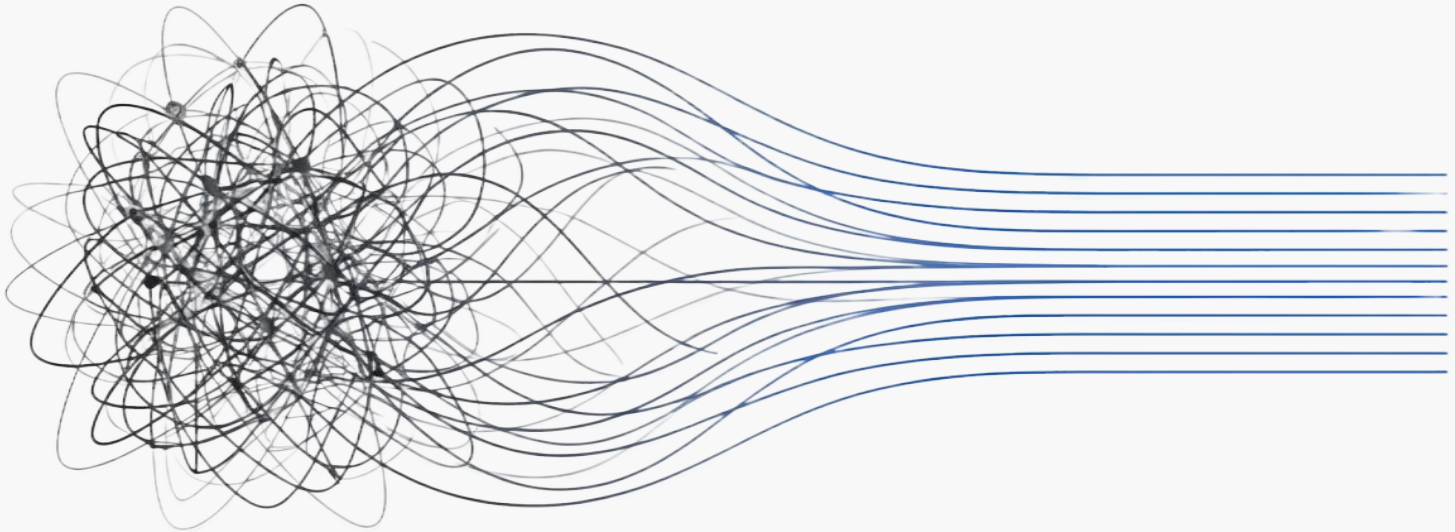


LightGCN

Jak mniej znaczy więcej w rekomendacjach grafowych?

Problematyka

Klasyczne sieci neuronowe na grafach (NGCF) były **bardzo skomplikowane.**

Sieci te miały mnóstwo skomplikowanych operacji takich jak

- 💠 nieliniowości,
- 💠 transformacje cech,

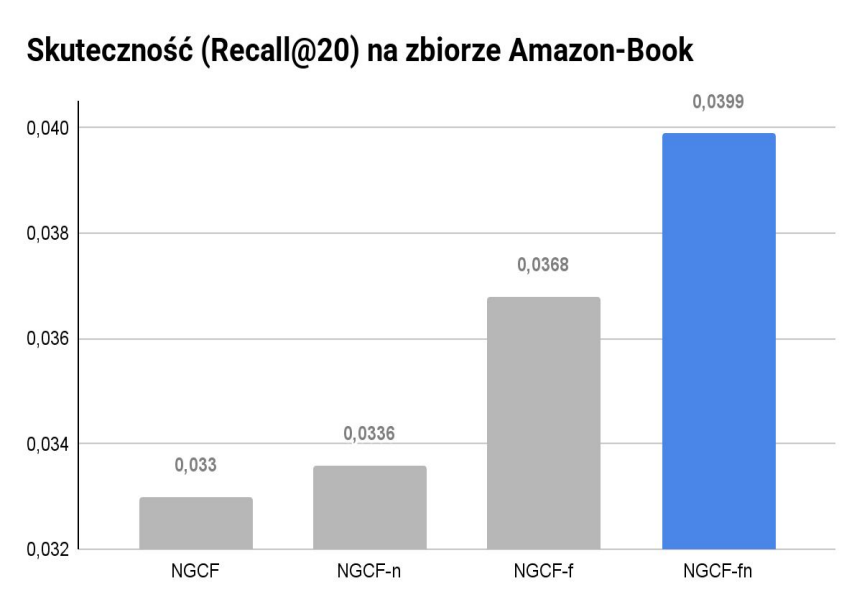
które są potrzebne przy rozpoznawaniu obrazów, **ale odkryto, że...**

