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

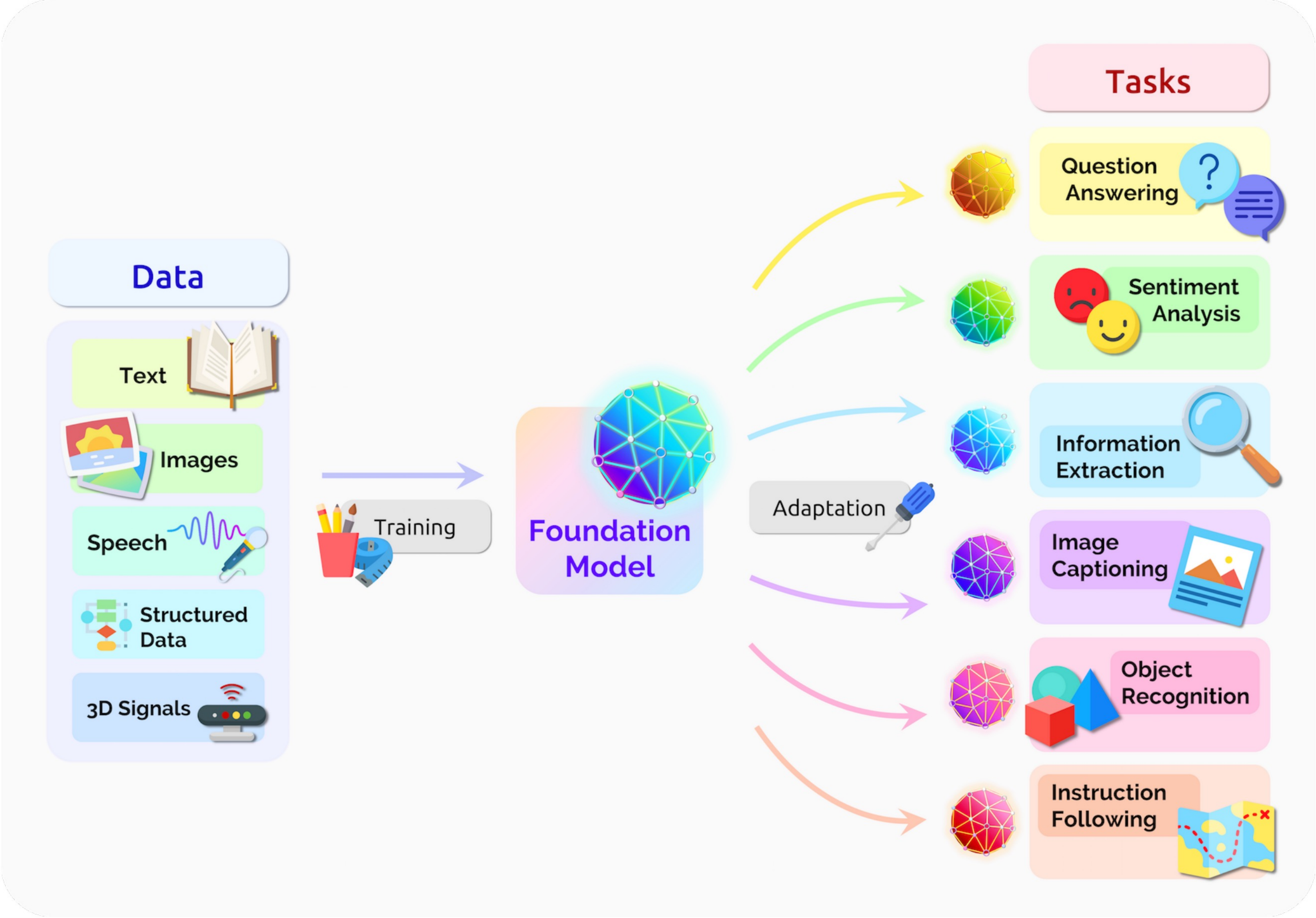
Jialong Wu

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

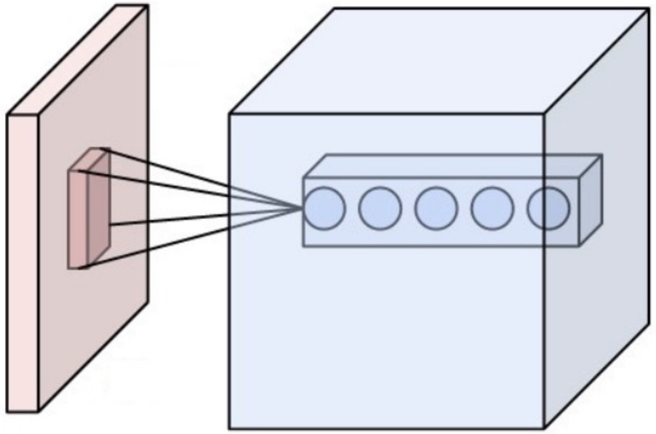
Image

Language

Time Series

Agent Trajectory

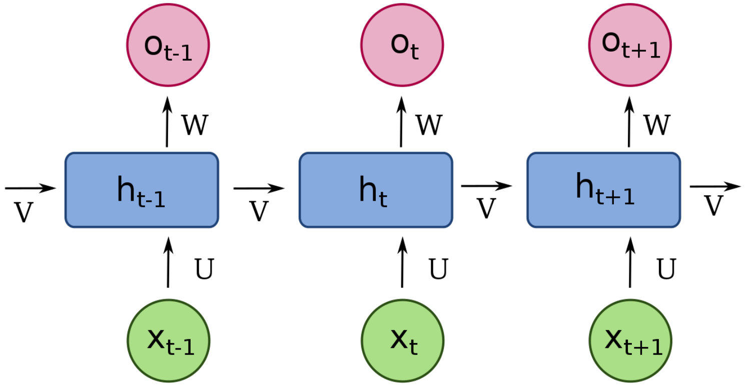
Universal Architecture



CNNs?

Locality

Shift Invariance ☹



RNNs?

Markov ☹

A Universal Architecture for General Proposes

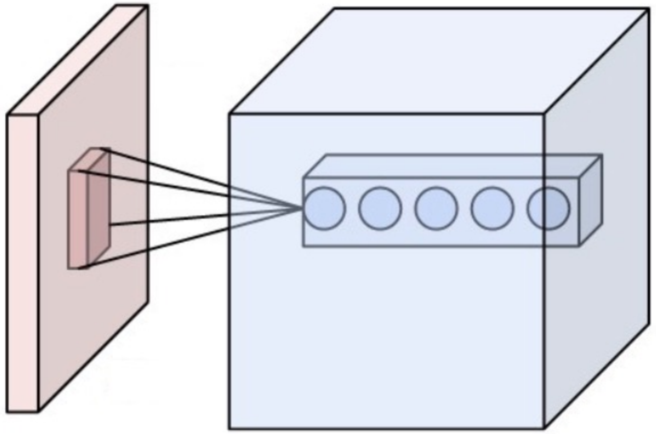
Image

Language

Time Series

Agent Trajectory

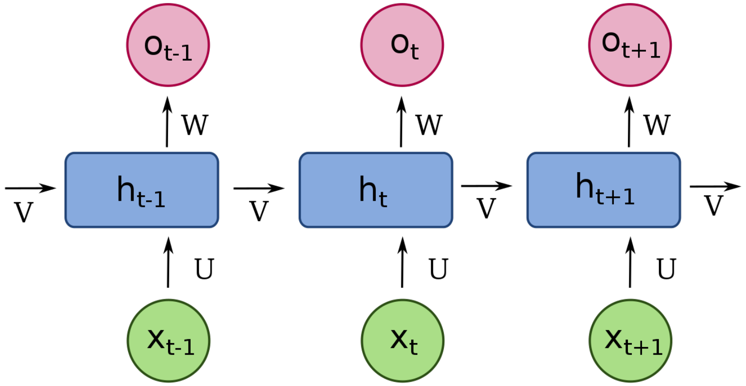
Universal Architecture



CNNs?

