Attention as Discrete-Time Markov Chains

Kamil Książek

December 8, 2025

Agenda

- Summary of the paper
- Markov chains
- PageRank
- Attention operations in terms of Markov chains
- Downstream tasks' performance

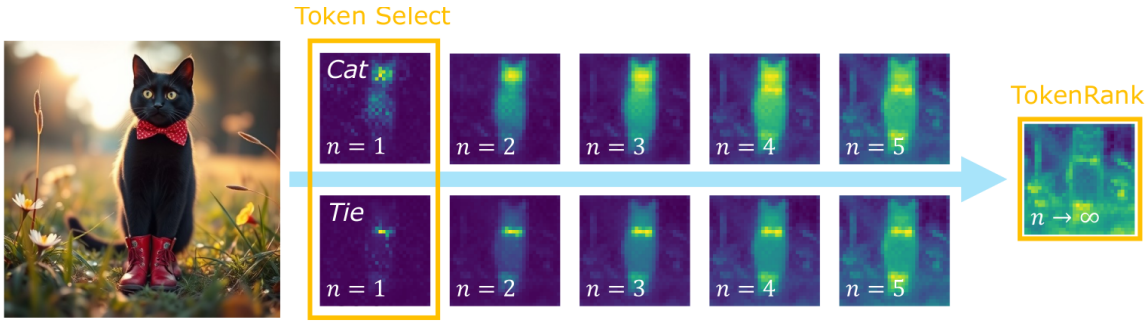
Attention *(as Discrete-Time Markov)* Chains

Yotam Erel[1*] **Olaf Dünkel**[2*] **Rishabh Dabral**[2]
Vladislav Golyanik[2] **Christian Theobalt**[2] **Amit H. Bermano**[1]
[1]Tel Aviv University [2]MPI for Informatics, SIC
https://yoterel.github.io/attention_chains/

Published at NeurIPS 2025 (poster).
One of the main weaknesses mentioned by the reviewers is the lack of experiments with NLP...

Summary

- They interpret the attention matrix as a **discrete-time Markov chain**.
- They define *TokenRank*, i.e. the steady state vector of the Markov chain, measuring **global token importance**.



n–th order attention bounce models higher-order attention effects, while a stationary vector ($n \to \infty$) globally captures the flow of attention into each token.

Markov chains

- Assume that X_n denotes a value of a certain process in the n–th time period.
- $\{X_n, n = 0, 1, 2, ...\}$ is a stochastic process that takes on a finite or countable number of possible values.
- If $X_n = i$, then the process is in state i at time n while P_{ij} is a fixed probability that it will be next in state j.

A **Markov chain** is a stochastic process such that for all states $i_0, i_1, ..., i_{n-1}$ and all $n \geqslant 0$

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0\} = P_{ij} \quad (1)$$

Markov chains

We have the following properties:

$$P_{ij} \geqslant 0, \quad i, j \geqslant 0, \quad \sum_{j=0}^{\infty} P_{ij} = 1, \quad i = 0, 1, ...$$

P denotes the matrix of one-step transition probabilities P_{ij}:

$$P = \begin{pmatrix} P_{00} & P_{01} & P_{02} & \dots \\ P_{10} & P_{11} & P_{12} & \dots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

Markov chains

- Let define by P_{ij}^n a probability that a process in state i will be in state j after n transitions, i.e. the n-step transition probability:

$$P_{ij}^n = P\{X_{n+k} = j | X_k = i\}, \quad i, j, n \geqslant 0.$$

- The matrix of n-step transition probabilities ($\boldsymbol{P}^{(n)}$) can be calculated by multiplying the matrix P by itself n times.

Markov chains

Theorem: Assume we have an irreducible ergodic Markov chain. Then, $\lim_{n\to\infty} P_{ij}^n$ exists and is independent of i. Also, if

$$\pi_j = \lim_{n\to\infty} P_{ij}^n, \quad j \geqslant 0,$$

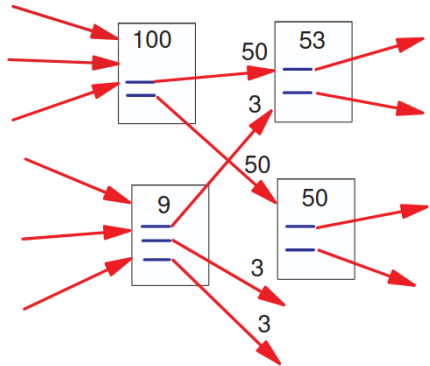
then π_j is the unique nonnegative solution of

$$\pi_j = \sum_{i=0}^{\infty} \pi_i \cdot P_{ij}, \quad j \geqslant 0 \quad \text{and} \quad \sum_{j=0}^{\infty} \pi_j = 1.$$

$\pi_j, j \geqslant 0$ are called **stationary probabilities**.