Krytyczne pytanie

Czy transformacja cech i nieliniowa aktywacja są niezbędne w systemach rekomendacyjnych?

Zaskakujące odkrycie

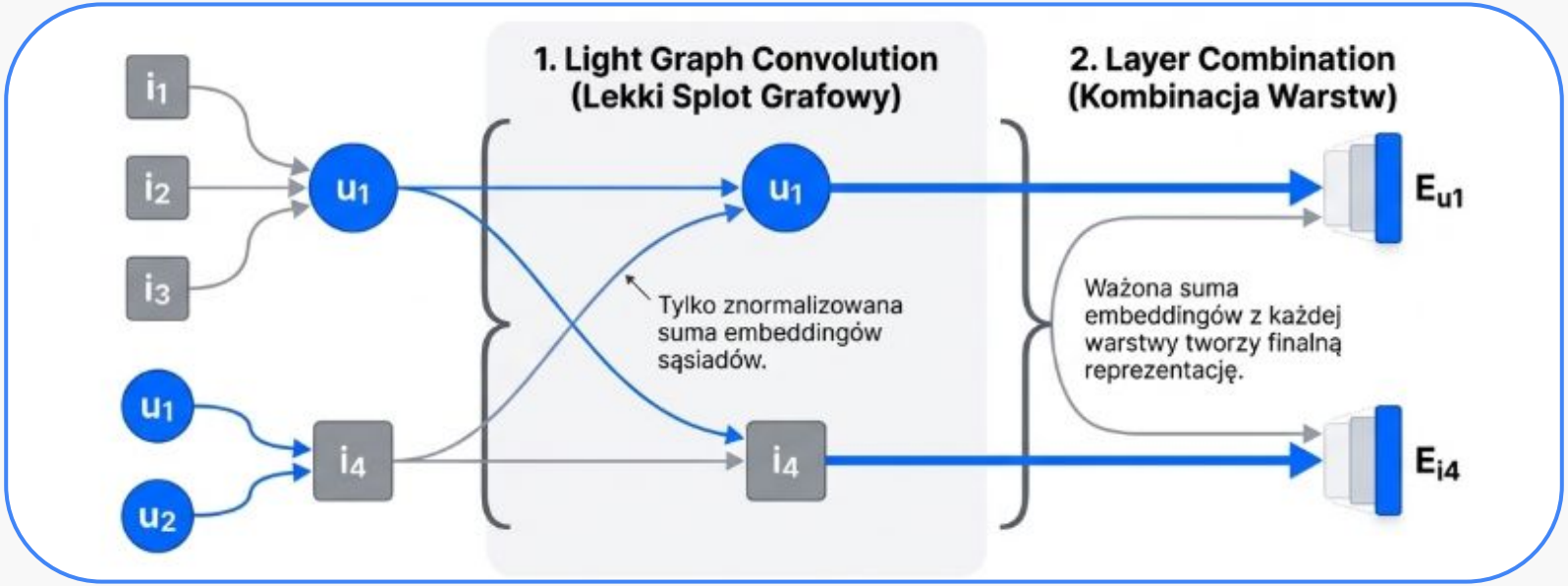
Przeprowadzono badanie ablacyjne, w którym usunięto z NGCF transformację cech (-f), nieliniowość (-n) oraz obie operacje (-fn). Wyniki przeczą intuicji.



Model NGCF-fn pozbawiony obu "ciężkich" operacji osiąga o 20,91% lepszy wynik Recall@20 na zbiorze Amazon-Book. Złożoność nie tylko nie pomaga, ale aktywnie szkodzi.