Locality

Shift Invariance ☹

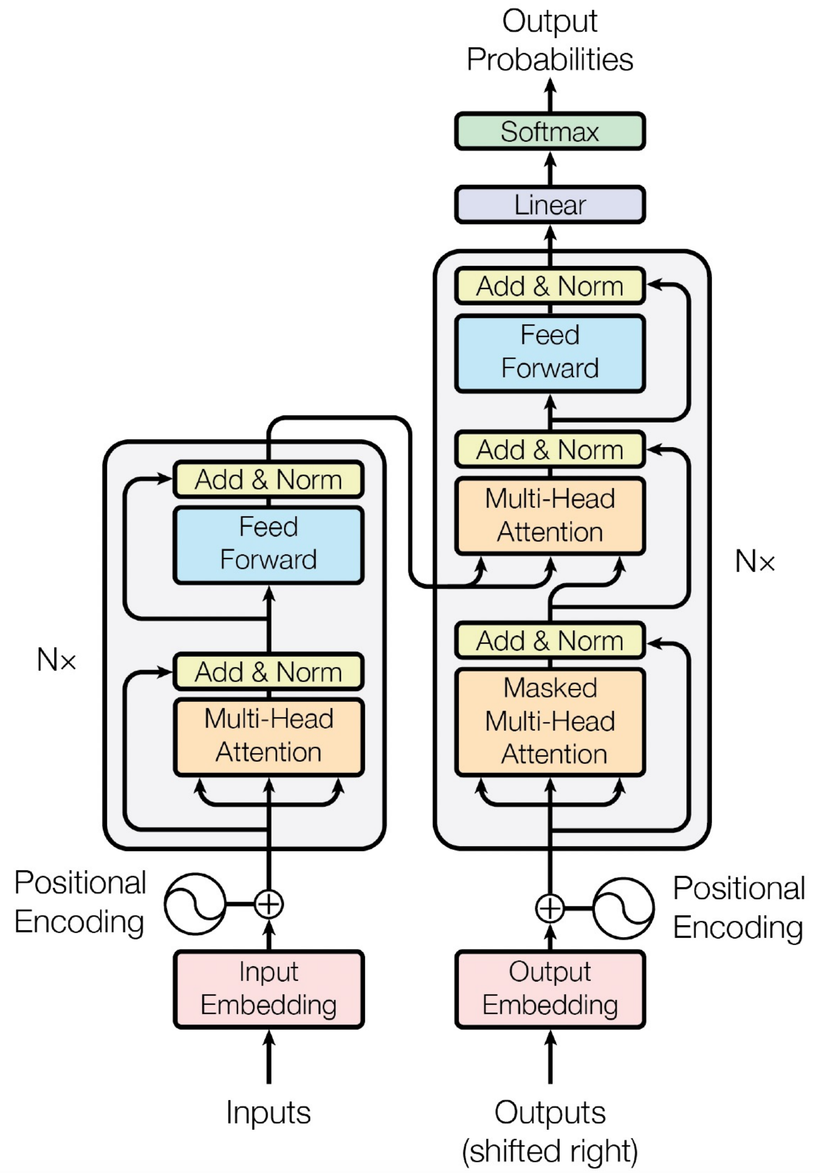


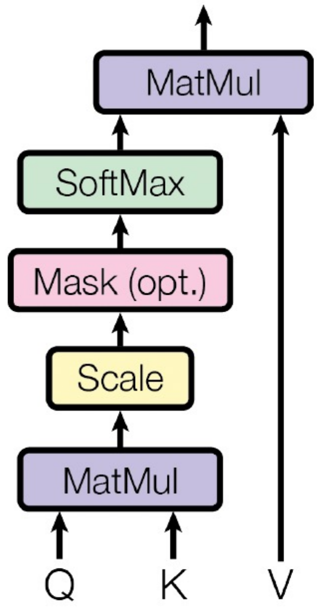
RNNs?

Markov ☹

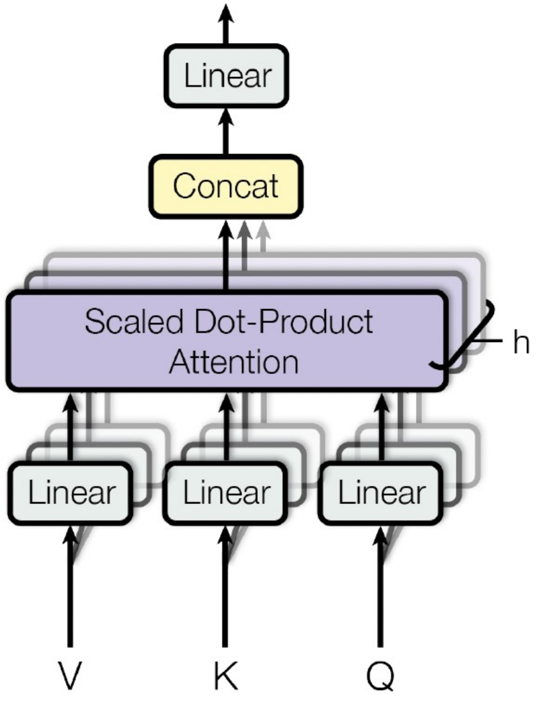
Specific Inductive Biases Limit the Model Universality

Transformers





Self-Attention



Multi-head **Self-Attention**

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

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**

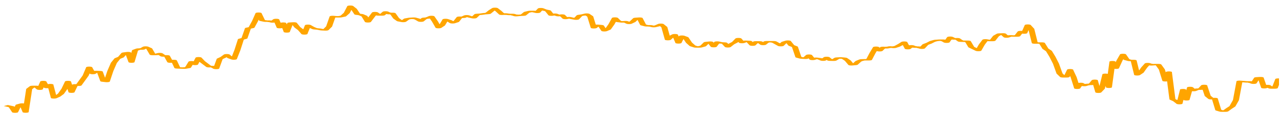
Language

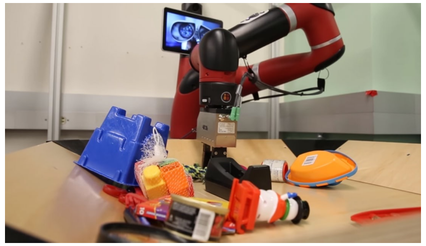
Relation among **Words**

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

Time Series

Relation among **Time Points**



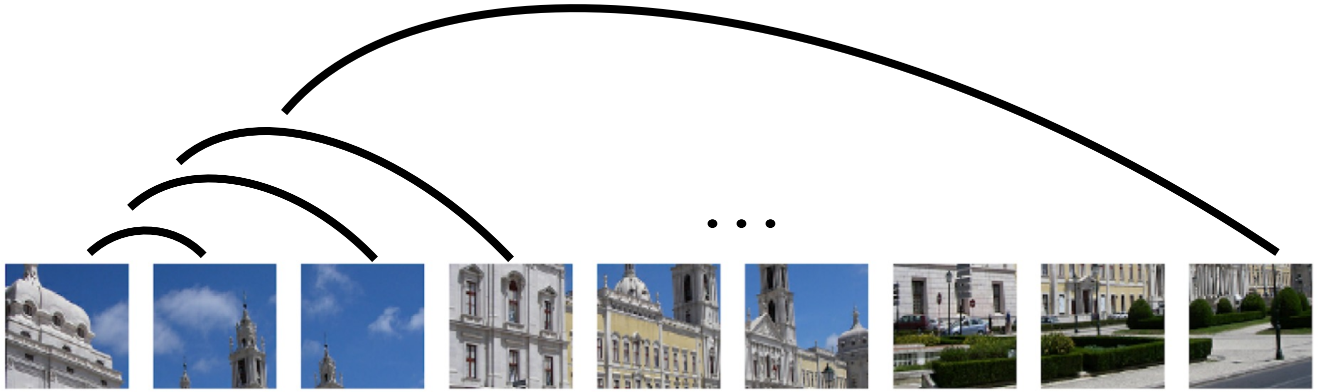


Agent Trajectory

Relation among **Agent-Environment Interactions**

… [S] [A] [R] …

Quadratic Complexity in Self-Attention



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

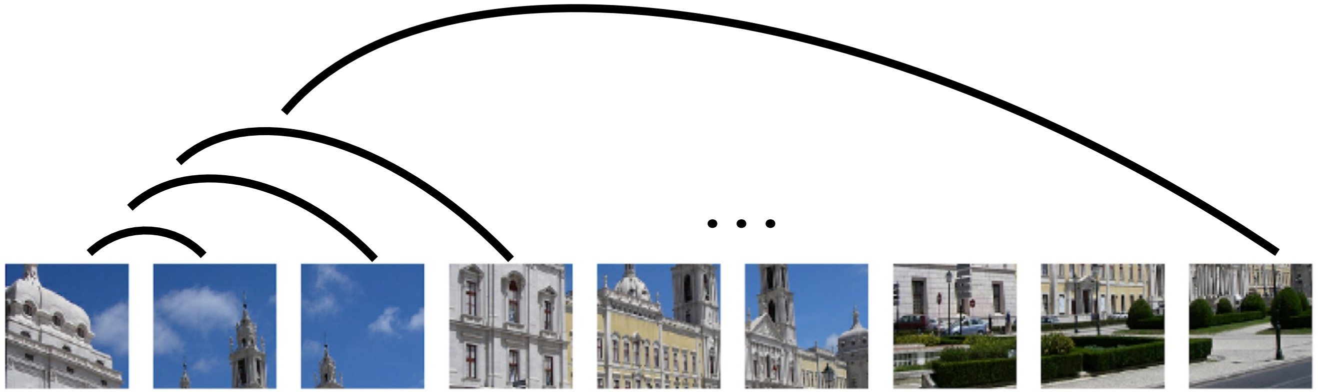
⇩

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

Long Sequence
Model Efficiency
Big Model ☹

Foundation
Model

Quadratic Complexity in Self-Attention



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

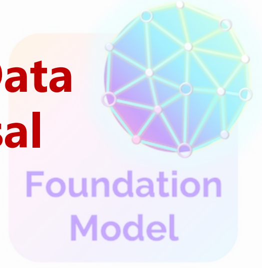
⇩

General Relation Modeling

[Quadratic Complexity]

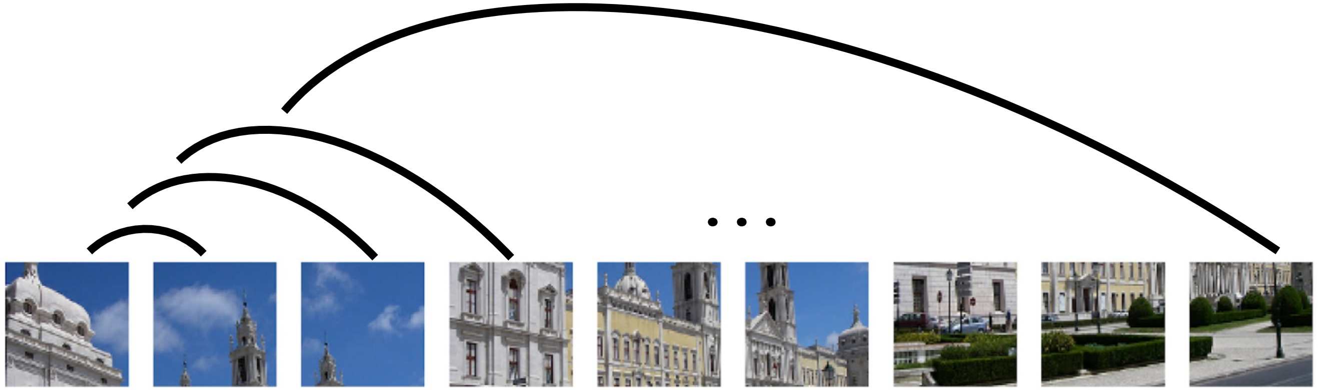
Long Sequence
Model Efficiency
Big Model ☹

Task & Data Universal



?