Author's idea

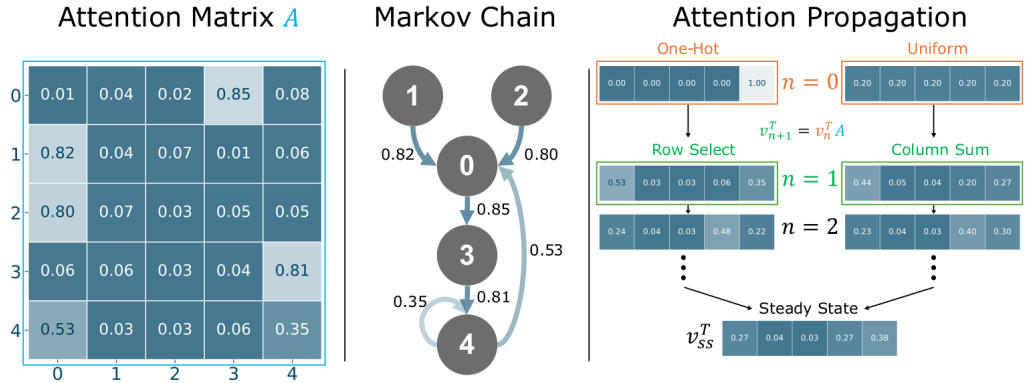
- Non-negative attention weights indicate **transition probabilities** between states that correspond to tokens.
- The analogue is **Google PageRank** algorithm, stating that simple counting of incoming hyperlinks in general does not correspond to page importance.



Simplified PageRank

Source: L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

Attention matrix as a discrete-time Markov chain

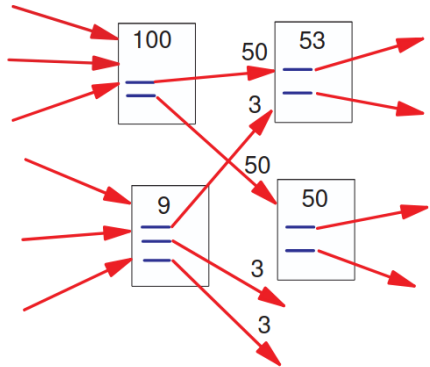


- Only strong connections are shown.
- In the second bounce ($n = 2$) more probability mass is directed into *state-3*.
- The steady state ranks *state-4* as the most important state globally.

PageRank

- $\boldsymbol{P}$ is a stochastic process where states are web pages and outgoing links define transition probabilities.
- $\boldsymbol{P}$ is normalized and pages without outgoing links are replaced with uniform vectors.
- $$\boldsymbol{P'} = \alpha \boldsymbol{P} + (1-\alpha)\frac{1}{m}\boldsymbol{e}\boldsymbol{e}^{\top}, \qquad (2)$$

  $\alpha \in (0,1)$ controls the probability of *teleportation* into a random state, m is the number of states, $\boldsymbol{e}$ is a column vector filled ones.



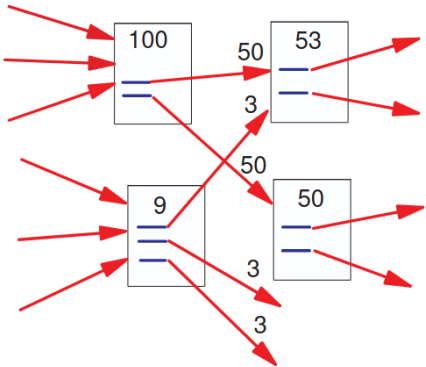
Simplified PageRank

Source: L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

PageRank

To get a unique steady-state vector $\vec{\pi}$, considered as a global PageRank vector, it its necessary to apply the power method on $\boldsymbol{P'}$:

$$i \preceq_{\vec{\pi}} j \quad \Leftrightarrow \quad \vec{\pi}_i \leqslant \vec{\pi}_j. \tag{3}$$



Simplified PageRank

Source: L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

Power method

- Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a right-stochastic matrix.
- To obtain the steady-state vector $\boldsymbol{v}_{ss}$ we can use the power method and iteratively compute:

$$\boldsymbol{v}_{n+1}^{\top} = \boldsymbol{v}_{n}^{\top} \boldsymbol{A}, \tag{4}$$

  where $\boldsymbol{v}_0^{\top}$ is an initial state summing to one.
- The calculations are stopped after a fixed number of iterations or when $\| \boldsymbol{v}_{n+1}^{\top} - \boldsymbol{v}_{n}^{\top} \|_2^2 < \tau$.
- To get the global PageRank vector, one has to apply the power method on $\boldsymbol{P}'$.