Rozwiązanie - LightGCN

W oparciu o te odkrycia powstał LightGCN. Model ten porzuca zbędne operacje i skupia się na jednej kluczowej idei GCN dla rekomendacji: agregacji cech sąsiadów w celu “wygładzenia” embeddingów.



Implementacja

Środowisko uruchomieniowe

1. Platforma i Sprzęt:

- **Środowisko:** Google Colab (Cloud Jupyter Notebooks).
- **Akceleracja:** GPU NVIDIA Tesla T4 (wymagane dla efektywnego treningu GNN).
- **Język:** Python 3.x.

2. Stos Technologiczny:

- **Model LightGCN:** Własna implementacja „od zera" w **TensorFlow/Keras**.
- **Model NGCF:** Implementacja oparta na frameworku **RecBole** (backend **PyTorch**).

3. Gwarancja powtarzalności:

- **Determinizm:** Sztywne ustawienie ziarna losowości (**SEED = 2020**) dla bibliotek **numpy**, **random**, **tensorflow** i **torch**.
- **Spójność danych:** Zaimplementowano identyczny algorytm doboru próby (Sparse/Dense) w obu notatnikach, eliminując wpływ losowego doboru danych na wyniki porównania.

Zbiór danych – Amazon-Book

1. Źródło i Charakterystyka:

- **Zbiór:** Amazon-Book.
- **Typ danych:** Implicit Feedback (binarny). Modelujemy sam fakt wystąpienia interakcji (zakup/kliknięcie), ignorując oceny w gwiazdkach.

2. Przetwarzanie:

- **Filtrowanie:** Usunięto użytkowników i przedmioty z mniej niż 5 interakcjami, aby zapewnić minimalną jakość sygnału.
- **Strategia doboru:** Zastosowano strategię Sparse (Losową). Wybrano losową próbkę użytkowników, aby zachować naturalną rzadkość danych i uniknąć sztucznego zagęszczenia (tzw. biasu "power users").

3. Statystyki Podzbioru:

- **Liczba użytkowników:** 20 000 (połowa wszystkich użytkowników z Amazon-Book).
- **Liczba przedmiotów:** ~90000.
- **Liczba interakcji**: ~900000.
- **Podział (Data Split):** 80% Trening / 10% Walidacja / 10% Test.

Krytyczne pytanie

Czy LightGCN uzyskał lepsze rezultaty względem NGCF?

Dowód - porównanie wyników

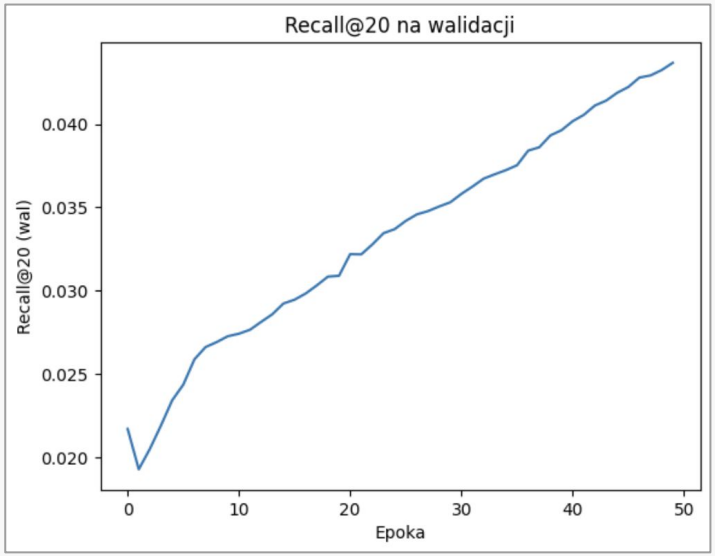
W bezpośrednim porównaniu na tych samych zbiorach danych i przy tej samej liczbie warstw, LightGCN osiąga znacznie lepsze wyniki na wielu płaszczyznach.