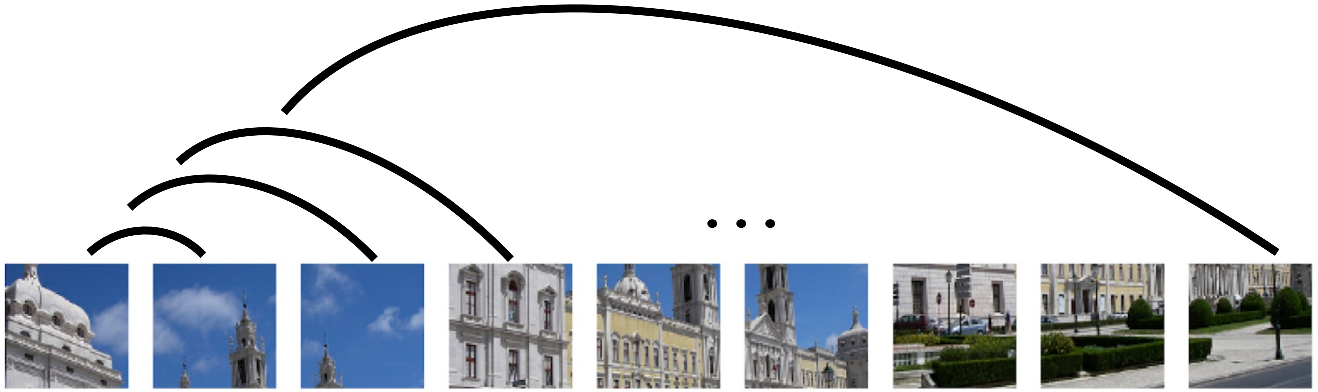
Quadratic Complexity in Self-Attention



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?

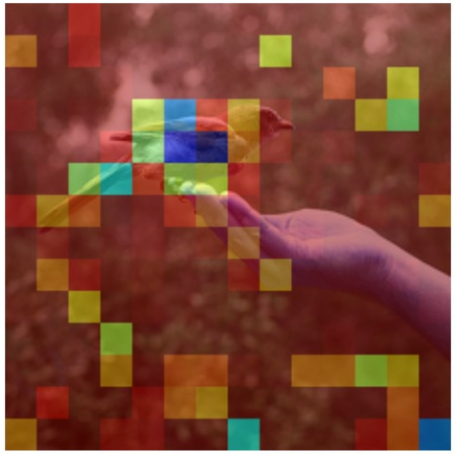
$(QK^T)V = Q(K^T V) \Rightarrow \mathcal{O}(n^2 d) \rightarrow \mathcal{O}(nd^2)$

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.

$$\phi(Q)(\phi(K)^T V)$$

$$+$$

Competition Mechanism

$$\Longleftrightarrow$$

$$\textbf{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

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.

$$\phi(Q)(\phi(K)^T V)$$

$$+$$

Competition Mechanism

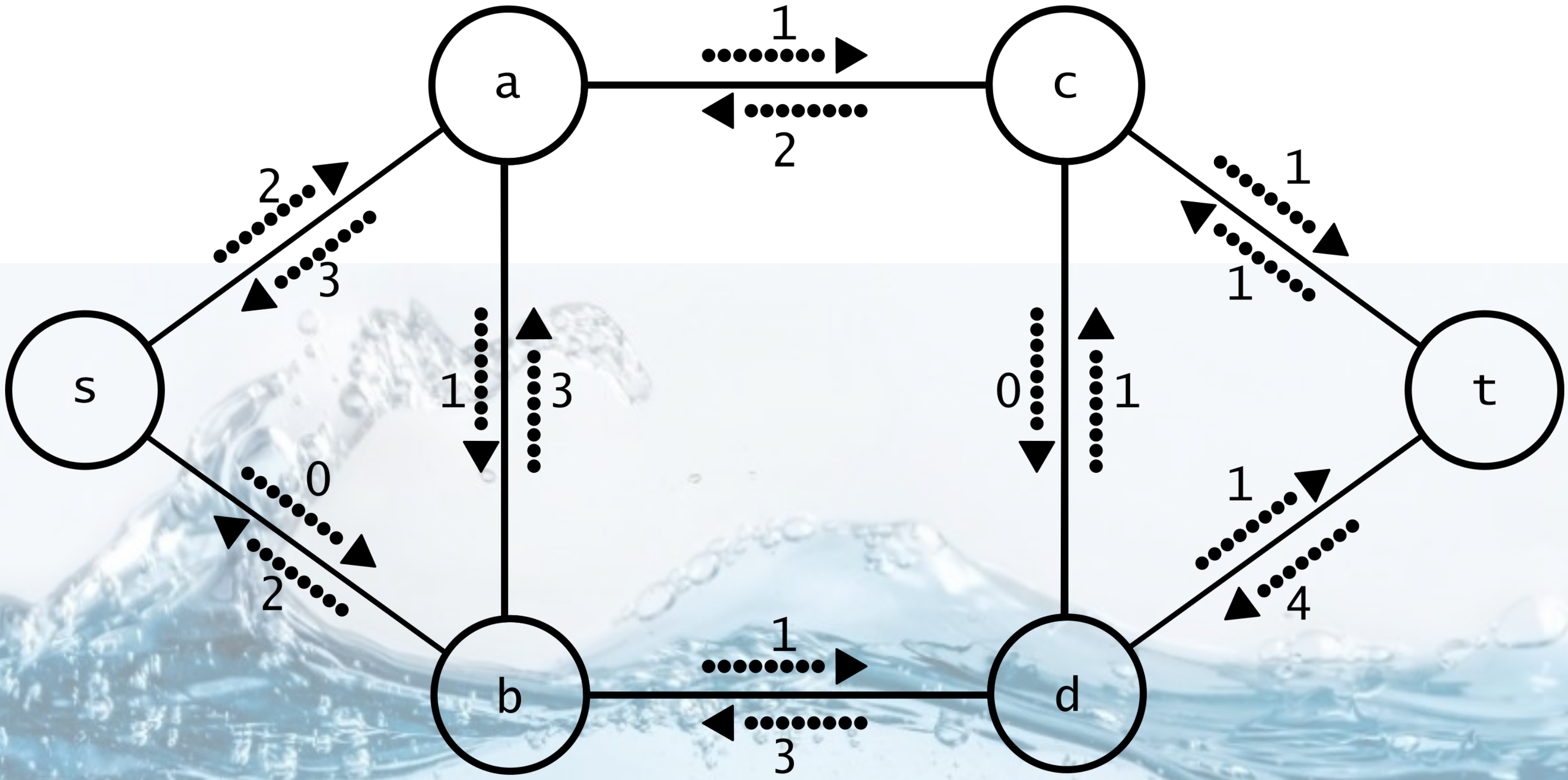
$$\Longleftrightarrow$$

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

"fixed resource will cause competition"

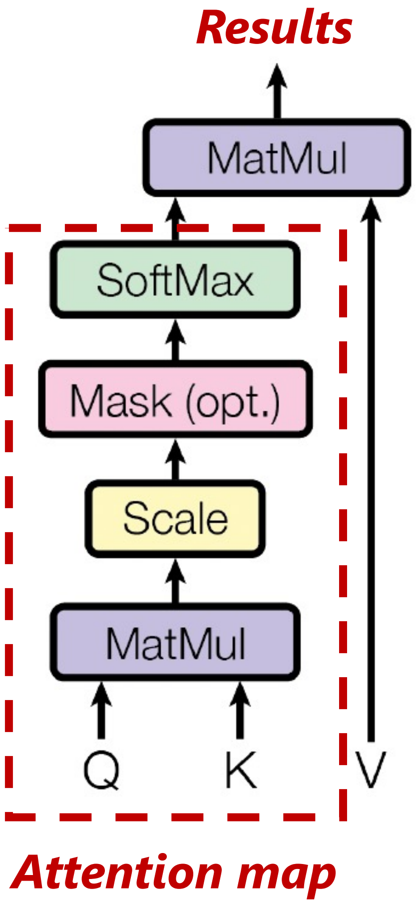
Bridle et al. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *NeurIPS 1989.*