Attention operations

- After softmax, the **attention matrix becomes a right-stochastic matrix**.
- **LIMITATION:** The method is limited to equal-size attention blocks (e.g. self-attention or hybrid attention).

MULTIPLYING attention matrices of consecutive blocks of a transformer is equivalent to chaining several Markov chains.

Attention operations

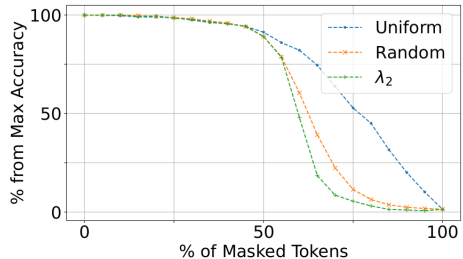
ADDING an identity matrix to an attention matrix and re-normalizing it is equivalent to computing a new chain $\boldsymbol{A}' = 0.5(\boldsymbol{I} + \boldsymbol{A})$. It does not change the steady state because $\boldsymbol{A}'\boldsymbol{v}_{ss} = \boldsymbol{v}_{ss}$.

SUMMING each column of $\boldsymbol{A}$ leads to a row vector $\boldsymbol{v}_s^\top$ that aggregates attention to each token from all other tokens. It is a first-order approximation in a global ranking of tokens' importance.

Attention operations

AVERAGING multiple attention matrices over different attention heads yields a new process:

- Transitioning probabilities are the mean of those from the original chains.
- Some heads are not informative.
- **Weighted averaging** over the heads using the **second largest eigenvalue** λ_2.
- Larger λ_2 values correspond to more stable *metastable* states, i.e. regions where attention tends to concentrate among tokens in semantically similar regions.



Different weighting schemes for the head dimension. λ_2 can be used as a useful tool for determining importance of heads.

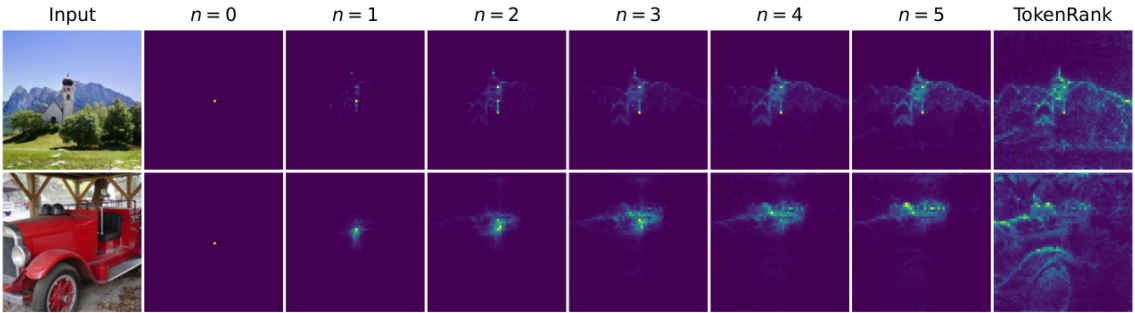
Multi-bounce attention

- To take into account the n–th order attention bounce for token i, we can apply the selection operation n times:

$$\boldsymbol{v}_{\{i,n+1\}}^{\top} = \boldsymbol{v}_{\{i,n\}}^{\top} \boldsymbol{A}. \tag{5}$$

- Applying multi-bounce attention will converge into a stationary vector (*TokenRank*) for any initial state, according to the assumptions of the theorem for Markov chains.
- Due to possible cycles or reduction of attention weight matrices, $\boldsymbol{A}$ should be adjusted using the PageRank equation (2).

