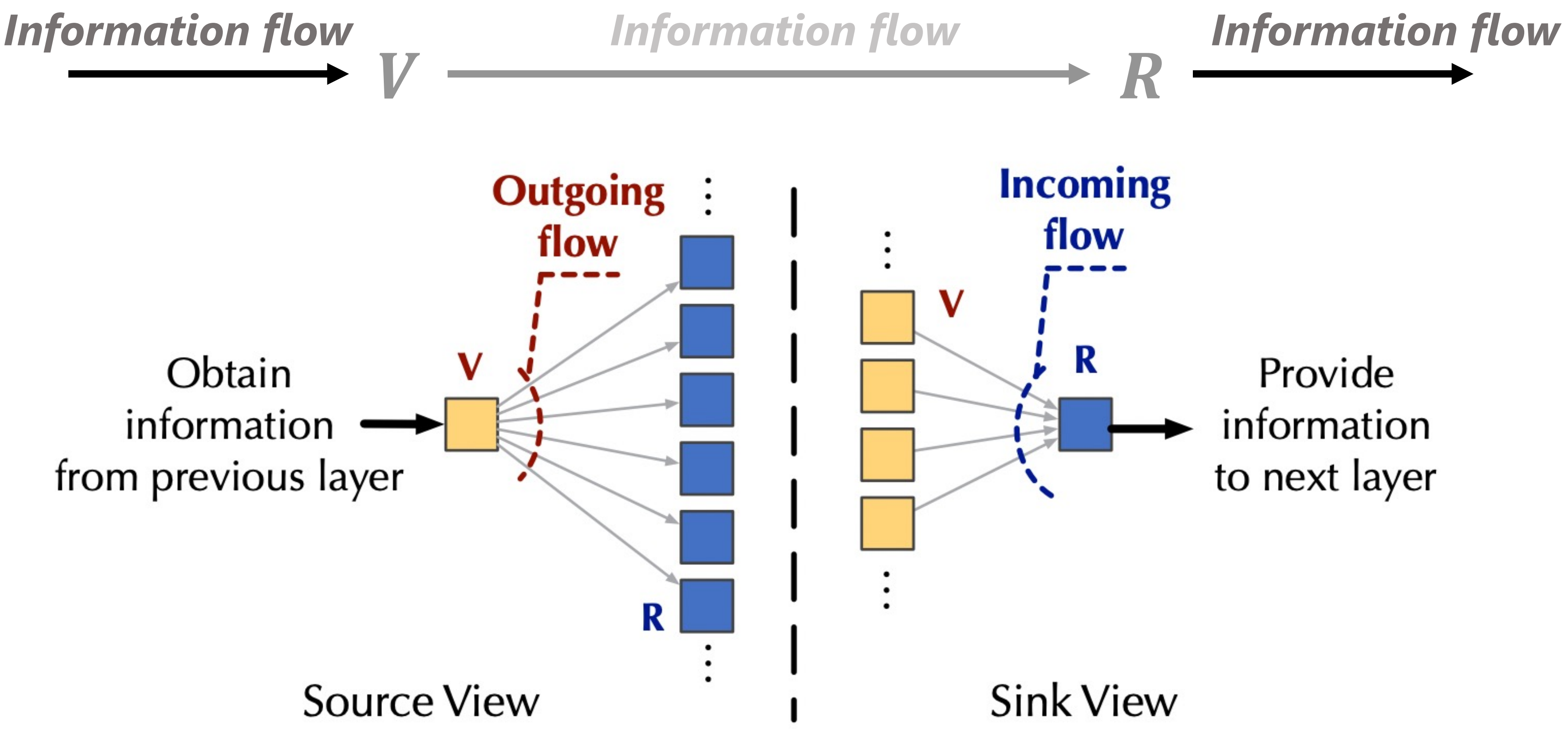




(a) Inner View

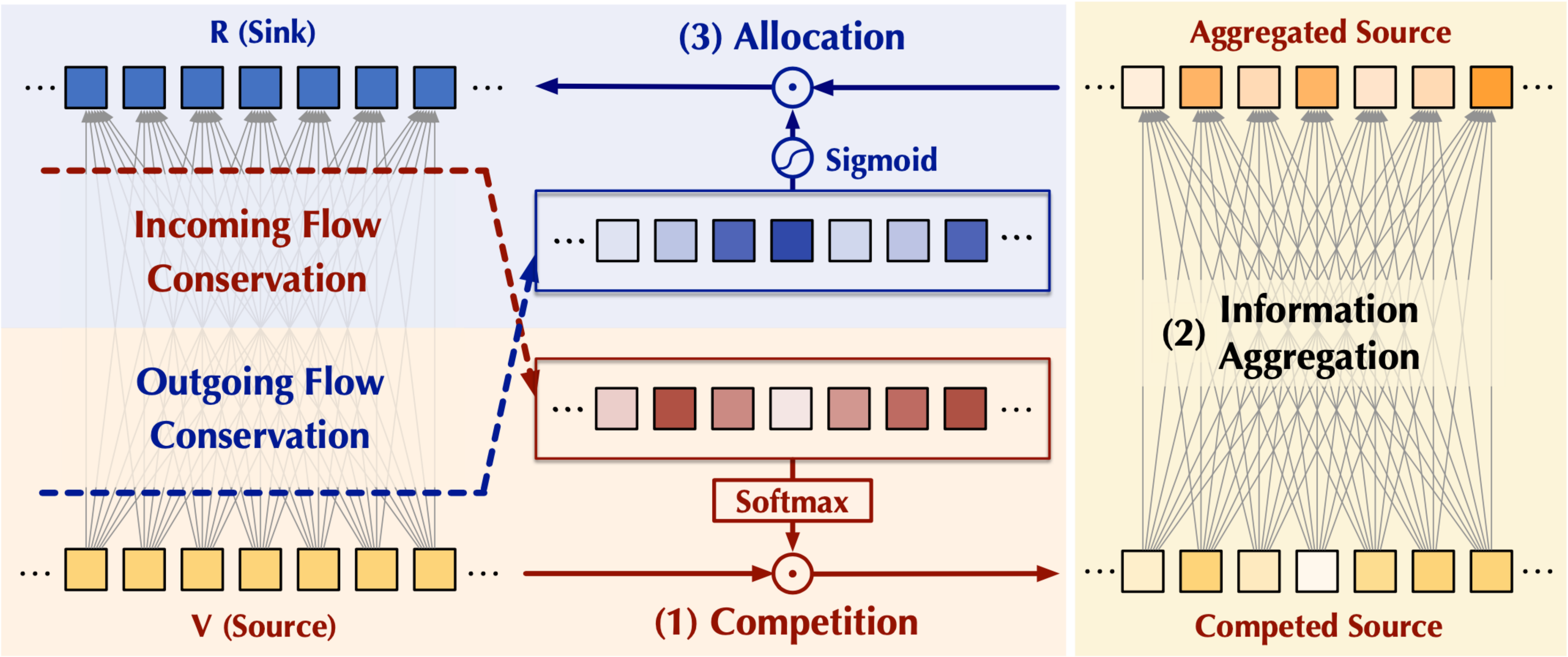
Attention: A Flow Network View





(b) Outer View