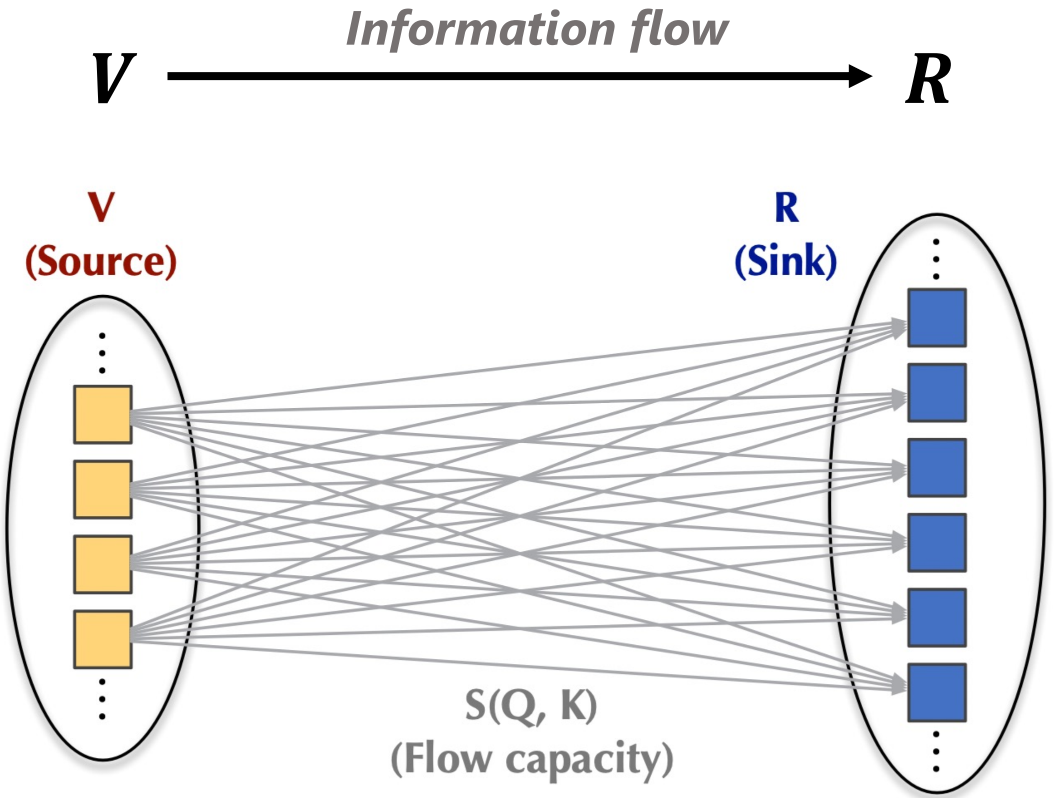
Flow Network Theory



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

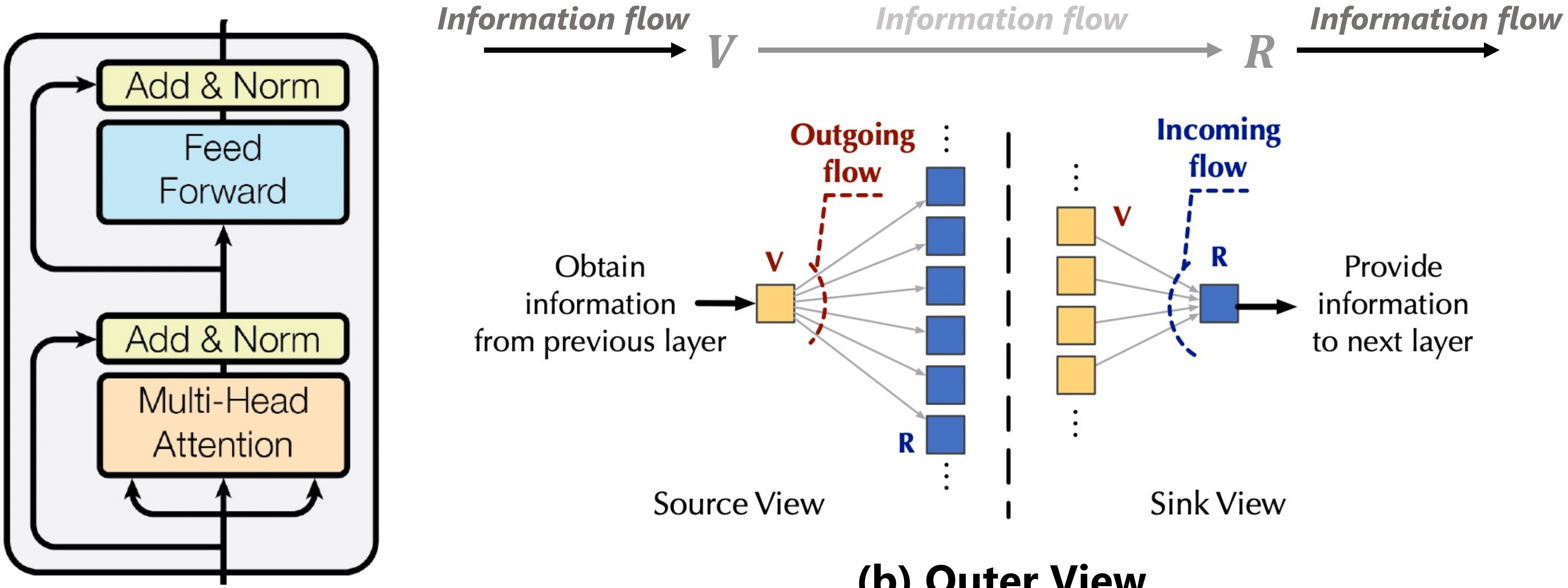
Attention: A Flow Network View





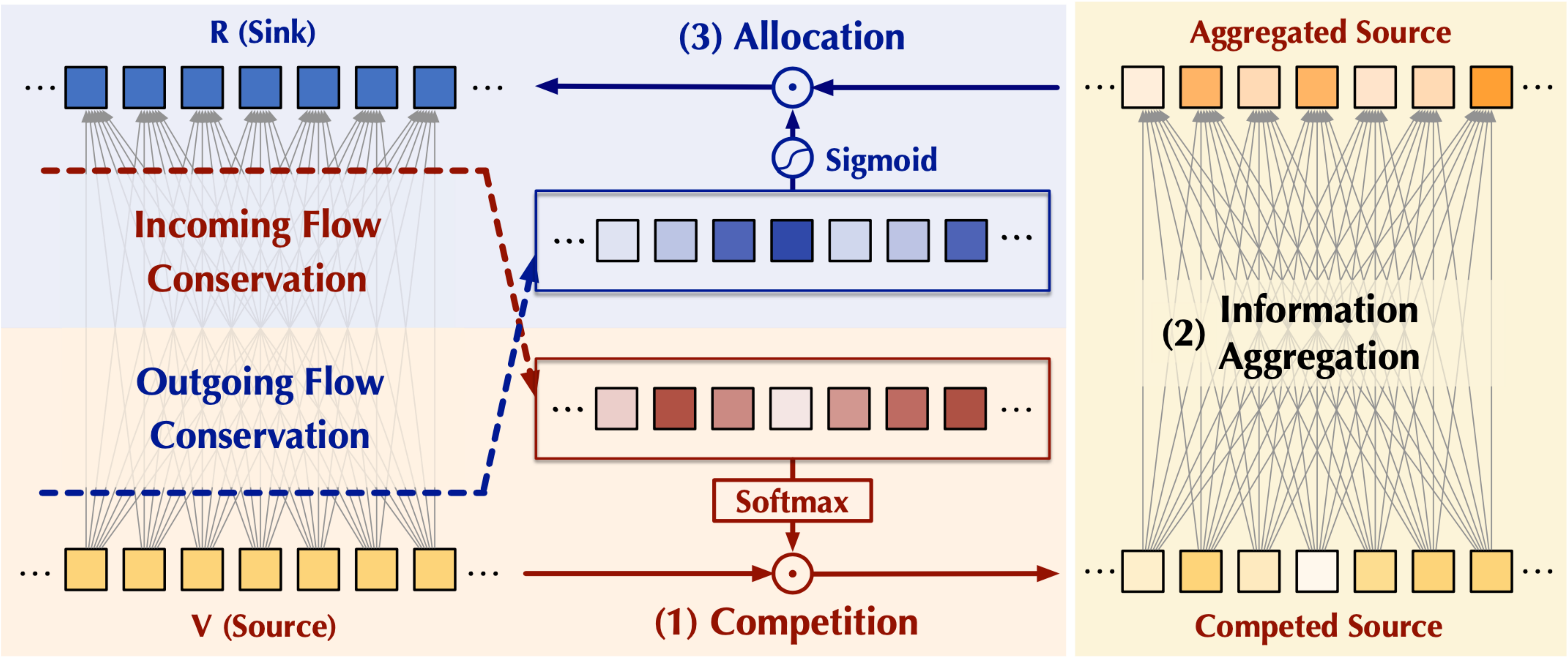