Multi-bounce attention



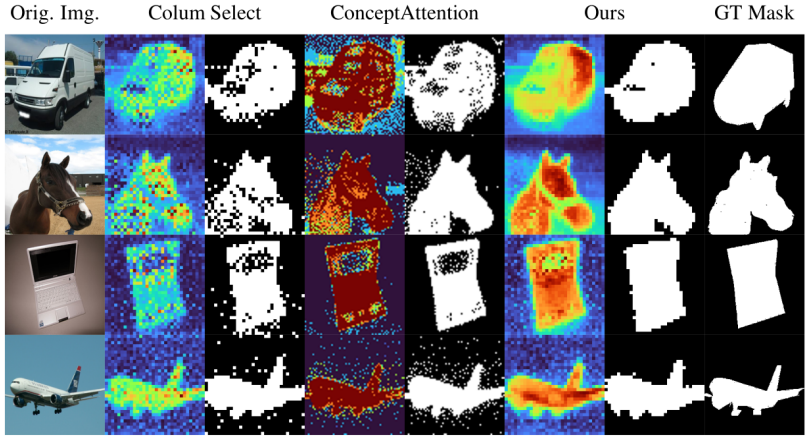
Visualization of the first bounces and the steady state (*TokenRank*). The Markov chain was defined by two attention matrices of two exemplary layers and heads of DINOv2. For $n = 0$, a one-hot vector was used.

Zero-shot segmentation

Method	Architecture	Acc ↑	mIoU ↑	mAP ↑
LRP (Binder et al., 2016)	ViT-B/16	51.09	32.89	55.68
Partial-LRP (Binder et al., 2016)	ViT-B/16	76.31	57.94	84.67
Rollout (Abnar and Zuidema, 2020)	ViT-B/16	73.54	55.42	84.76
ViT Attention (Dosovitskiy et al., 2021)	ViT-B/16	67.84	46.37	80.24
GradCam (Selvaraju et al., 2017)	ViT-B/16	64.44	40.82	71.60
DiffSeg Tian et al. (2024)	SD1.4	65.41	52.12	-
TextSpan (Gandelsman et al., 2024)	ViT-H/14	75.21	54.50	81.61
TransInterp (Chefer et al., 2021b)	ViT-B/16	79.70	61.95	86.03
DINO Attention (Caron et al., 2021)	ViT-S/8	81.97	69.44	86.12
DAAM (Tang et al., 2023)	SDXL UNet	78.47	64.56	88.79
FLUX Cross Attention (Helbling et al., 2025)	FLUX DiT	74.92	59.90	87.23
FLUX row-select	FLUX DiT	73.96	54.65	82.64
FLUX column-select	FLUX DiT	80.55	64.02	87.20
Concept Attention (Helbling et al., 2025)	FLUX DiT	83.07	**71.04**	90.45
Ours w/o λ_2	FLUX DiT	<u>84.00</u>	70.02	<u>94.28</u>
Ours	FLUX DiT	**84.12**	<u>70.20</u>	**94.29**

Results for ImageNet segmentation.

Zero-shot segmentation



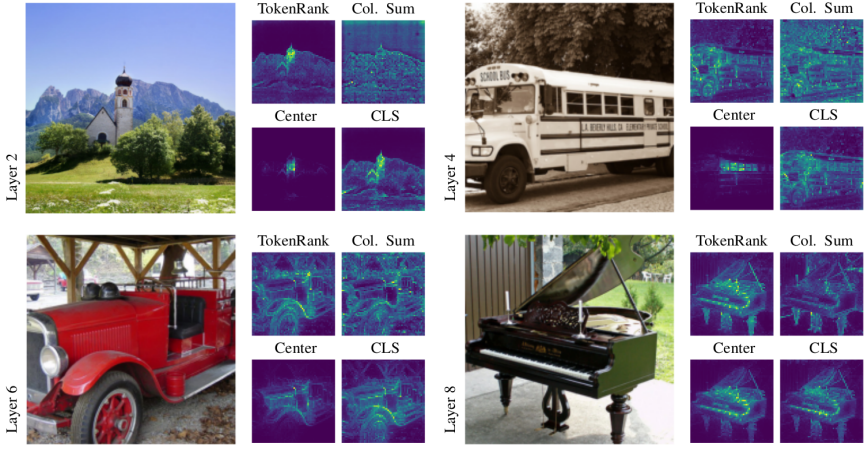
Presentation of raw attention output and binary segmentation masks for different methods. As a baseline, column select operation (outgoing attention for tokens that directly attend a given token) is depicted.

Partial segmentation

By setting the initial one-hot vector to different text tokens we can segment the corresponding image parts.

The text prompt was, as follows: *cute black cat standing up wearing red boots and a bow tie, photorealistic, masterpiece, macro wildlife photography, dewy grass, closeup low angle, wildflowers, sun, shadows, depth of field, desaturated, film grain, low quality*.

Visualizations of attention maps



Authors emphasize: *TokenRank considers indirect attention paths to evaluate where attention is flowing into*. The figure depicts visualizations after averaging over heads for four different layers of DINOv1. CLS token was explicitly trained to capture global attention for DINOv1.