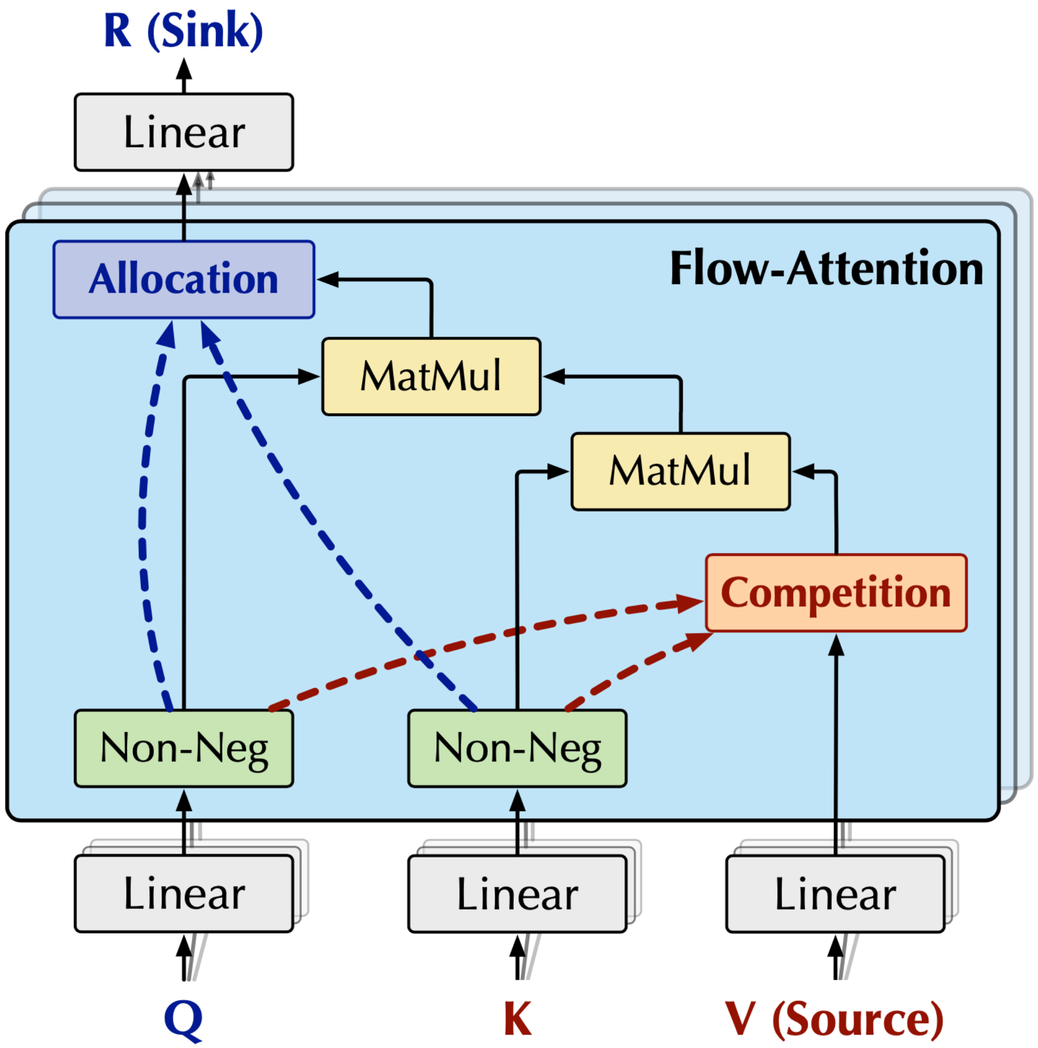
Conservation in Attention



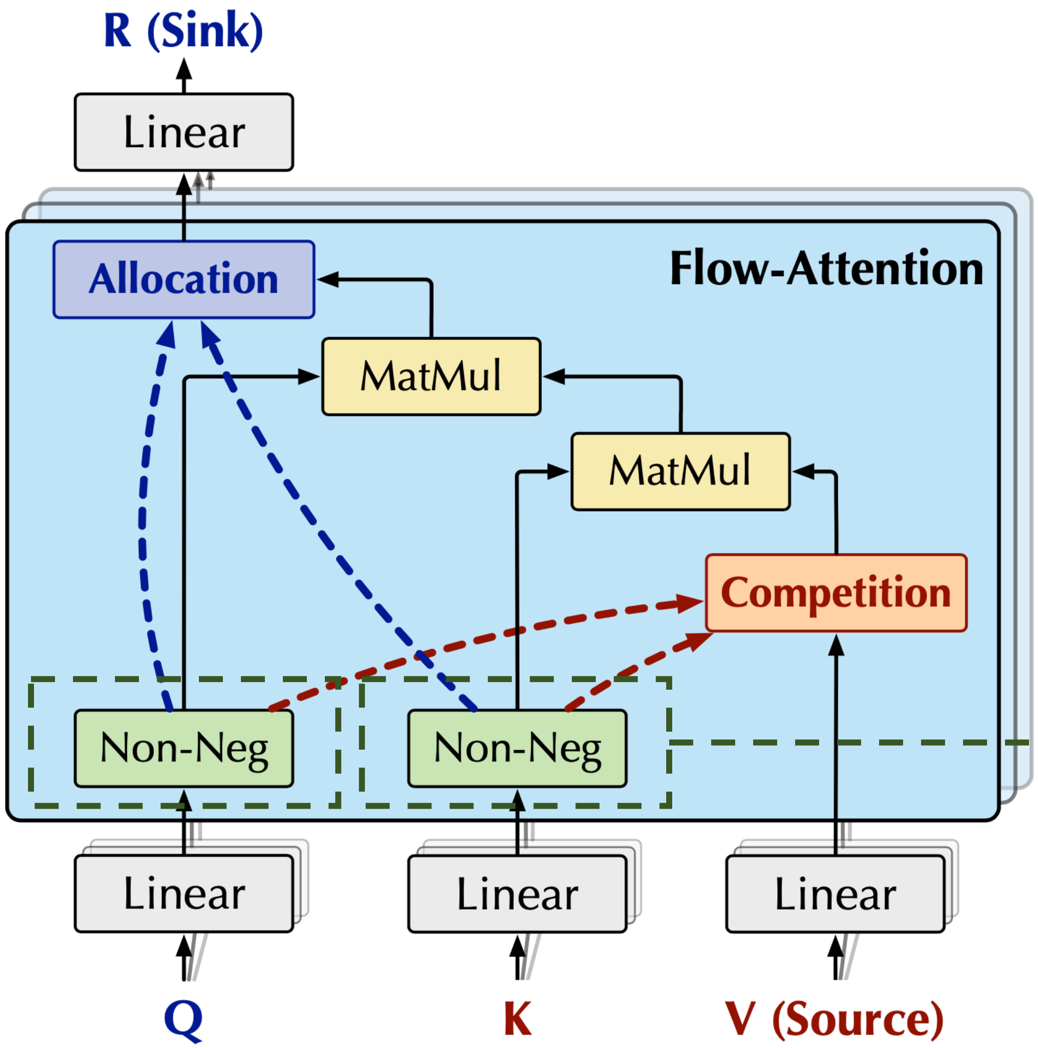
[Incoming Flow Conservation]: *Competition among Source tokens*

