УДК 004.89:004.93

ТОКЕНІЗАЦІЯ ЯК ЗАСІБ ТРАНСФОРМАЦІЇ ІНФОРМАЦІЇ

Оченашко М.О.
e-mail: maksym.ochenashko@nure.ua
Науковий керівник– д.т.н., проф., Гороховатський В.О.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

Tokenization is a critical mathematical transformation function that bridges computational linguistics and data protection. By developing rigorous models for tokenization, researchers quantify the degree of semantic preservation and computational efficiency in large language models or privacy technologies. This research explores a comprehensive theoretical framework for understanding tokenization as a fundamental information processing mechanism with cross-domain applications. By investigating various tokenization techniques, researchers aim to enhance both the accuracy and efficiency of computational models while addressing security and privacy concerns in diverse technological fields.

Токенізація – фундаментально важлива функція математичного перетворення, що поєднує апарат обчислювальної лінгвістики та захисту даних в системах штучного інтелекту [1, 2].

На підставі інформаційно-теоретичного аналізу та формального системного моделювання можна продемонструвати, як різні схеми токенізації оптимізуються для конкуруючих цілей, зберігаючи при цьому основні властивості даних. Ця формалізація забезпечує теоретичну основу для впровадження токенізації в обчислювальну лінгвістику та обробку даних із збереженням конфіденційності.

Токенізація здійснює фундаментальні перетворення в інформаційних системах, яка відображає послідовності необроблених даних в дискретні обчислювальні одиниці. Хоча застосування різняться в різних галузях, основна математична структура залишається незмінною: функція відображення, яка перетворює вхідний простір у простір токенів, зберігаючи при цьому специфічні властивості інформації [2].

Визначимо токенізацію формально як функцію

$$T: S \to U^*,$$

яка відображає елементи з вихідного простору S у послідовності у просторі токенів U^*, а простір U представляє словник можливих токенів. Функція T може бути далі розкладена на композицію підфункцій: $T = C \circ B \circ A$, де $A: S \to V$ представляє сегментацію, $B: V \to W$ представляє нормалізацію, а $C: W \to U^*$ представляє остаточне відображення у простір токенів.

Властивість токенізації зберігати інформацію може бути кількісно оцінена через умовну ентропію $H(S|T(S))$, яка вимірює втрату інформації під

час перетворення. Ідеальна токенізація мінімізує значення умовної ентропії, зберігаючи при цьому параметр керованої кардинальності простору токенів $|U|$.

З системної точки зору, токенізація у великих мовних моделях LLM являє собою підсистему попередньої обробки, яка перетворює неструктурований текст у структуровані обчислювальні одиниці. Ця підсистема повинна оптимізуватися для досягнення декількох конкуруючих цілей: мінімізації розміру словника, максимізації збереження семантики та підтримки обчислювальної ефективності.

Кодування байт-пар (BPE) можна моделювати як ітераційну оптимізаційну задачу

$$arg\ max(x, y) \in V \sum_{i=1} \mid D \mid count(xy, Di),$$

де D – навчальний корпус, а count(xy, D_i) – кількість входжень сусідніх лексем x та y у документі D_i [1, 2]. Цей жадібний алгоритм створює марковську модель, що наближає розподіл мови до базового розподілу.

Ефективність системи можна виміряти коефіцієнтом стиснення $\rho = \frac{|S|}{|T(S)|}$, що представляє середню кількість вихідних символів, закодованих на токен. Вищі значення вказують на більш ефективне кодування, але потенційно меншу вірність вихідного матеріалу.

У контексті захисту даних токенізація повинна задовольняти додаткові обмеження, пов'язані з інформаційною безпекою. Функція токенізації T повинна бути стійкою до попереднього зображення, що робить обчислювально неможливим отримання вихідних даних без доступу до захищеної функції відображення. Щоб система токенізації задовольняла вимогам GDPR щодо псевдонімізації [1], вона повинна демонструвати такі властивості як стійкість до колізій та незалежність від контексту.

Стійкість до колізій можна визначити як $P(T(s_1) = T(s_2)|s_1 \neq s_2) < \epsilon$ для деякого незначного ϵ, а незалежність від контексту формалізувати як $P(T(s) = t|C) \approx P(T(s) = t)$ для будь-якого контексту C. Форматно-зберігаюча токенізація вводить додаткові обмеження на простір виводу, вимагаючи, щоб $T: S \to S'$, де S' має спільні структурні властивості з S. Це можна формалізувати за допомогою регулярних мов, де і S, і S'описуються одним і тим же регулярним виразом [2].

У LLM підсистема токенізації створює інформаційне вузьке місце, яке змушує зменшувати розмірність вхідного простору. Пропускна здатність цього каналу може бути кількісно оцінена за допомогою взаємної інформації $I(S; T(S))$. Для системи з розміром словника $|U|$ і максимальною довжиною послідовності n, теоретична верхня межа інформації, яку можна передати, становить $n \log_2 |U|$ біт. Однак, реальна ефективна ємність, як правило, нижча через семантичні та статистичні обмеження.

У системах, що зберігають конфіденційність, токенізація створює навмисну інформаційну асиметрію між авторизованими і неавторизованими учасниками системи. Ця асиметрія може бути змодельована за допомогою матриць контролю доступу та обмежень інформаційних потоків, коли конфіденційні компоненти даних систематично ізолюються від загальних потоків обробки даних.

Підхід системного аналізу показує, що токенізація функціонує як критичний інтерфейс між неструктурованими даними та обчислювальною обробкою. У системах комп'ютерного зору принципи токенізації використовуються для формування результативних просторів ознак [3-5].

У технологіях LLM токенізація створює структурований простір представлення, який дозволяє нейронним мережам обробляти природну мову. У застосунках, що зберігають конфіденційність, токенізація встановлює межі безпеки, які відокремлюють конфіденційну інформацію, зберігаючи при цьому корисність даних. Майбутні дослідження націлені на вивчення адаптивних систем токенізації, які динамічно налаштовують свої властивості на основі вхідних характеристик і системних вимог. Такі системи потенційно можуть оптимізувати функцію токенізації для конкретних доменів, мов або обмежень конфіденційності, зберігаючи при цьому математичні гарантії збереження інформації і властивості безпеки.

Список використаних джерел:

1. Data Sharing Under the General Data Protection Regulation / A. Vlahou et al. Hypertension. 2021. Vol. 77, no. 4. P. 1029–1035. URL: https://doi.org/10.1161/hypertensionaha.120.16340 (дата звернення: 01.03.2025).

2. Radford, A, (2019), Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report*.

3. Gorokhovatskyi, V., Tvoroshenko, I., Yakovleva, O. (2024) Transforming image descriptions as a set of descriptors to construct classification features, Indonesian Journal of Electrical Engineering and Computer Science, 33 (1), 113-125.

4. Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., and Zeghid M. (2024) Improving the effectiveness of image classification structural methods by compressing the description according to the information content criterion, Computers, Materials & Continua, vol. 80, no. 2, pp. 3085-3106.

5. Gorokhovatskyi V., Tvoroshenko I., Yakovleva O., Hudáková M., and Gorokhovatskyi O. (2024) Application a committee of Kohonen neural networks to training of image classifier based on description of descriptors set, IEEE Access, vol. 12, pp. 73376-73385.