NGCF vs LightGCN

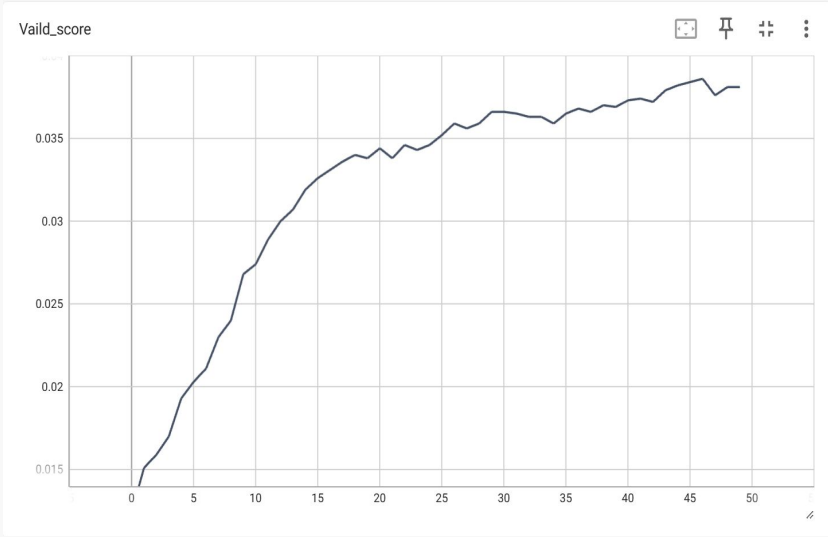
Kategoria	NGCF	LightGCN	Względna poprawa
Czas trenowania sieci	~150min	**~75min**	**-50%**
Recall@20	0.0361	**0.0431**	**+19,3%**