[Outgoing Flow Conservation]: *Competition among Sink tokens*

Flow-Attention

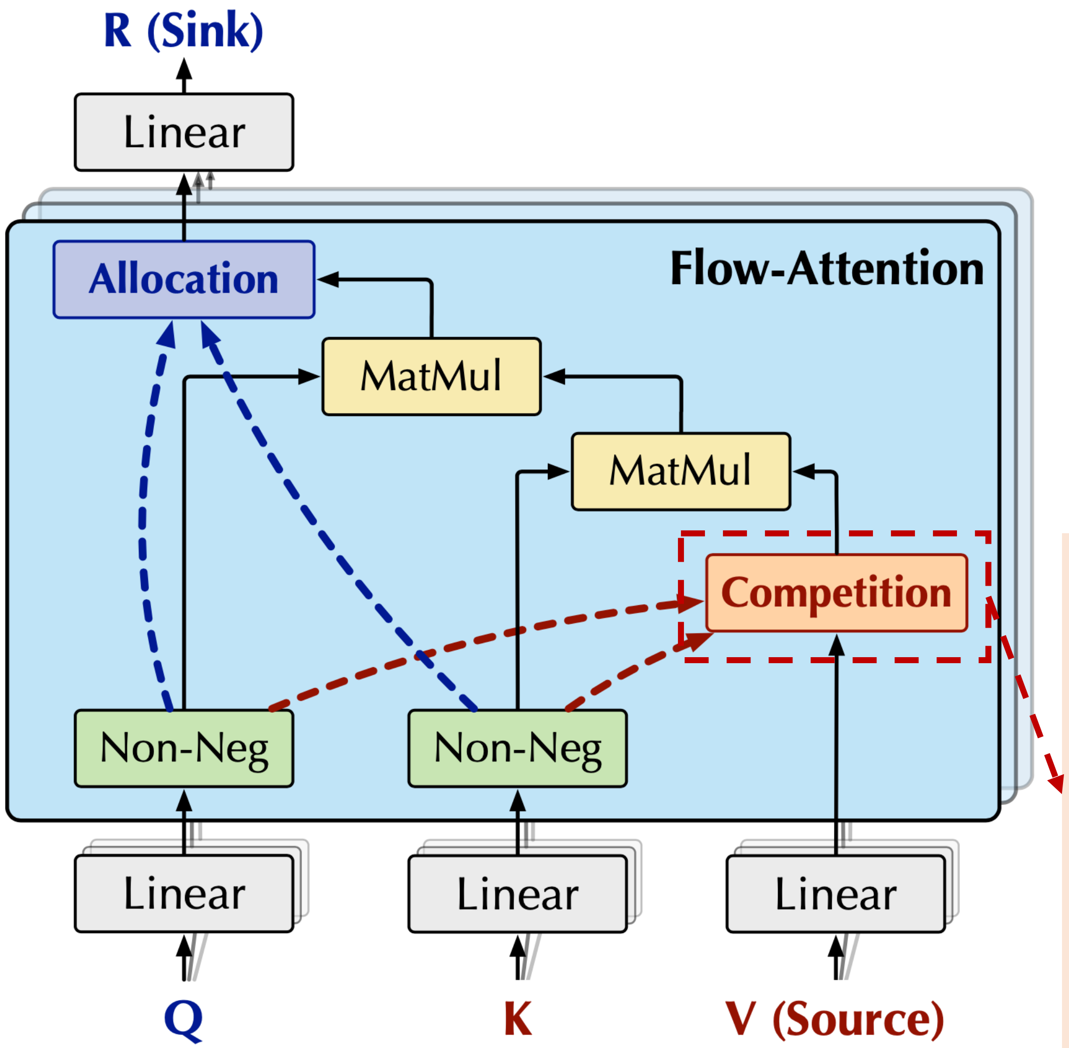


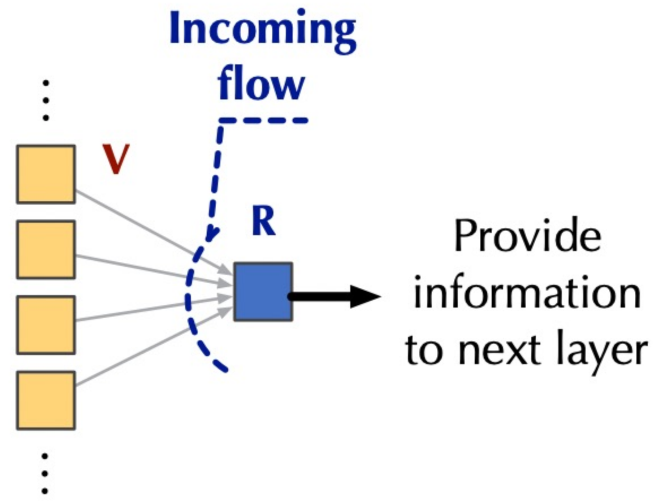
Flow-Attention



$$\phi(\cdot) = \mathrm{Sigmoid}(\cdot) \text{ or } \phi(\cdot) = \mathrm{ELU}(\cdot) + 1.0$$

Flow-Attention



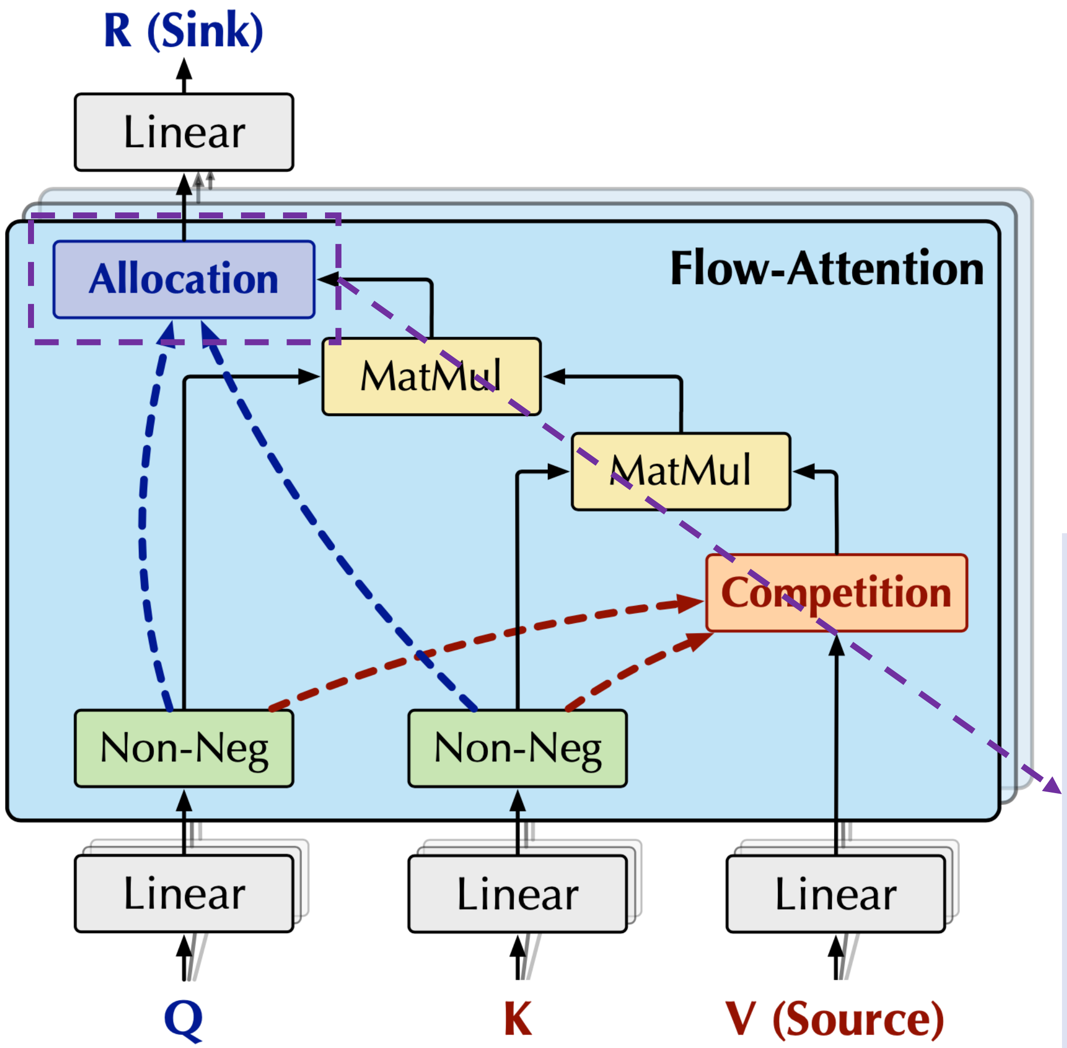


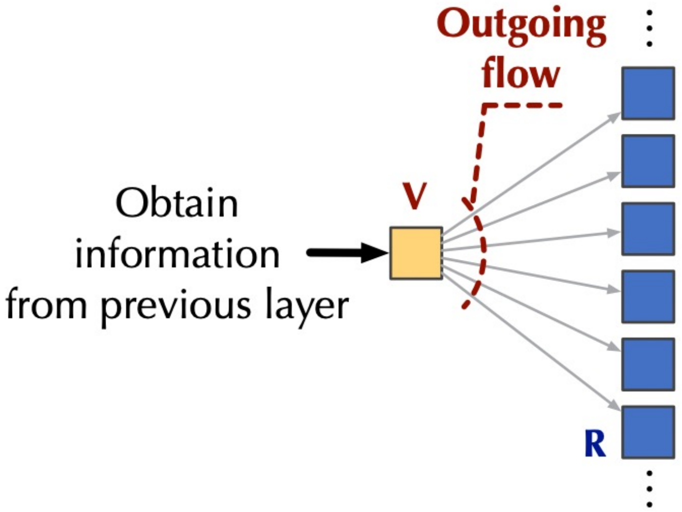
Incoming flow: $I_i = \phi(Q_i)\sum_j \phi(K_j)^T$