(a) Inner View

Attention: A Flow Network View



(b) Outer View

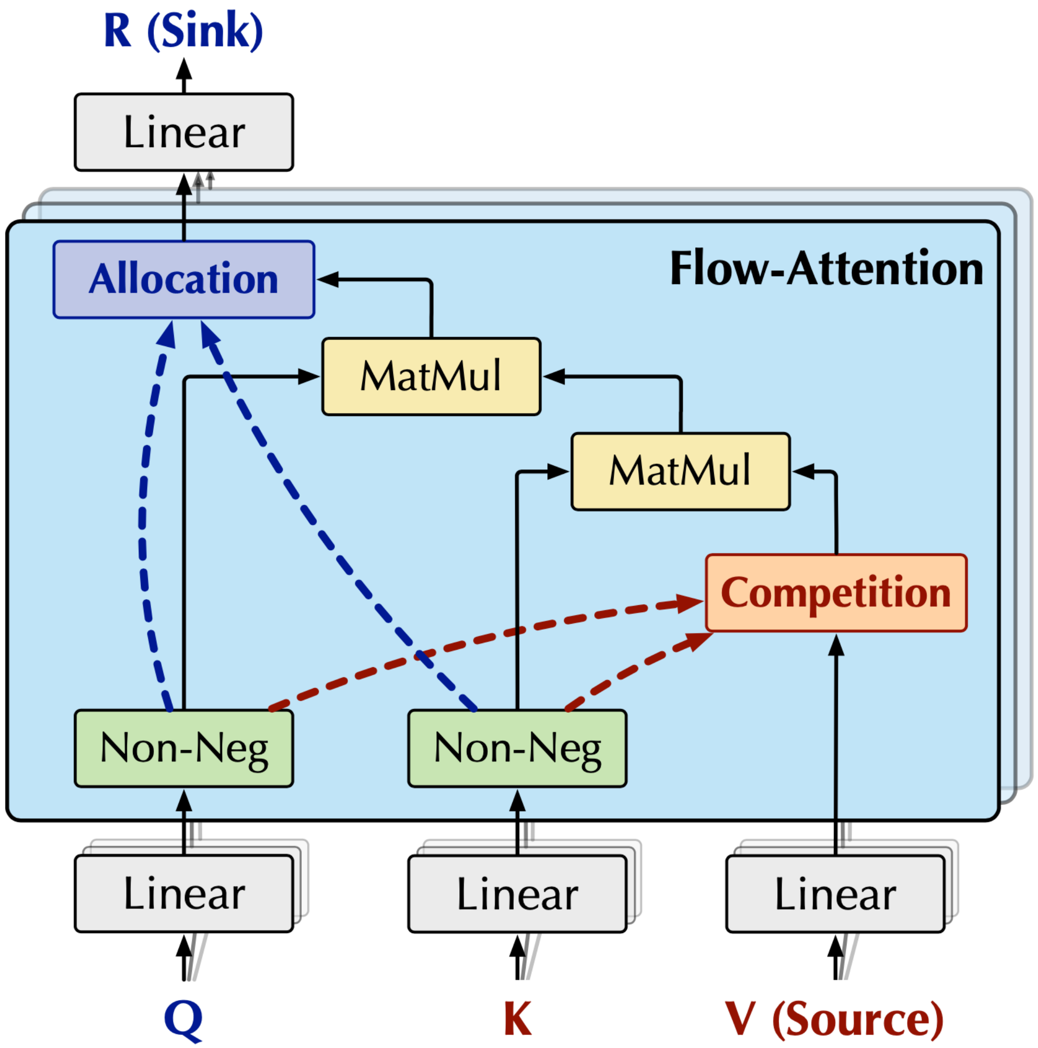
Conservation in Attention



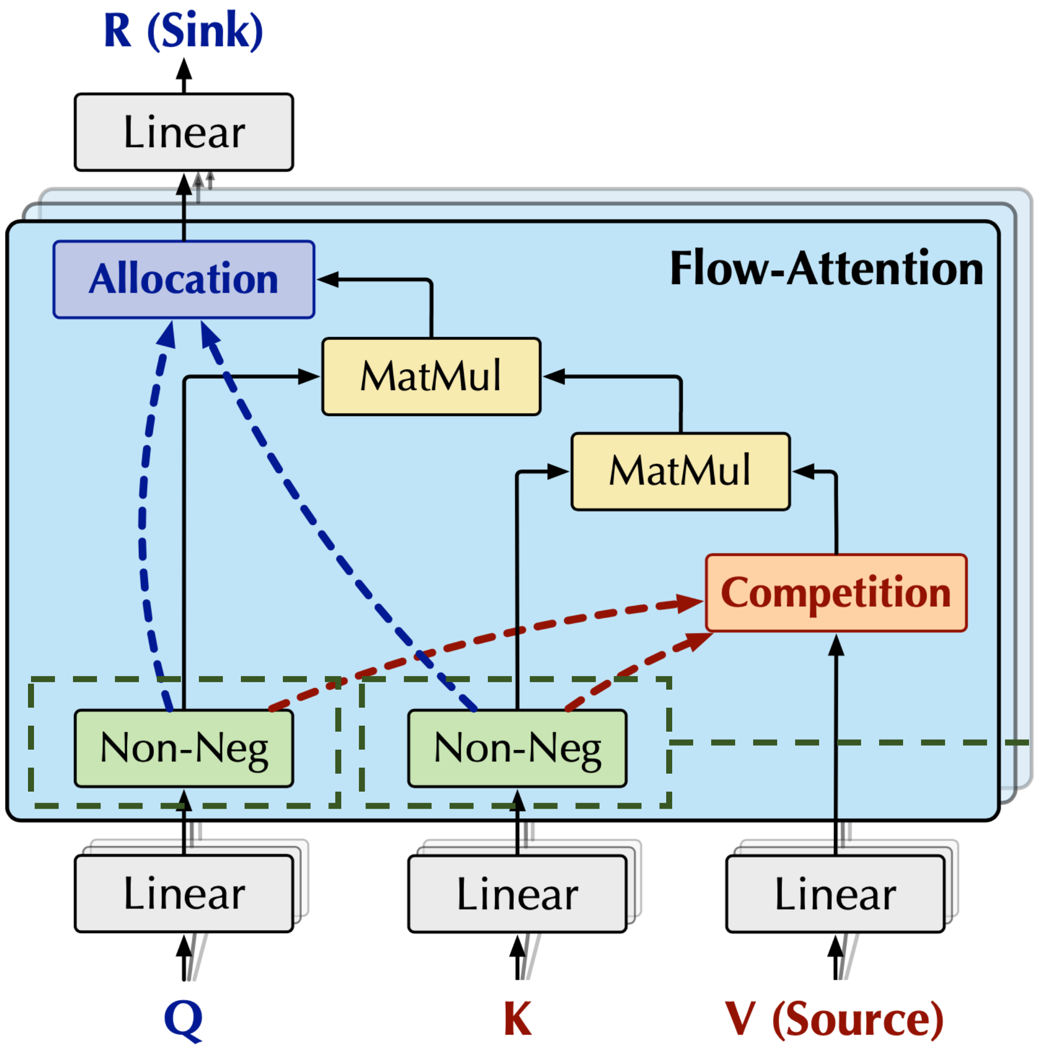
[Incoming Flow Conservation]: *Competition among Source tokens*

