**Politechnika Rzeszowska im. Ignacego Łukasiewicza**

Wydział Elektrotechniki i Informatyki

KATEDRA

POLITECHNIKA
RZESZOWSKA
im. IGNACEGO ŁUKASIEWICZA

PRACA INŻYNIERSKA

DAWID KOWALCZUK

# PRZYGOTOWANIE PRACY DYPLOMOWEJ W SYSTEMIE LaTeX

PROMOTOR:
dr hab. inż. Maciej Kusy, prof. PRz

Rzeszów 2026

## OŚWIADCZENIE AUTORA PRACY

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie, i nie korzystałem ze źródeł innych niż wymienione w pracy.

.....

PODPIS

**Rzeszów University of Technology**

Faculty of Electrical Engineering and Computer Science

DEPARTMENT OF

POLITECHNIKA
RZESZOWSKA
im. IGNACEGO ŁUKASIEWICZA

ENGINEERING THESIS

DAWID KOWALCZUK

# THESIS IN LaTeX

SUPERVISOR:
Assoc. Prof. D.Sc. Eng.
Maciej Kusy

Rzeszów 2026

Serdecznie dziękuję . . . tu ciąg dalszych podziękowań np. dla promotora, żony, sąsiada itp.

# Spis treści

# 1. Wstęp

LaTeXci([Dil00], [Lam92]).

## 1.1. Wprowadzenie do problematyki

Analiza danych nieustrukturyzowanych, w szczególności obrazów, stanowi istotny obszar rozwoju współczesnych systemów informatycznych. W sektorach takich jak medycyna, przemysł czy archiwizacja cyfrowa, dane wizualne stanowią coraz większą część przetwarzanych informacji. Przez ostatnią dekadę standardem w ich analizie były metody wizji komputerowej oparte na splotowych sieciach neuronowych. Choć rozwiązania te sprawdzają się w zadaniach klasyfikacji, okazują się niewystarczające w bardziej złożonych procesach wymagających głębokiego zrozumienia semantyki sceny oraz relacji między obiektami.

Istotną zmianę w podejściu do przetwarzania obrazu przyniosła adaptacja architektury Transformer, pierwotnie dedykowanej przetwarzaniu języka naturalnego. Umożliwiło to rozwój modeli multimodalnych, które potrafią przetwarzać sygnał wizualny i tekstowy we wspólnej przestrzeni wektorowej. Rozwiązania te, reprezentowane przez architektury takie jak BLIP czy LLaVA, pozwalają na automatyzację zadań eksperckich, w tym generowanie opisów radiologicznych na podstawie zdjęć RTG czy ekstrakcję danych z dokumentacji technicznej.

Mimo potencjału dużych modeli fundamentowych, ich wdrożenie w specyficznych środowiskach produkcyjnych wiąże się z ograniczeniami technicznymi i prawnymi. Modele trenowane na ogólnych zbiorach danych często generują opisy nieprecyzyjne merytorycznie w dziedzinach wąskospecjalistycznych, takich jak diagnostyka obrazowa czy botanika. Dodatkowym wyzwaniem jest konieczność uzyskania danych wyjściowych w ustrukturyzowanych formatach, takich jak JSON lub CSV, co jest niezbędne do integracji systemów sztucznej inteligencji z bazami danych. Istotną barierą są również regulacje dotyczące prywatności, które często wymuszają przetwarzanie danych lokalnie, na sprzęcie o ograniczonych zasobach pamięciowych, co wyklucza użycie zewnętrznych interfejsów API.

W tym kontekście kluczowym zagadnieniem inżynierskim staje się opracowanie efektywnych metod adaptacji istniejących modeli otwartoźródłowych do specyficznych wymagań domenowych.

## 1.2. Cel pracy i pytania badawcze:

Głównym celem pracy jest zbadanie efektywności adaptacji modeli typu Vision Transformer do realizacji zaawansowanych zadań generowania opisów w domenach specjalistycznych. Praca ma charakter badawczo-wdrożeniowy i koncentruje się na empirycznej weryfikacji wpływu różnych strategii uczenia maszynowego na zdolność modelu do przyswajania wiedzy eksperckiej oraz formalnej.

Cel ten zostanie zrealizowany poprzez następujące działania szczegółowe:

– **Analiza porównawcza architektur:** Zbadanie wpływu rozmiaru oraz budowy modelu na jakość adaptacji do nowych zadań przy ograniczonym zbiorze treningowym.