Incoming flow conservation: $\dfrac{\phi(\boldsymbol{Q})}{\boldsymbol{I}}$

Conserved outgoing flow: $\widehat{\boldsymbol{O}} = \phi(\boldsymbol{K})\sum_i \dfrac{\phi(Q_i)^T}{I_i}$

Flow-Attention



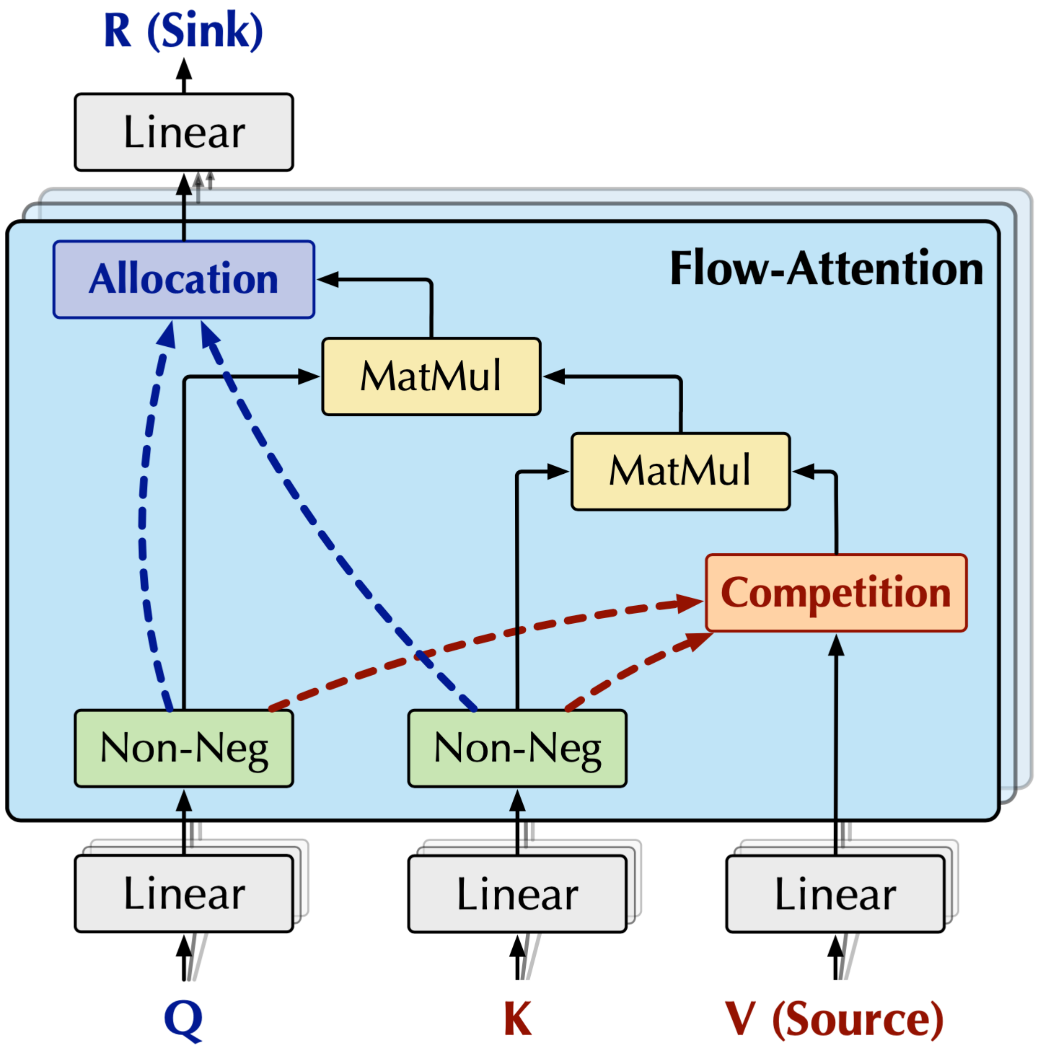


Outgoing flow: $O_i = \phi(K_i) \sum_j \phi(Q_j)^T$

Outgoing flow conservation: $\dfrac{\phi(\boldsymbol{K})}{\boldsymbol{O}}$

Conserved incoming flow: $\hat{\boldsymbol{I}} = \phi(\boldsymbol{Q}) \sum_j \dfrac{\phi(K_j)^T}{O_j}$

Flow-Attention

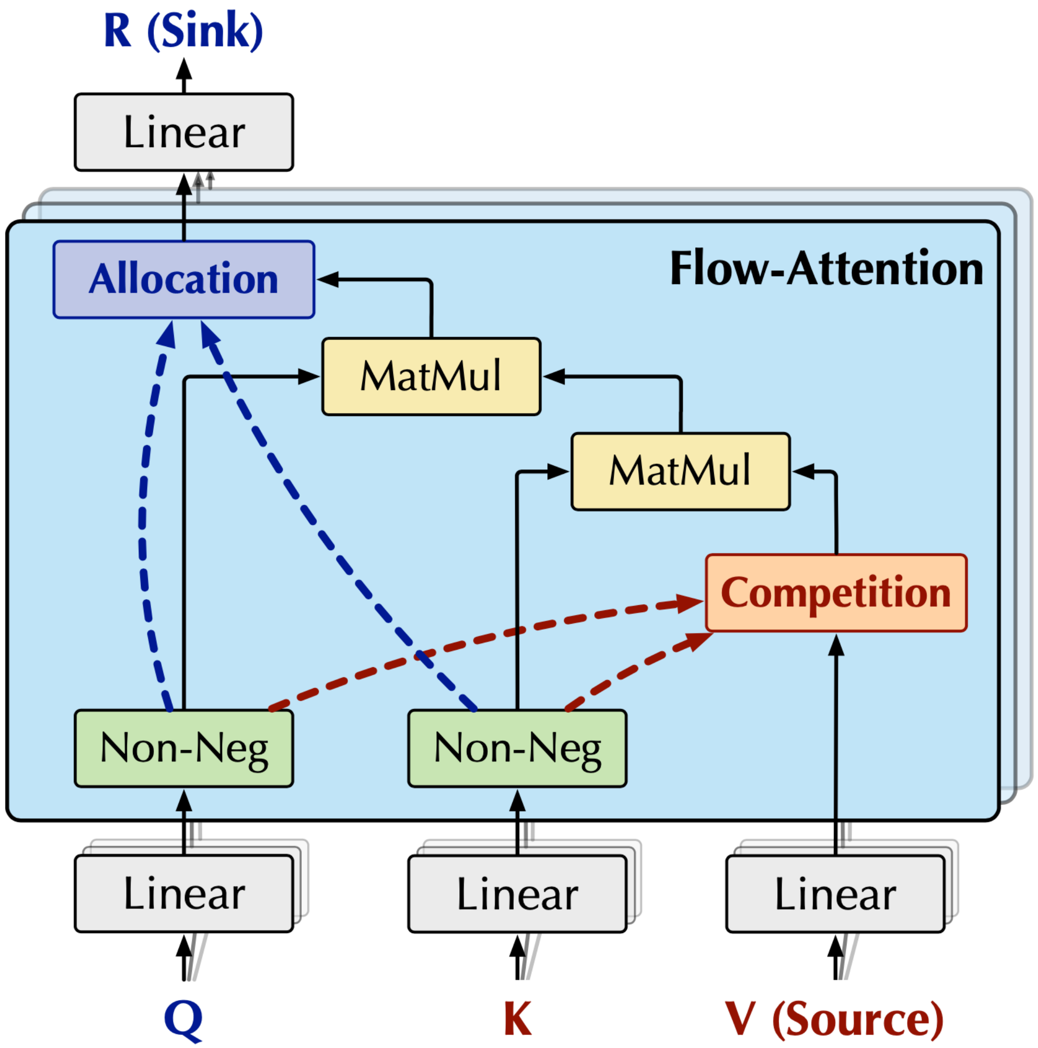


$$\text{Competition: } \widehat{\mathbf{V}} = \underline{\mathrm{Softmax}(\widehat{\mathbf{O}})} \odot \mathbf{V}$$

