
Dynamiczne kodowanie słownikowe LZSS

Realizacja pierwszego etapu projektu

Domański Piotr, Pietrowcew Jakub, Skórka Kornel

2021-11-20

Analiza literatury

Podczas tego etapu dokonano zapoznania z opisem metod słownikowych w dwóch książkach:

- A. Przelaskowski, **“Kompresja danych. Podstawy, metody bezstratne, kodery obrazów”**
- K. Sayood, **“Kompresja danych – wprowadzenie”**

Najważniejsze zebrane przez nas informacje:

- Metody słownikowe polegają na zastępowaniu ciągu symboli zmiennej długości poprzez słowo kodowe zwane indeksem.
- Przeglądając słownik poszukujemy w nim frazy identycznej do sekwencji wejściowej, a indeks będący słowem kodowym wskazuje nam odpowiednią pozycję w słowniku.
- Efektywność kodowania jest tym większa, im dłuższa jest dopasowanie sekwencji oraz im krótszy jest indeks
- Dynamiczne kodowanie słownikowe LZSS jest modyfikacją metody LZ77. Metoda LZ77 opiera się na strukturze okna przesuwne. Okno składa się z większej części będącej słownikiem, który zawiera ostatnio zakodowane dane źródłowe. Drugą część okna stanowi bufor w którym umieszczane są dane źródłowe przeznaczone do zakodowania. Indeks w metodzie LZ77 stanowią: pozycja frazy w słowniku, długość znalezionej najdłuższej dopasowania frazy, pierwszy symbol występujący po kodowanym łańcuchu.
- Modyfikacje względem LZ77, które wprowadza LZSS to:
 - dwa rodzaje indeksów rozróżniane za pomocą pojedynczego bitu sygnalizującego. Rodzaj indeksu zależy od długości dopasowanej frazy. Przy czym warto nadmienić, że rozróżnienie na długie i krótkie frazy zależy od zastosowania (w tym od długości słownika i bufora).
 - Słownik w postaci drzewa binarnego – ułatwia to przeszukiwanie słownika, nie wpływa to na efektywność kompresji, ale przyspiesza działanie algorytmu.
- Na efektywność kompresji wpływ ma ustalony rozmiar słownika oraz bufora. Rozmiary te powinny być dostosowane do kodowanych danych. Podczas implementacji przetestowane zostaną różne rozmiary słownika i bufora tak aby uzyskać lepszą efektywność kompresji.

Wybór narzędzi programistycznych

Do stworzenia kodera oraz dekodera zamierzemy skorzystać z języka programowania C++, natomiast do generowania wizualizacji efektów, np. histogramu, zamierzamy wykorzystać język programowania Python - bibliotekę Matplotlib. Taki wybór narzędzi programistycznych pozwoli nam na efektywną implementację, wydajne kodowanie i dekodowanie oraz prostotę przy generacji wizualizacji efektów.

Wybór metod oceny efektywności

W celu oceny efektywności zaimplementowanego rozwiązania, dla każdego zestawu danych planujemy sprawdzić i przeanalizować następujące wielkości:

1. stosunek rozmiaru danych wejściowych do danych zakodowanych
2. czas kodowania
3. czas dekodowania
4. wartość przepływności