Dowód - porównanie wyników cd

Porównanie skuteczności (Recall@20) na zbiorze Amazon-Book



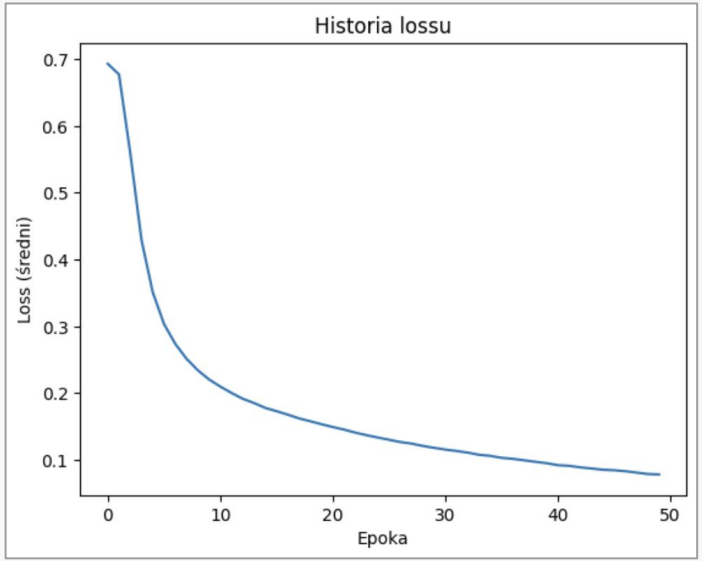
LightGCN



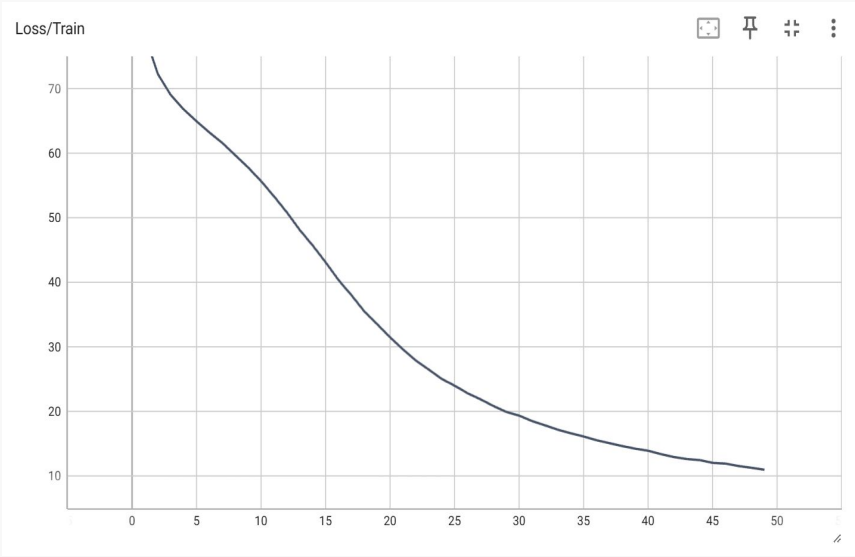
GNCF

Dowód - porównanie wyników cd

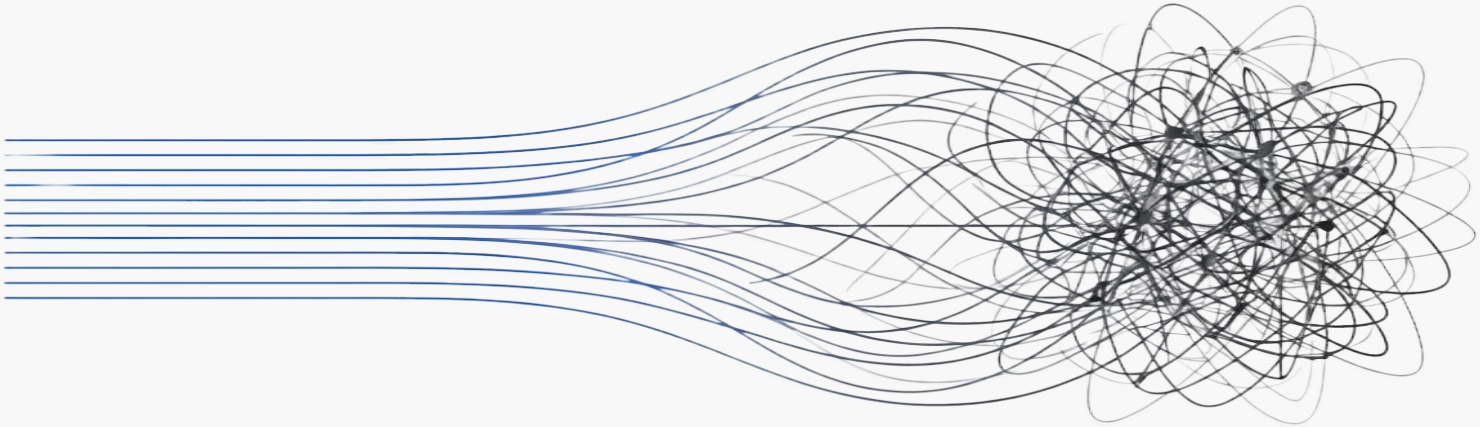
Porównanie wykresu funkcji straty (BPR) na zbiorze Amazon-Book



LightGCN



GNCF



Prosty model > **Skomplikowany model**

Wierne odtworzenie publikacji naukowej

Niestety, jednym z największych problemów współczesnej nauki o danych jest "Kryzys Reprodukowalności" (The Reproducibility Crisis).

"Piekło Zależności": Różne wersje bibliotek (np. TensorFlow 2.x vs 1.x, PyTorch) mają inne domyślne ustawienia i implementacje matematyczne.

Ukryte Detale ("Secret Sauce"): Publikacje naukowe często pomijają szczegóły inżynierskie, t.j. dokładna metoda preprocessingu danych , sposób losowania negatywów czy strategia tworzenia batchy.

Loteria Inicjalizacji: Losowy dobór wag początkowych ma ogromny wpływ na to, czy model utknie w słabym minimum lokalnym, czy osiągnę świetny wynik.

Niespójność Metryk: Różne definicje tych samych miar (np. różne warianty wzoru na NDCG) lub stosowanie uproszczonej ewaluacji.

Dziękujemy za uwagę

Michał Tomaszewski
Damian Tomczyk
Szymon Wierzchoś