$$\text{Aggregation: } \mathbf{A} = \frac{\phi(\mathbf{Q})}{\mathbf{I}}(\phi(\mathbf{K})^{\mathsf{T}}\widehat{\mathbf{V}})$$

$$\text{Allocation: } \mathbf{R} = \underline{\mathrm{Sigmoid}(\widehat{\mathbf{I}})} \odot \mathbf{A},$$

Flow-Attention



$$\text{Competition: } \widehat{\mathbf{V}} = \text{Softmax}(\widehat{\mathbf{O}}) \odot \mathbf{V}$$

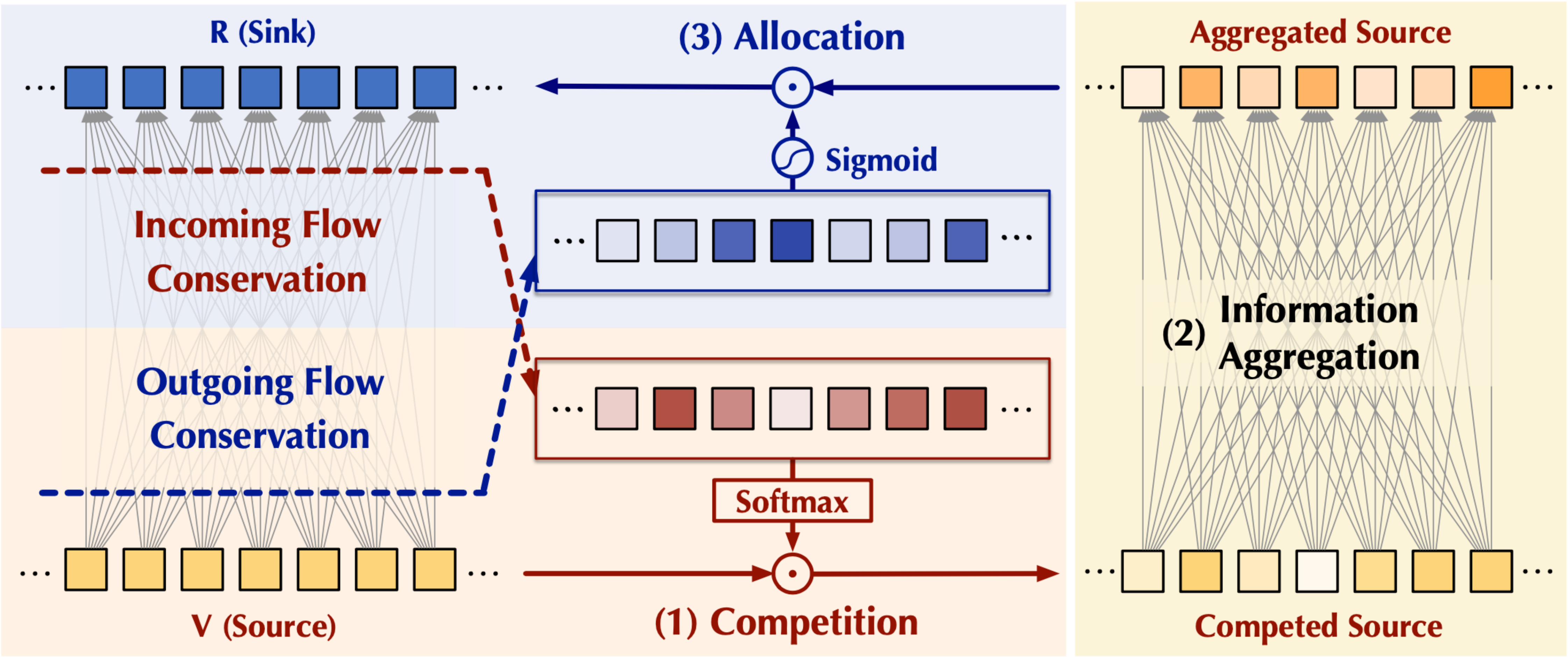
$$\text{Aggregation: } \mathbf{A} = \frac{\phi(\mathbf{Q})}{\mathbf{I}}(\phi(\mathbf{K})^{\mathsf{T}}\widehat{\mathbf{V}})$$

$$\text{Allocation: } \mathbf{R} = \text{Sigmoid}(\widehat{\mathbf{I}}) \odot \mathbf{A},$$

Successfully bring the *Competition Mechanism*

Into Attention design to avoid trivial attention

Efficiency and Universality



[Efficiency]: *All the calculations are* ***in linear complexity.***

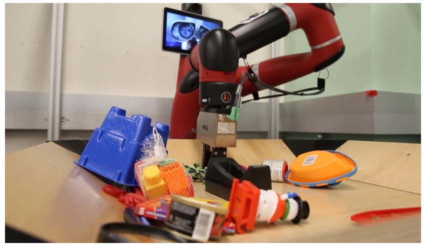
[Universality]: *The whole design is based on flow network* ***without specific inductive biases.***

