Mechanistical interpretability

Michał Burzyński

Celem tych zadań jest zachęcenie Was do ogarnięcia podstaw działania transformerów i kilku innych narzędzi, z których będziemy dużo korzystali podczas zajęć. Na początku warsztatów opiszę dokładnie jak one działają, jednak dobrze by było mieć już jakąś intuicję (szczególnie jeśli ktoś nigdy z nimi nie pracował). Żeby dostać się na warsztaty, **nie trzeba rozwiązać wszystkich zadań** — choć im więcej, tym lepiej! :)

W razie pytań, sugestii czy innych spraw, śmiało piszcie na maila: michal28burzynski@gmail

1 Wprowadzenie

Zadanie 1.1 (2 pkt)**.** Napisz kilka słów o sobie: dlaczego wybrałeś/aś akurat te zajęcia? Czy miałeś/aś już styczność z transformerami lub ogólnie pojętym uczeniem maszynowym? Czy pojęcie *mechanistic interpretability* jest Ci znane?

2 Podstawy Architektury Transformera

Zadanie 2.1 (3 pkt)**.** Przeczytaj artykuł https://jalammar.github.io/illustrated-transformer/ i opisz własnymi słowami, jak działają transformery. Nie musisz wchodzić w głębokie szczegóły matematyczne, ale zwróć szczególną uwagę na rolę mechanizmu attention.

Zadanie 2.2 (2 pkt)**.** Dlaczego w architekturze transformera konieczne jest użycie tzw. *Positional Encoding*? Co by się stało z modelem i jego rozumieniem tekstu, gdybyśmy całkowicie pominęli ten krok?

Zadanie 2.3 (2 pkt)**.** Czym różni się zwykły mechanizm *Self-Attention* (używany w enkoderze) od *Masked Self-Attention* (używanego w dekoderze)? Dlaczego to maskowanie jest absolutnie niezbędne podczas generowania tekstu?

Zadanie 2.4 (2 pkt)**.** Zamiast jednej dużej operacji uwagi, transformery używają mechanizmu *Multi-Head Attention*. Jaką konkretną przewagę daje modelowi podzielenie tej operacji na wiele niezależnych "głów"?

Zadanie 2.5 (3 pkt)**.** We wzorze na *Scaled Dot-Product Attention*: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ pojawia się dzielenie przez pierwiastek z wymiaru klucza ($\sqrt{d_k}$). Dlaczego to skalowanie jest ważne dla poprawnego działania i stabilności podczas treningu?

Zadanie 2.6 (2 pkt)**.** Skoro mechanizm *Self-Attention* odpowiada za komunikację i znajdowanie relacji między tokenami w sekwencji, to jaka jest główna rola warstw *Feed-Forward* (MLP), które znajdują się bezpośrednio po każdej warstwie attention?

3 Mechanistyczna Interpretowalność

Zadanie 3.1 (5 pkt)**.** Przeczytaj wstęp i wczesne rozdziały artykułu https://transformer-circuits.pub/2021/framework/index.html (nie musisz czytać całości, chyba że głębiej Cię to interesuje) i odpowiedz zwięźle na poniższe pytania:

1. Artykuł opisuje *residual stream* jako "shared memory bandwidth". Czym dokładnie jest ten strumień i w jaki sposób attention heads oraz warstwy MLP komunikują się za jego pomocą?
2. Czym różnią się od siebie obwody QK oraz OV? Za co odpowiada każdy z nich?
3. W jaki sposób poszczególne attention heads w tej samej warstwie wchodzą ze sobą w interakcję?
4. Dlaczego autorzy artykułu, budując swój framework, na początku analizują modele attention-only, całkowicie pomijając warstwy MLP?
5. Autorzy analizują modele o rosnącej złożoności. Czego fundamentalnie NIE potrafi zrobić model z tylko jedną warstwą uwagi (1-layer attention-only), co z powodzeniem realizuje model dwuwarstwowy?