Linear probing

Authors train a linear classifier on top of all proposed attention visualizations for all layers, heads and images for *Imagenette*. They consider three foundation models, without taking the CLS token.

Authors emphasize: *TokenRank visualizations results in a cleaner signal for image classification than previous approaches that only consider one bounce of attention.*

Linear probing

Accuracy	DINOv1	DINOv2	CLIP
Rand. Token	67.33 ± 3.12	63.32 ± 6.04	57.65 ± 2.66
Center Token	66.32 ± 3.08	75.34 ± 1.85	68.42 ± 2.66
Column Sum	75.64 ± 2.74	90.76 ± 2.21	73.73 ± 2.98
TokenRank	$\mathbf{77.31 \pm 2.50}$	$\mathbf{92.73 \pm 1.51}$	$\mathbf{73.88 \pm 2.86}$
CLS Token	81.53 ± 2.44	94.07 ± 0.97	72.46 ± 3.40

Linear probing of steady state vectors after aggregating heads.

Method	uniform	random	λ_2
CLIP	**49.07**	44.07	**49.07**
DINOv1	53.50	47.21	**53.55**
DINOv2	71.62	50.29	**72.36**

Unconditional image generation with denoising diffusions

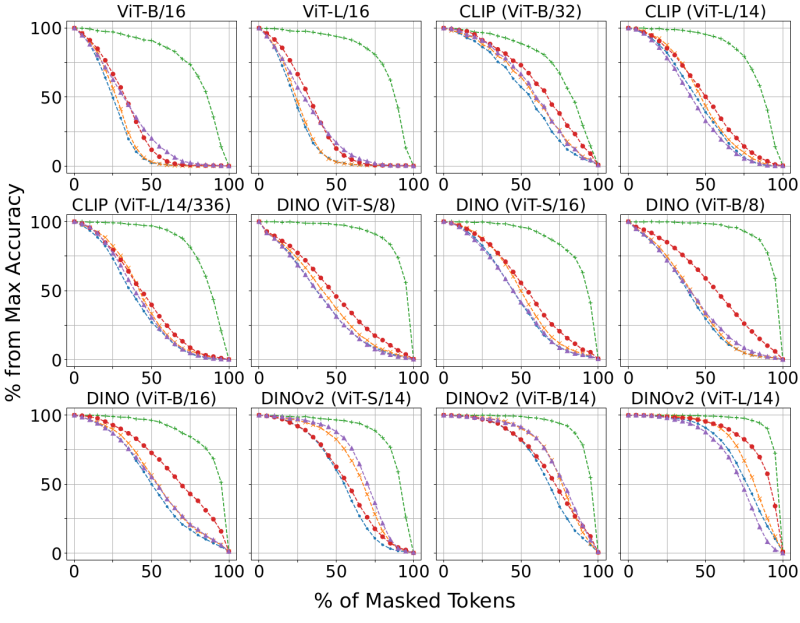
Self-Attention Guidance uses the self-attention weights to form a mask indicating important spatial tokens. Then, it adversarially blurs the masked regions and drives the denoising process away from it. Here, authors use SD1.5.

Token masking

Classification accuracy degrades faster using TokenRank than other baselines.

Results for masking most influential tokens.

AUC ↓	ViT	CLIP	DINOv1	DINOv2
Rand. Token	0.79	0.80	0.88	0.89
Center Token	0.33	0.47	0.45	0.70
Column Sum	0.27	0.49	0.47	0.71
CLS Token	0.33	0.53	0.56	0.70
TokenRank	**0.26**	**0.46**	**0.44**	**0.64**


ViT-B/16
ViT-L/16
CLIP (ViT-B/32)
CLIP (ViT-L/14)
CLIP (ViT-L/14/336)
DINO (ViT-S/8)
DINO (ViT-S/16)
DINO (ViT-B/8)
DINO (ViT-B/16)
DINOv2 (ViT-S/14)
DINOv2 (ViT-B/14)
DINOv2 (ViT-L/14)
% from Max Accuracy
% of Masked Tokens
TokenRank
Column Sum
Rand. Token
CLS Token
Center Token

Conclusions

- The authors interpret the attention matrix as a discrete-time Markov chain.
- TokenRank, i.e. the steady state vector of the Markov chain, measures global token importance.
- Taking into account higher-order interactions between tokens results in better performance on various downstream tasks.

Thank you for your **attention**!

SOURCES:

1. Y. Erel et al, Attention (as Discrete-Time Markov) Chains, The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
2. S. Ross, Introduction to Probability Models, 10th Edition, Elsevier, 2010.