Flowformer Experiments

Image

Language

Time Series

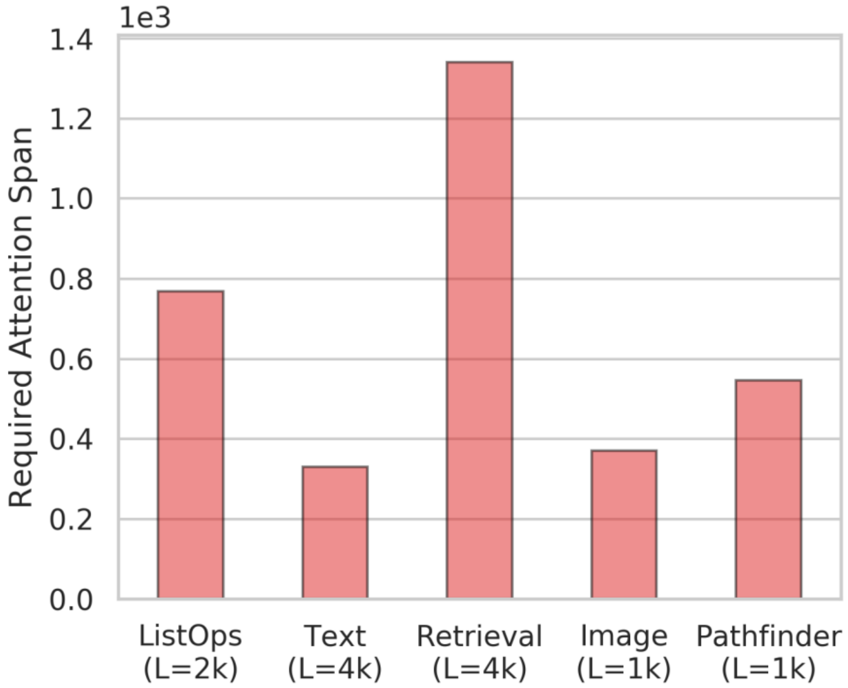


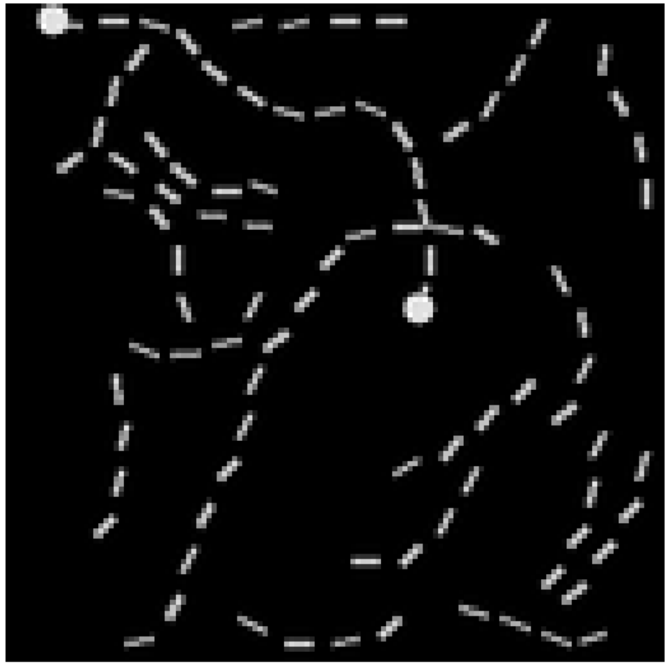
Agent Trajectory

BENCHMARKS	TASK	VERSION	LENGTH
LRA (2020c)	SEQUENCE	NORMAL	1000∼4000
WIKITEXT (2017)	LANGUAGE	CAUSAL	512
IMAGENET (2009)	VISION	NORMAL	49∼3136
UEA (2018)	TIME SERIES	NORMAL	29∼1751
D4RL (2020)	OFFLINE RL	CAUSAL	60

- Extensive tasks (covering 5 mainstream tasks)
- Normal and causal versions
- Various sequence lengths (29-4000)
- Extensive baselines (20+)

Long Sequence Modeling on LRA





Long sequence dataset with interesting tasks

Example 1: Pathfinder

INPUT: [MAX 4 3 [MIN 2 3] 1 0 [MEDIAN 1 5 8 9, 2]] **OUTPUT:** 5

Example 2: Listops

https://github.com/google-research/long-range-arena

Long Sequence Modeling on LRA

Model	ListOps ↑	Text ↑	Retrieval ↑	Image ↑	Pathfinder ↑	Avg ↑
Local Attention (Tay et al., 2021)	15.82	52.98	53.39	41.46	66.63	46.06
Linear Trans. (Katharopoulos et al., 2020)	16.13	**65.90**	53.09	42.34	75.30	50.55
Reformer (Kitaev et al., 2020)	37.27	56.10	53.40	38.07	68.50	50.67
Sparse Trans. (Child et al., 2019)	17.07	63.58	59.59	**44.24**	71.71	51.24
Sinkhorn Trans. (Tay et al., 2020b)	33.67	61.20	53.83	41.23	67.45	51.29
Linformer (Wang et al., 2020)	35.70	53.94	52.27	38.56	76.34	51.36
Performer (Choromanski et al., 2021)	18.01	65.40	53.82	42.77	**77.05**	51.41
Synthesizer (Tay et al., 2020a)	36.99	61.68	54.67	41.61	69.45	52.88
Longformer (Beltagy et al., 2020)	35.63	62.85	56.89	42.22	69.71	53.46
Transformer (Vaswani et al., 2017)	36.37	64.27	57.46	42.44	71.40	54.39
BigBird (Zaheer et al., 2020)	36.05	64.02	59.29	40.83	74.87	55.01
cosFormer (Zhen et al., 2022)	37.90	63.41	61.36	43.17	70.33	55.23
Flowformer w/o Competition	36.80	63.48	61.66	42.39	71.90	55.25
Flowformer w/o Allocation	37.00	63.78	61.33	42.52	73.26	55.58
Flowformer	**38.70**	64.29	**62.24**	43.20	73.95	**56.48**

State-of-the-art performance (Avg Acc 56.48%)

Ablation study to verify model effectiveness

Long Sequence Modeling on LRA

Model Speed	Inference (Steps per second)				Train (Steps per second)			
Sequence length	1K	2K	3K	4K	1K	2K	3K	4K
Transformer (Vaswani et al., 2017)	81.83	25.26	-	-	22.12	7.50	-	-
Local Attention (Tay et al., 2021)	98.28	96.51	94.60	95.60	46.75	43.05	35.42	30.34
Linear Trans. (Katharopoulos et al., 2020)	97.33	96.14	94.03	93.69	<u>48.66</u>	**48.78**	<u>41.66</u>	35.44
Reformer (Kitaev et al., 2020)	60.92	60.30	39.37	26.98	46.07	22.93	14.34	9.56
Sparse Trans. (Child et al., 2019)	78.30	23.33	-	-	21.74	7.30	-	-
Sinkhorn Trans. (Tay et al., 2020b)	91.42	92.21	92.72	80.67	45.93	36.21	28.11	23.83
Linformer (Wang et al., 2020)	96.56	**96.84**	94.74	93.59	45.57	44.11	37.28	31.58
Performer (Choromanski et al., 2021)	**99.60**	96.80	**96.52**	**96.42**	47.34	48.30	41.00	36.14
Synthesizer (Tay et al., 2020a)	65.44	-	-	-	5.16	-	-	-
Longformer (Beltagy et al., 2020)	73.56	-	-	-	13.09	-	-	-
BigBird (Zaheer et al., 2020)	82.50	54.12	37.83	29.34	27.34	16.95	12.00	9.33
cosFormer (Zhen et al., 2022)	96.46	95.58	95.19	94.69	46.50	45.24	39.49	35.09
Flowformer	<u>98.83</u>	96.21	<u>95.65</u>	<u>95.82</u>	**49.76**	47.18	**41.93**	**36.79**

High efficiency (comparable to Performer)

State-of-the-art performance (56.48 v.s. 51.41)

Language Modeling on Wikitext-103

Model	Perplexity ↓
Transformer (2017)	33.0
Linear Trans. (2020)	38.4
Reformer (2020)	33.6
Performer (2021)	37.5
TRF-Transformer (2021)	33.6
TRF-Transformer-GATE (2021)	31.3
cosFormer (2022)	34.1
Flowformer w/o Competition	31.2
Flowformer w/o Allocation	32.2
Flowformer	**30.8**