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

Flow-Attention

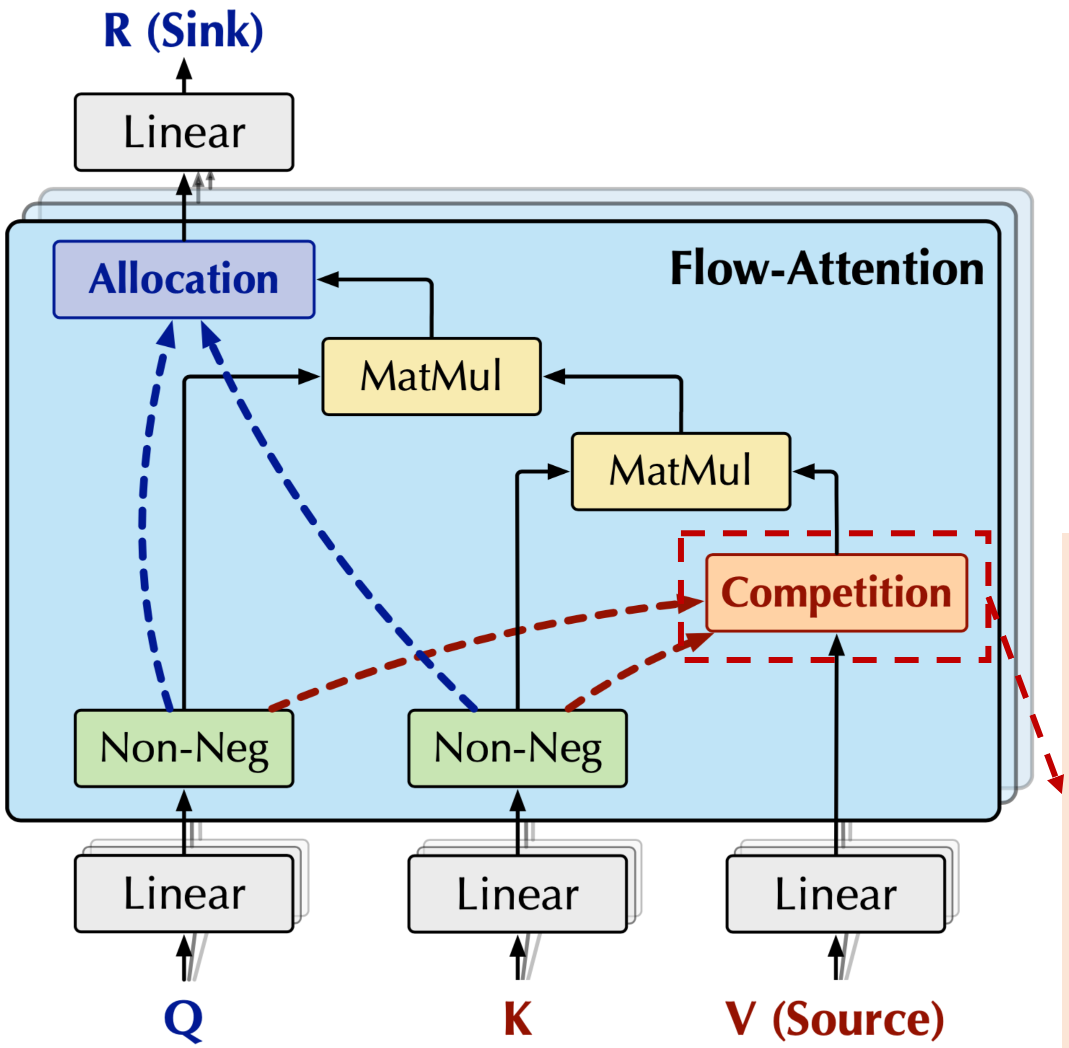


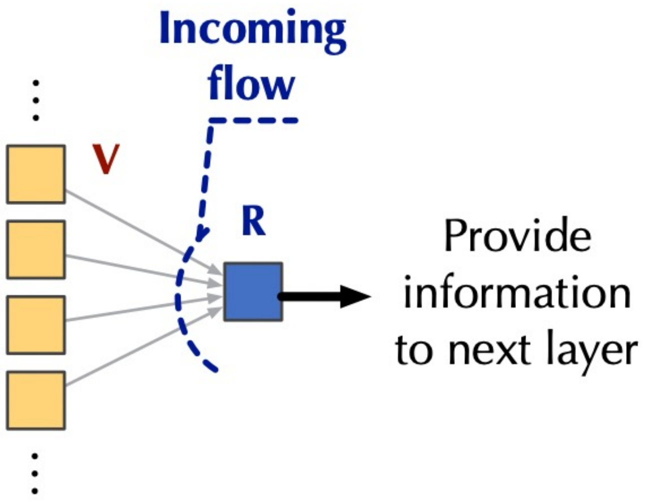
Flow-Attention



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

Flow-Attention



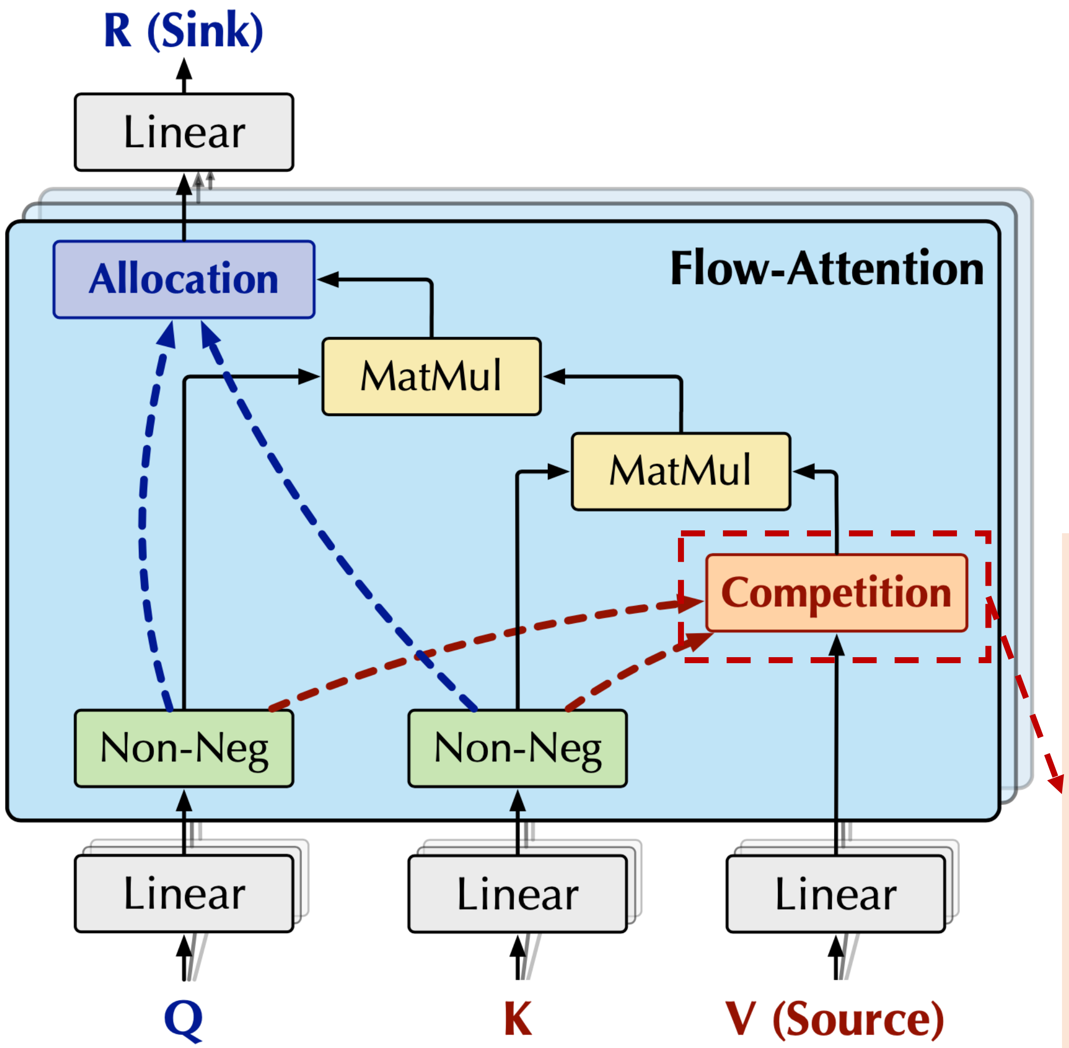


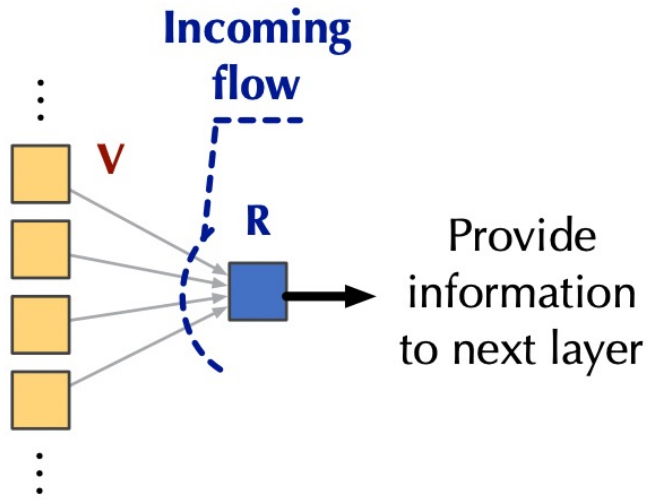
Incoming flow: $I_i = \phi(Q_i)\sum_j \phi(K_j)^T$

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

Incoming flow: $\dfrac{\phi(Q_i)}{I_i}\sum_j \phi(K_j)^T = \dfrac{I_i}{I_i} = 1$

Flow-Attention



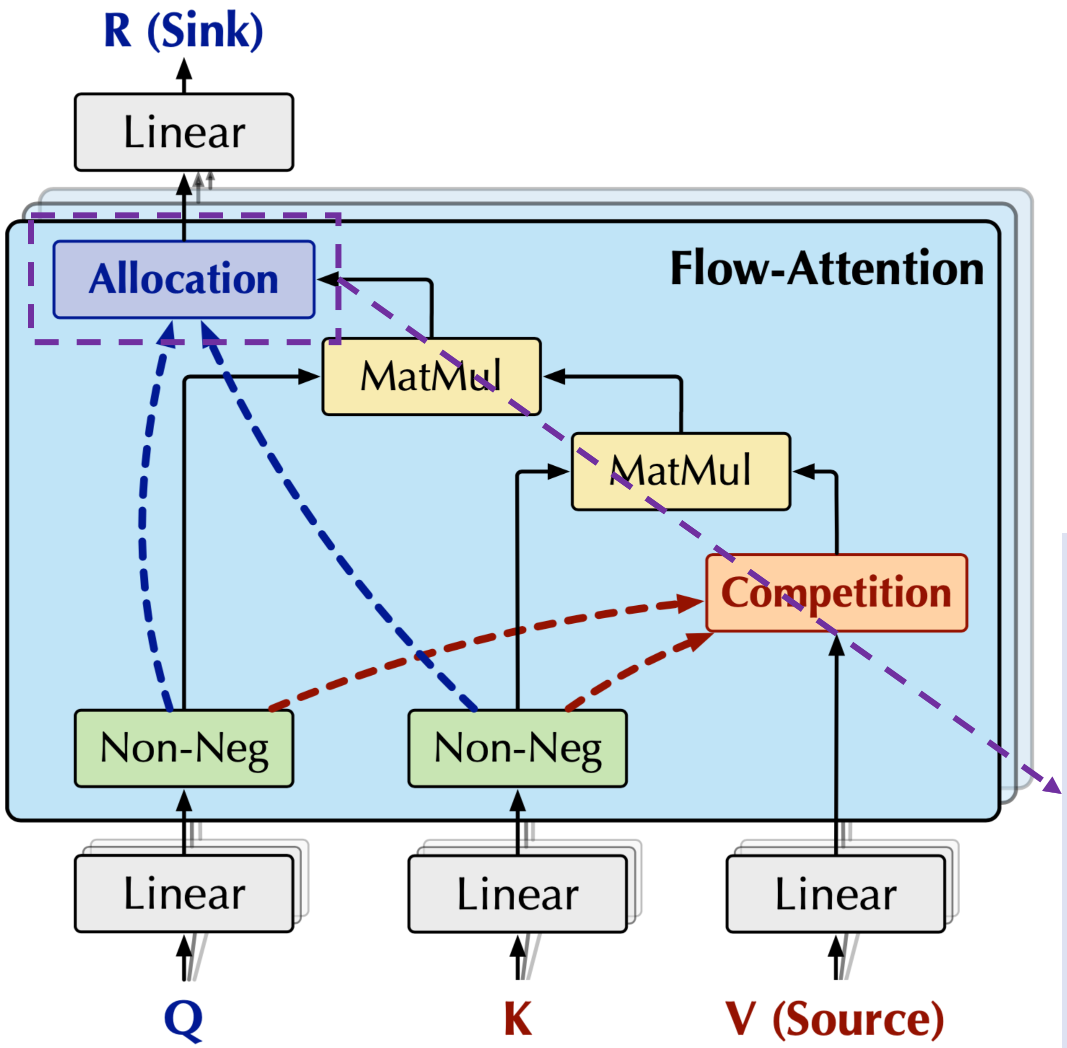


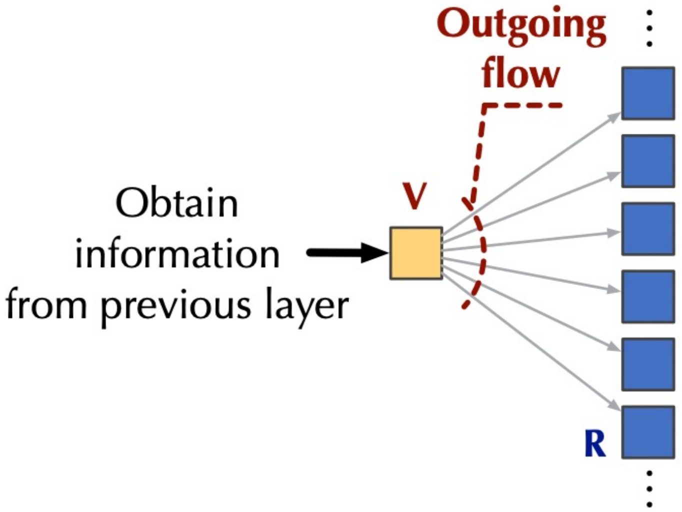
Incoming flow: $I_i = \phi(Q_i)\sum_j \phi(K_j)^T$

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

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

Flow-Attention



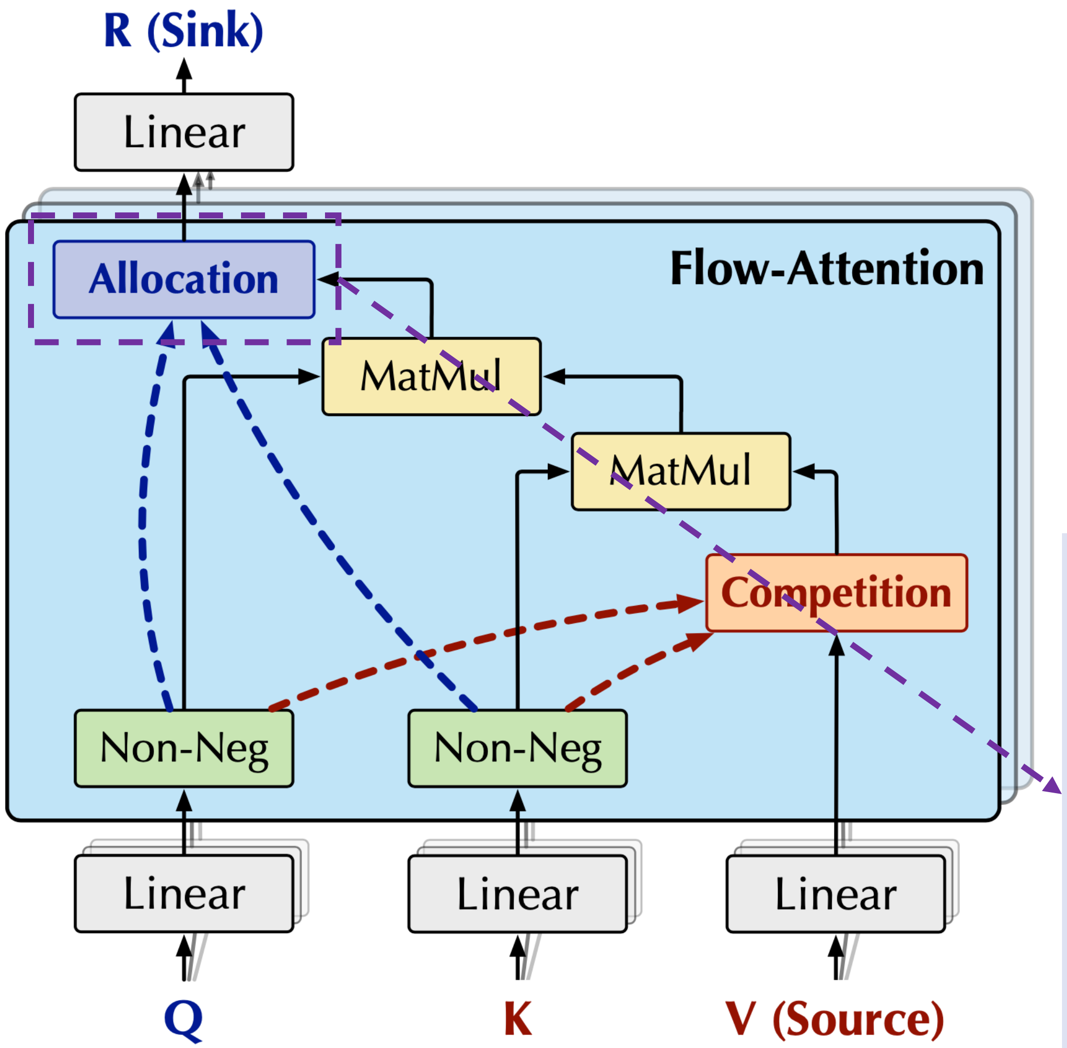


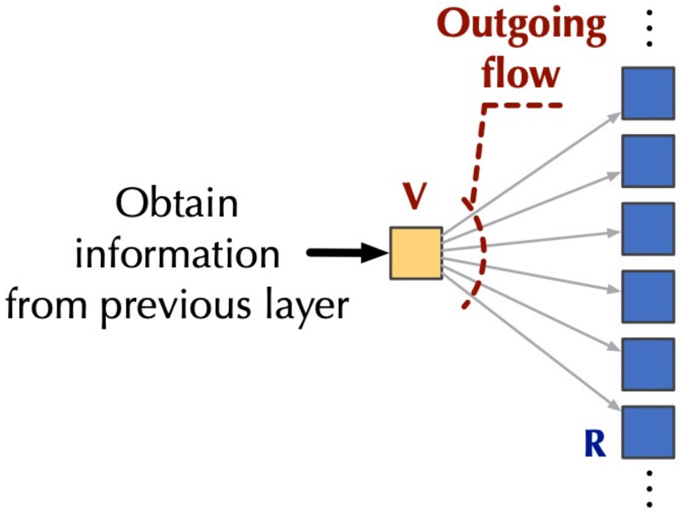
Outgoing flow: $O_i = \phi(K_i) \sum_j \phi(Q_j)^T$

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

Outgoing flow: $\frac{\phi(K_i)}{O_i} \sum_j \phi(Q_j)^T = \frac{O_i}{O_i} = 1$

Flow-Attention



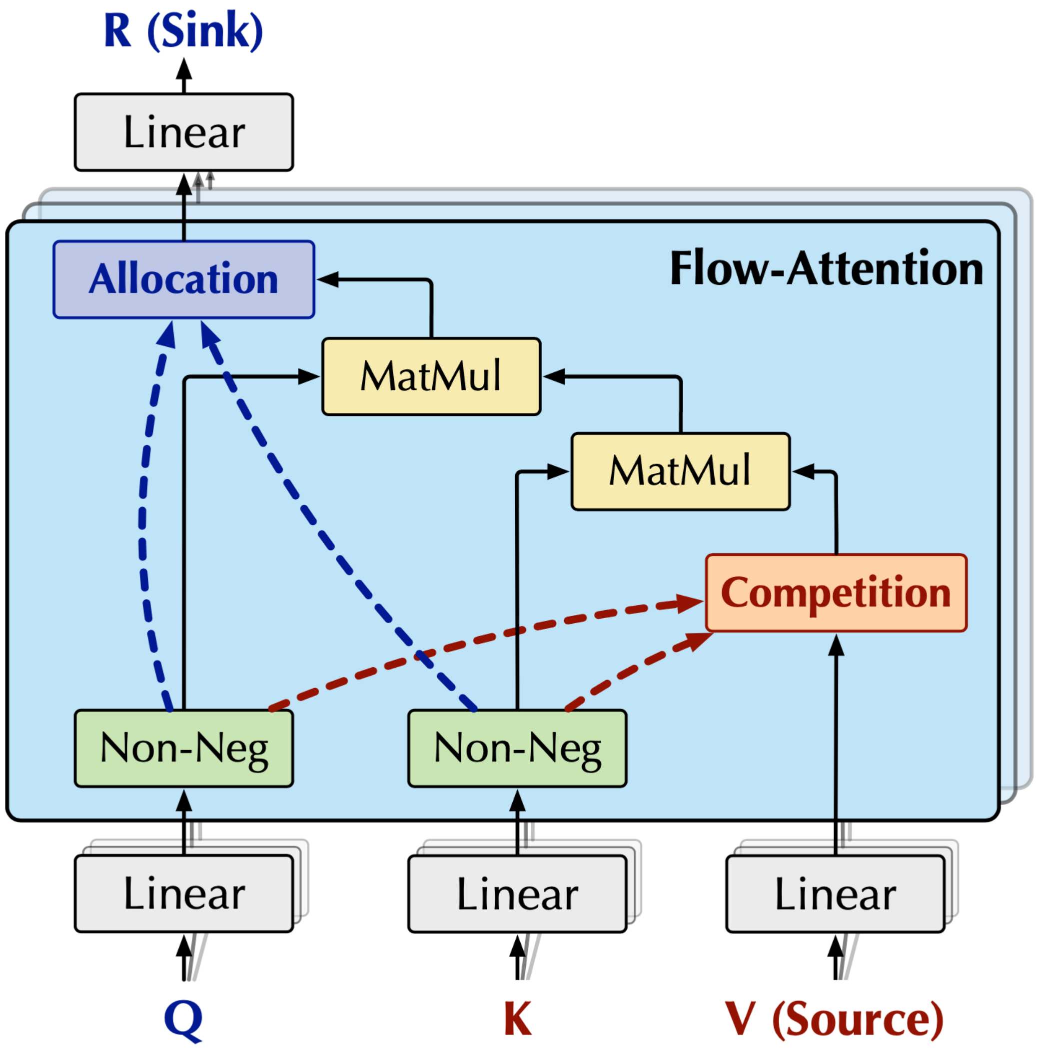


Outgoing flow: $O_i = \phi(K_i)\sum_j \phi(Q_j)^T$

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

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

Flow-Attention

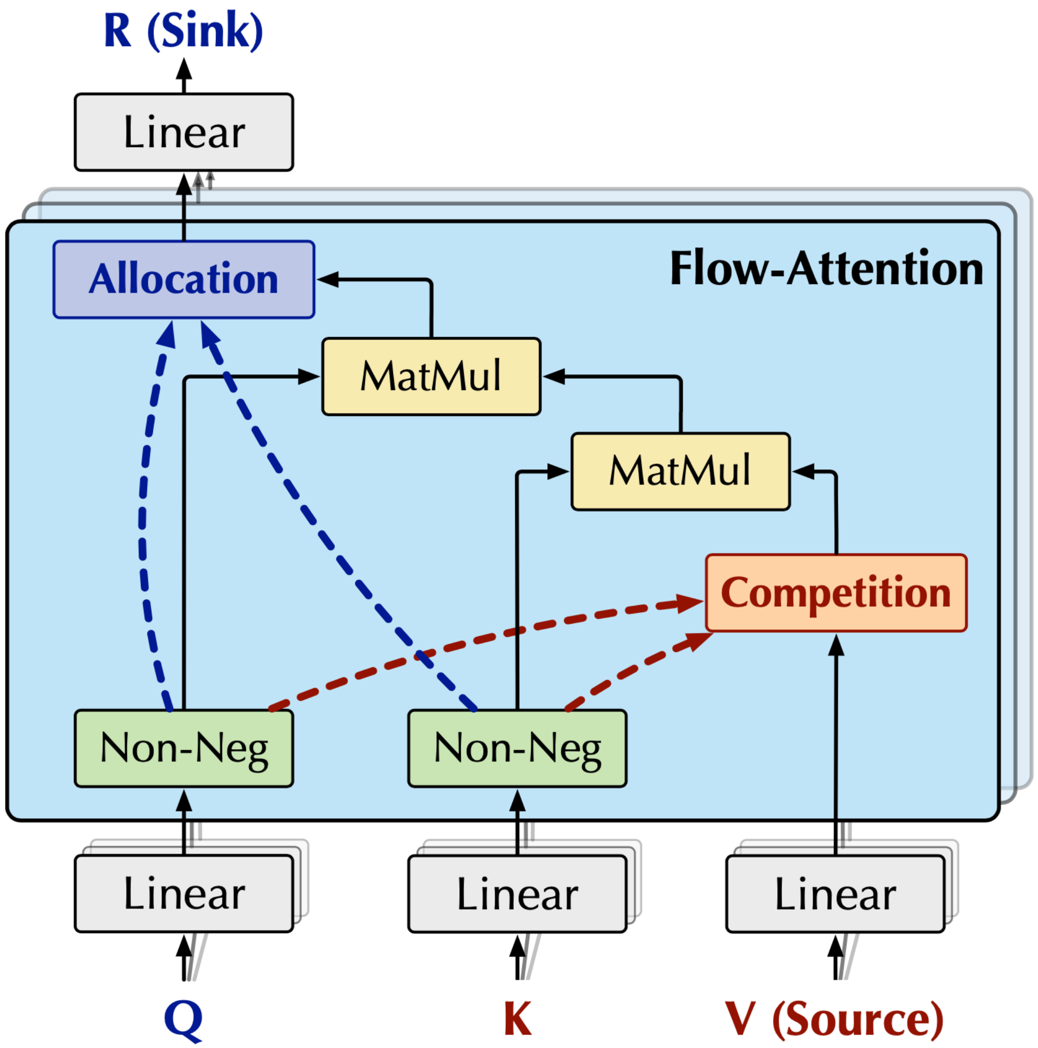


$$\text{Competition: } \widehat{\mathbf{V}} = \text{Softmax}(\widehat{\mathbf{O}}) \odot \mathbf{V}$$