Strong performance in **causal task**

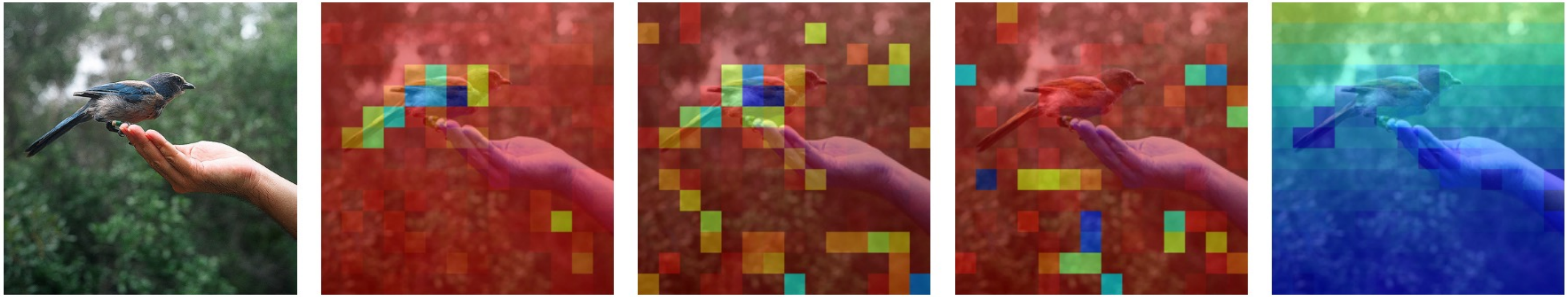
Vision Recognition on ImageNet-1K

Model	Complex.	Params (MB)	Flops (G)	Top-1 Acc.	Top-5 Acc.
ViT-base (2021)	$\mathcal{O}(n^2d)$	86	55.4	77.9	/
ViT-large (2021)	$\mathcal{O}(n^2d)$	307	190.7	76.5	/
Full Attn. (2017)	$\mathcal{O}(n^2d)$	41	6.7	78.7	94.3
Linear Trans. (2020)	$\mathcal{O}(nd^2)$	41	6.3	79.0	94.1
Reformer (2020)	$\mathcal{O}((n\log n)d)$	37	6.0	79.6	94.7
Longformer (2020)	$\mathcal{O}(nd^2)$	38	6.3	77.6	93.1
Performer (2021)	$\mathcal{O}(nd^2)$	41	6.3	78.1	93.2
Nyströmformer (2021)	$\mathcal{O}(nd^2)$	41	6.3	77.2	93.0
YOSO-E (2021)	$\mathcal{O}(nd^2)$	41	5.8	79.0	94.3
SOFT (2021)	$\mathcal{O}(nd^2)$	37	5.8	79.2	94.5
cosFormer (2022)	$\mathcal{O}(nd^2)$	41	6.3	68.3	88.0
Flowformer	$\mathcal{O}(nd^2)$	41	6.3	**80.6**	**94.9**
DeiT-S (2021)	$\mathcal{O}(n^2d)$	22	4.6	79.8	**95.0**
DeiT+Flowformer	$\mathcal{O}(nd^2)$	22	**4.2**	**80.0**	94.8

Surpass previous attentions

Speed up well-designed Transformers

Vision Recognition on ImageNet-1K



Input Frame (Bird) | **Flowformer (Ours)** | Canonical Transformer | Linear Transformer | cosFormer

Input Frame (Birdhouse) | **Flowformer (Ours)** | Canonical Transformer | Linear Transformer | cosFormer

Flowformer can naturally visualize the attention map $\mathrm{Softmax}(\widehat{\boldsymbol{O}})$

Vision Recognition on ImageNet-1K

Class: Airliner　　Class: Bird　　Class: Birdhouse

Visualization of the allocation weights $\mathrm{Sigmoid}(\hat{\boldsymbol{I}})$

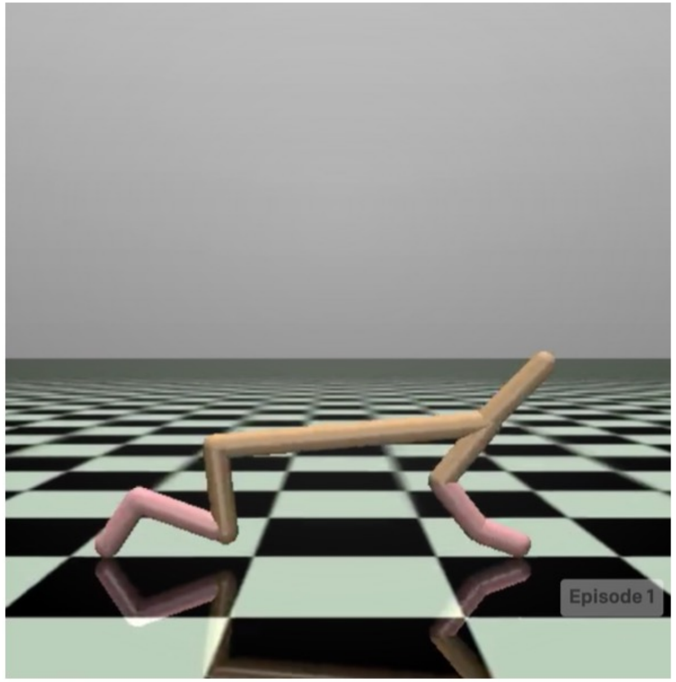
Time Series Classification on UEA

Dataset / Model	Classical methods			Deep Models										
				RNN	TCN	Transformer and its Efficient Variants								
	DTW (1994)	XGBoost (2016)	Rocket (2020)	LSTM (1997)	Unsuper. (2019)	Trans. (2017)	Linear. (2020)	Re. (2020)	Long. (2020)	Per. (2021)	YOSO-E (2021)	SOFT (2021)	cos. (2022)	Flow. (ours)
EthanolConcentration	32.3	43.7	45.2	32.3	28.9	32.7	31.9	31.9	32.3	31.2	31.2	32.3	33.5	33.8
FaceDetection	52.9	63.3	64.7	57.7	52.8	67.3	67.0	68.6	62.6	67.0	67.3	64.8	67.1	67.6
Handwriting	28.6	15.8	58.8	15.2	53.3	32.0	34.7	27.4	39.6	32.1	30.9	28.9	34.7	33.8
Heartbeat	71.7	73.2	75.6	72.2	75.6	76.1	76.6	77.1	78.0	75.6	76.5	77.1	75.6	77.6
JapaneseVowels	94.9	86.5	96.2	79.7	98.9	98.7	99.2	97.8	98.9	98.1	98.6	98.3	99.2	98.9
PEMS-SF	71.1	98.3	75.1	39.9	68.8	82.1	82.1	82.7	83.8	80.9	85.2	83.2	80.9	83.8
SelfRegulationSCP1	77.7	84.6	90.8	68.9	84.6	92.2	92.5	90.4	90.1	91.5	91.1	91.1	91.8	92.5
SelfRegulationSCP2	53.9	48.9	53.3	46.6	55.6	53.9	56.7	56.7	55.6	56.7	53.9	55.0	55.6	56.1
SpokenArabicDigits	96.3	69.6	71.2	31.9	95.6	98.4	98.0	97.0	94.4	98.4	98.9	98.4	98.8	98.8
UWaveGestureLibrary	90.3	75.9	94.4	41.2	88.4	85.6	85.0	85.6	87.5	85.3	88.4	85.6	85.0	86.6
Average Accuracy	67.0	66.0	72.5	48.6	70.3	71.9	72.4	71.5	72.0	71.9	72.2	71.5	72.2	**73.0**

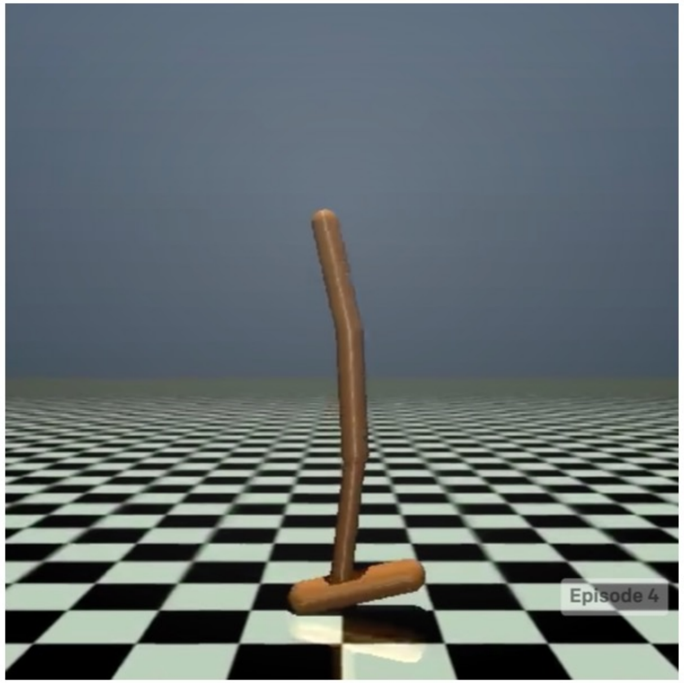
Extensive **data types** and **comparing baselines**

The **only** deep model surpasses Rocket.