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

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

Flow-Attention



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

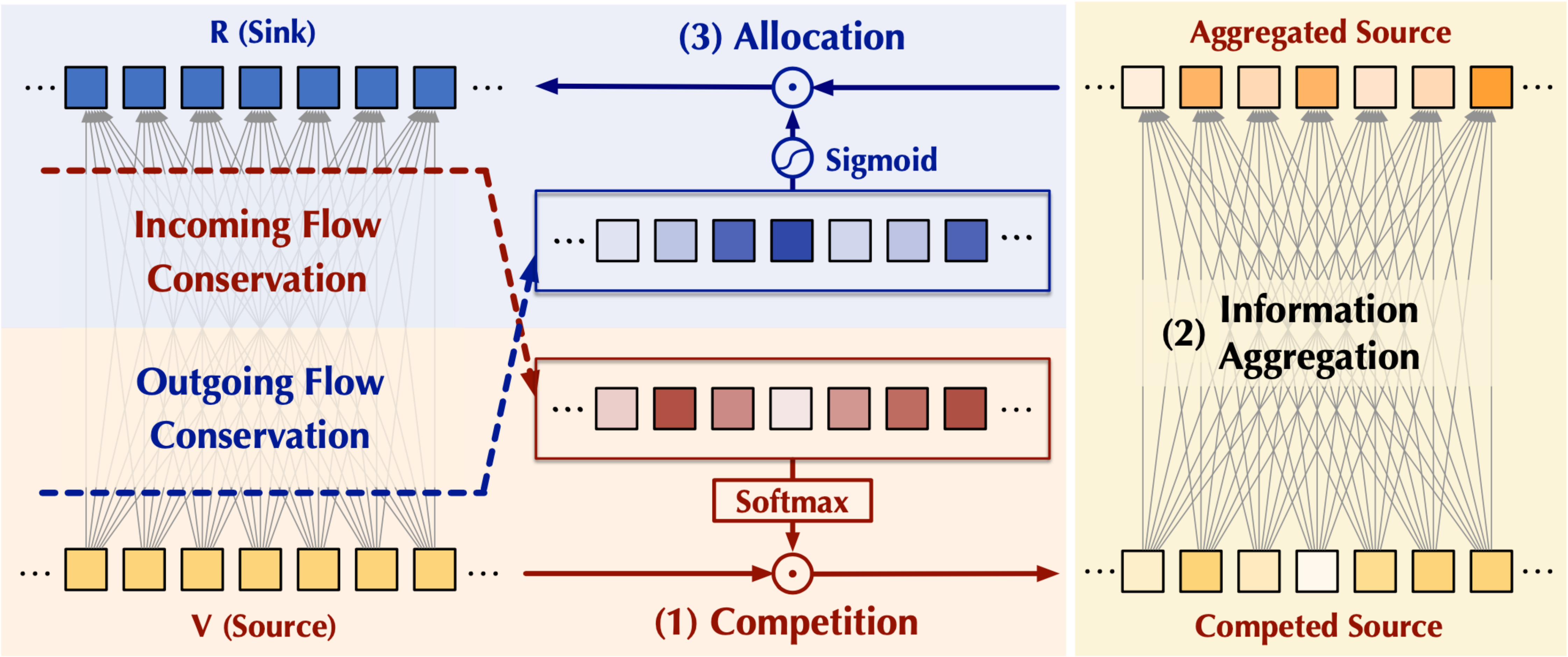
$$\text{Aggregation: } \mathbf{A} = \frac{\phi(\mathbf{Q})}{\mathbf{I}} (\phi(\mathbf{K})^{\mathsf{T}} \widehat{\mathbf{V}})$$

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

Successfully bring the *Competition Mechanism*

Into Attention design to avoid trivial attention

Efficiency and Universality



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

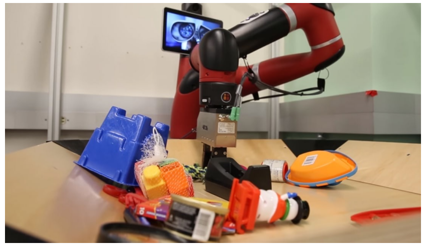
[Universality]: *The whole design is based on flow network* ***without specific inductive biases.***

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+)

Flowformer Experiments

Task	Metrics	Flowformer	Performer	Reformer	Vanilla Transformer
Long Sequence Modeling (LRA)	Avg Acc (%) ↑	**56.48**	51.41	50.67	OOM
Vision Recognization (ImageNet-1K)	Top-1 Acc (%) ↑	**80.6**	78.1	79.6	78.7
Language Modeling (WikiText-103)	Perplexity ↓	**30.8**	37.5	33.6	33.0
Time series classification (UEA)	Avg Acc (%) ↑	**73.0**	71.5	71.9	71.9
Offline RL (D4RL)	Avg Reward ↑ Avg Deviation ↓	**73.5** $\pm$ 2.9	63.8 $\pm$ 7.6	63.9 $\pm$ 2.9	72.2 $\pm$ **2.6**

Strong performance on all five mainstream tasks within the linear complexity.

Summary

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

Long Sequence
Model Efficiency
Big Model ☹

Foundation Model

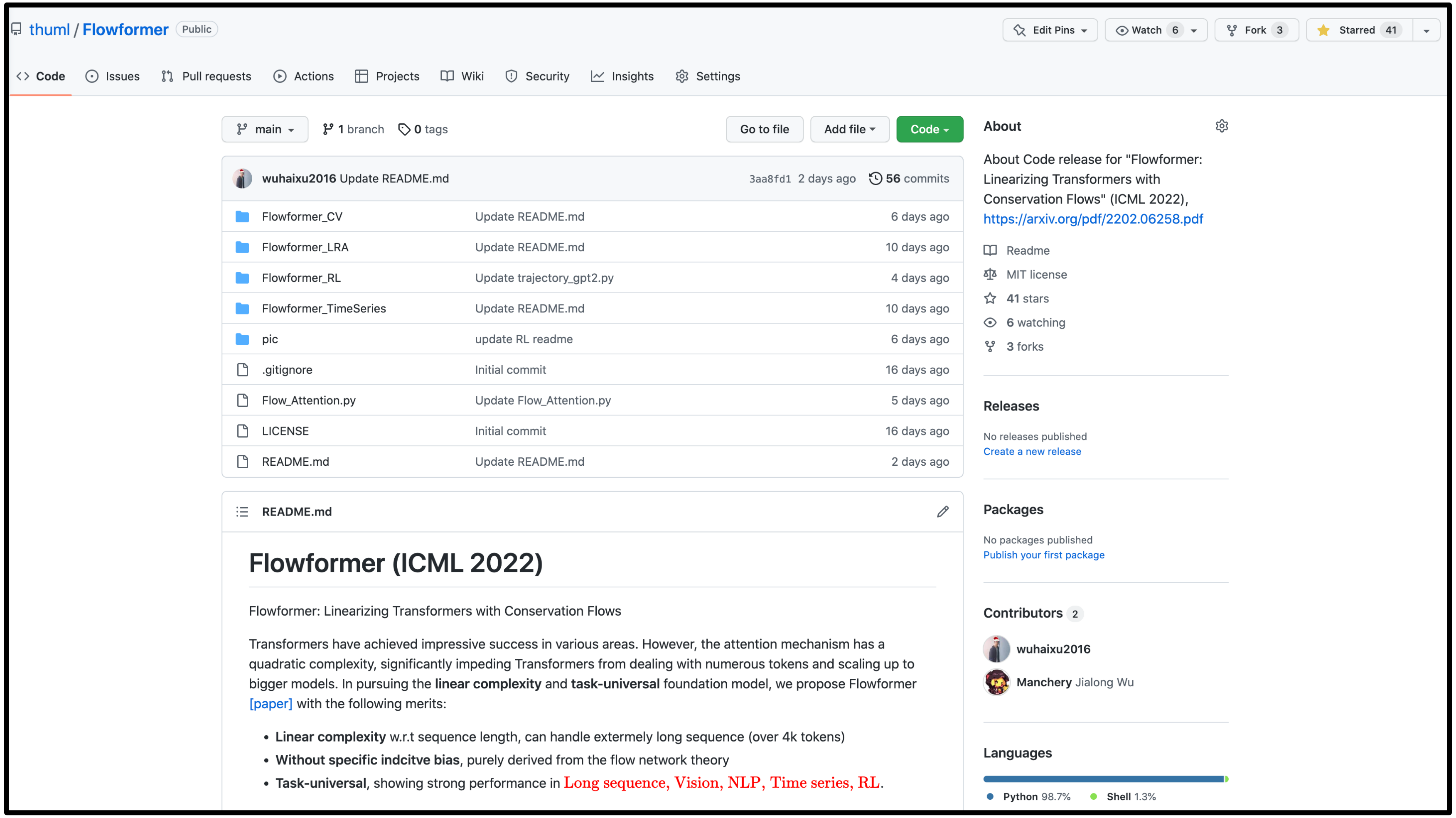
Flowformer

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

长按关注，获取最新资讯