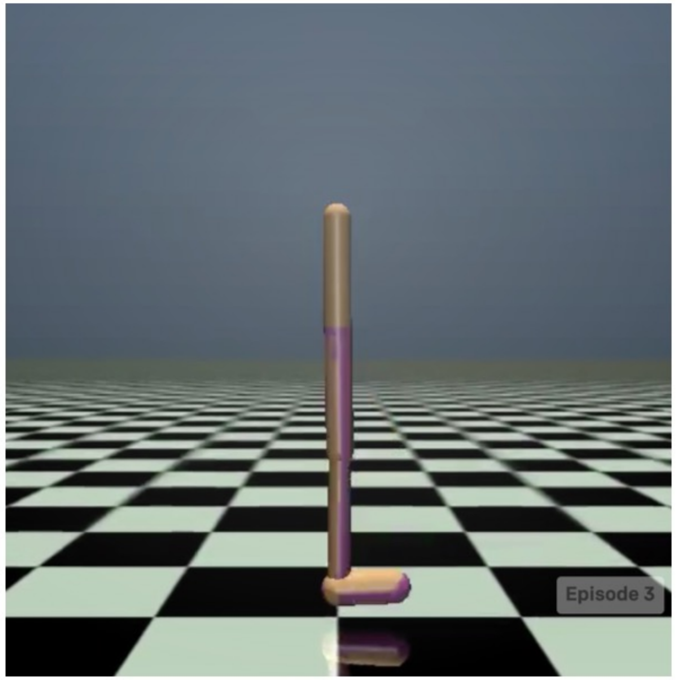
Offline Reinforcement Learning on D4RL



HalfCheetah



Hopper



Walker

Complex **offline control** task in the **autoregressive** protocol.

… [S] [A] [R] …

Offline Reinforcement Learning on D4RL

Environment	BC (1989)	AWAC (2020)	DT (2021a)	Linear Trans. (2020)	Reformer (2020)	Performer (2021)	cosFormer (2022)	Flowformer (Ours)
Medium-Expert								
HalfCheetah	55.2	42.8	83.8±3.3	78.2±3.2	81.5±1.6	85.1±2.1	85.5±2.9	90.8±0.4
Hopper	52.5	55.8	104.0±2.5	107.2±0.9	104.2±9.8	93.5±13.9	98.1±7.4	109.9±1.0
Walker	107.5	74.5	107.7±0.6	67.2±27.3	71.4±1.8	72.6±2.4	100.5±14.5	108.0±0.4
Medium								
HalfCheetah	42.6	43.5	42.4±0.1	42.3±0.2	42.2±0.1	42.1±0.2	42.1±0.3	42.2±0.2
Hopper	52.9	57.0	64.2±1.1	58.7±0.4	59.9±0.7	59.7±7.5	59.8±3.8	66.9±2.5
Walker	75.3	72.4	70.6±3.2	57.9±10.6	65.8±4.9	63.3±10.7	71.4±1.2	71.7±2.5
Medium-Replay								
HalfCheetah	36.6	40.5	34.6±0.6	32.1±1.5	33.6±0.7	31.7±0.9	32.8±3.6	34.7±1.5
Hopper	18.1	37.2	79.7±7.4	74.3±7.0	66.1±2.6	64.6±24.2	59.3±16.5	75.5±14.5
Walker	26.0	27.0	62.9±5.0	62.1±7.4	50.1±3.5	61.3±6.7	60.5±9.9	62.0±3.1
Avg Reward	51.9	50.1	72.2±**2.6**	64.4±6.5	63.9±2.9	63.8±7.6	67.8±7.6	**73.5**±2.9

Competitive performance in the comparison with *Decision Transformer*.

Chen et al. Decision Transformer: Reinforcement Learning via Sequence Modeling. *NeurIPS 2021.*

Summary

General Relation Modeling

Quadratic Complexity

Long Sequence
Model Efficiency
Big Model ☹

Task & Data Universal

Foundation Model

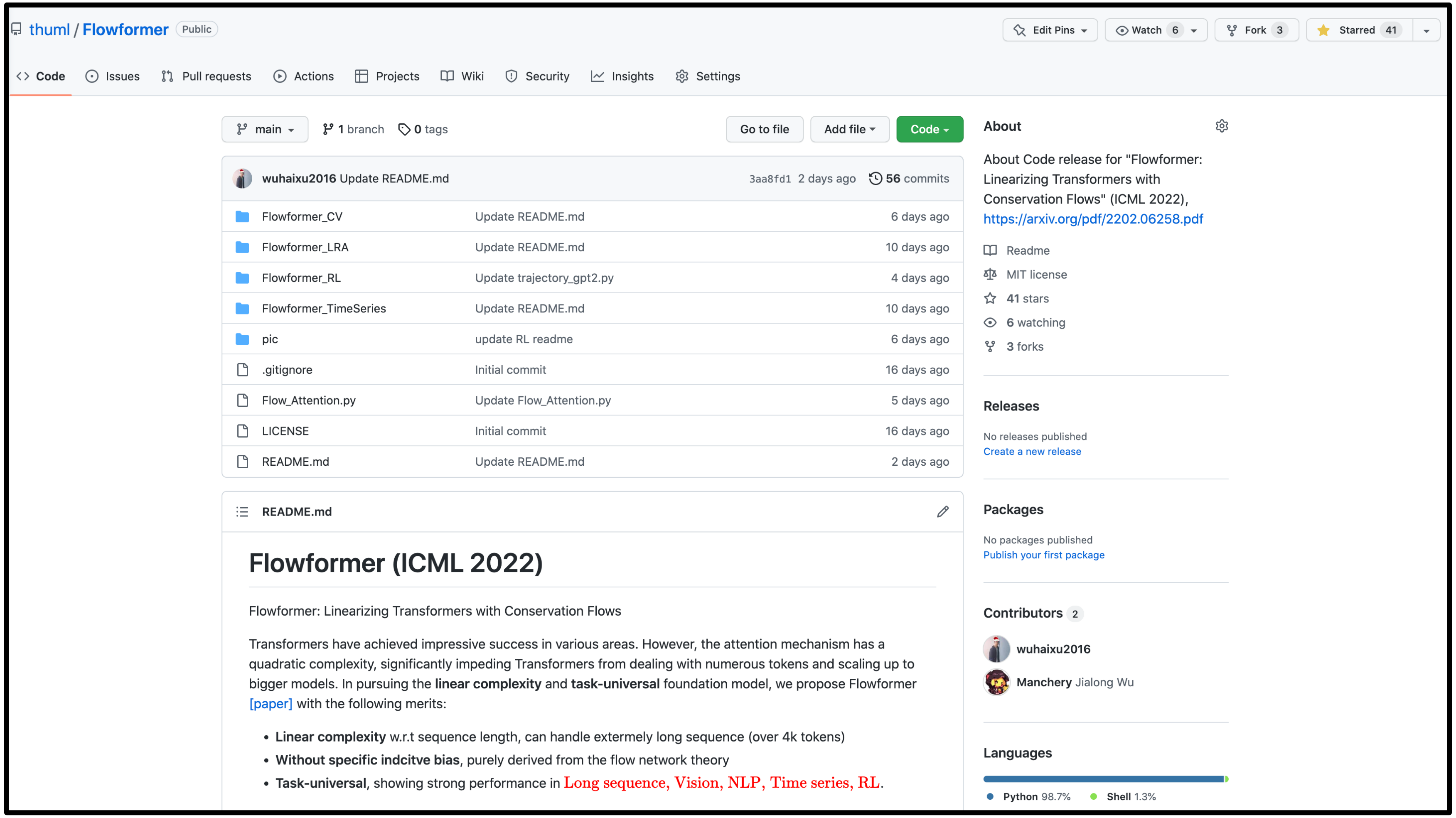
Flowformer

Linear complexity w.r.t. sequence length

Based on flow network & without specific inductive biases

Strong performance in Long Sequence, CV, NLP, Time Series, RL

Open Source



https://github.com/thuml/Flowformer

Complete benchmarks & datasets & scripts

Thank You!

whx20@mails.tsinghua.edu.cn