– **Ewaluacja metod treningowych:** Porównanie klasycznego pełnego dostrajania, zwanego Full Fine-Tuning, z nowoczesnymi metodami efektywnymi parametrowo, takimi jak LoRA i QLoRA. Celem jest ustalenie, czy metody redukujące zapotrzebowanie na pamięć VRAM wiążą się z degradacją zdolności dyskryminacyjnych modelu w zadaniach wymagających wysokiej precyzji.

– **Weryfikacja zdolności strukturalnych:** Zbadanie możliwości nauczenia modelu wizyjnego roli parsera, czyli generowania poprawnych składniowo plików JSON bezpośrednio z obrazu, co stanowi alternatywę dla klasycznych potoków OCR.

– **Optymalizacja wydajności:** Określenie kompromisu pomiędzy czasem treningu i zużyciem zasobów sprzętowych a jakością końcową modelu.

## 1.3. Zakres pracy:

Praca obejmuje spektrum działań inżynierskich, począwszy od przygotowania danych, poprzez eksperymenty uczenia maszynowego, aż po wdrożenie rozwiązania w formie aplikacji.

Warstwa teoretyczna zawiera przegląd literatury dotyczącej ewolucji modeli językowo-wizyjnych, ze szczególnym uwzględnieniem zjawisk takich jak zapaść modalności wizyjnej, znana w literaturze jako Vision Token Collapse, oraz metod przeciwdziałania katastrofalnemu zapominaniu wiedzy.

W części badawczej wykorzystane zostaną wybrane modele dostępne na licencji Open Source. Eksperymenty zostaną przeprowadzone na autorskich lub specjalnie przygotowanych podzbiorach danych, reprezentujących dwa odmienne wyzwania: domenę medyczną lub przyrodniczą, gdzie kluczowa jest precyzja wizualna, oraz domenę inżynierską, wymagającą zachowania struktury logicznej danych wyjściowych. W ramach procedury badawczej zaimplementowane i przetestowane zostaną różne konfiguracje treningowe, uwzględniające techniki kwantyzacji oraz mieszanej precyzji obliczeń.

Zakres pracy w warstwie aplikacyjnej obejmuje zaprojektowanie i implementację prototypu systemu w architekturze klient-serwer. Część backendowa, oparta na języku Python i bibliotece PyTorch, będzie odpowiedzialna za inferencję modeli i obsługę żądań. Część frontendowa dostarczy interfejs graficzny umożliwiający użytkownikowi końcowemu interakcję z systemem, wybór modelu oraz wizualizację wyników w formie tekstowej lub ustrukturyzowanej.

# 2. Podstawy teoretyczne (Przegląd literatury)

## 2.1. Architektury multimodalne (Vision-Language Model)

Tradycyjne podejście do uczenia maszynowego traktowało analizę obrazu i przetwarzanie tekstu jako dwa odrębne problemy inżynieryjne, rozwiązywane za pomocą niekompatybilnych architektur. Systemy wizyjne, oparte głównie na splotowych sieciach neuronowych, specjalizowały się w ekstrahowaniu cech przestrzennych z macierzy pikseli w celu klasyfikacji obiektów (np. przypisania etykiety "samochód"). Z kolei systemy językowe skupiały się na analizie sekwencyjnej i modelowaniu gramatyki. Takie rozdzielenie uniemożliwiało tworzenie systemów zdolnych do pełnego zrozumienia sceny, w której treść wizualna jest nierozerwalnie związana z opisem semantycznym. Klasyfikacja obrazu daje jedynie zbiór etykiet, natomiast dopiero wygenerowanie opisu w języku naturalnym pozwala na uchwycenie relacji, akcji i atrybutów obiektów widocznych na zdjęciu.

Rozwiązaniem tego problemu są modele multimodalne typu Vision-Language (VLM). Ich zadaniem jest integracja danych o skrajnie różnej strukturze: ciągłego sygnału wizualnego (wartości intensywności pikseli) oraz dyskretnego sygnału tekstowego (tokeny słownikowe). Wyzwanie inżynieryjne polega tutaj na stworzeniu mechanizmu, który pozwoli komputerowi "zrozumieć", że matematyczna reprezentacja obrazu psa jest tożsama z matematyczną reprezentacją słowa "pies". Modele VLM realizują to poprzez rzutowanie obu typów danych do wspólnej, wielowymiarowej przestrzeni wektorowej. W tej przestrzeni wektory reprezentujące obraz i odpowiadający mu tekst znajdują się blisko siebie, co pozwala na wykonywanie operacji logicznych łączących wzrok z językiem, takich jak generowanie podpisów czy wyszukiwanie obrazów za pomocą zapytań tekstowych.

### 2.1.1. Ewolucja paradygmatów: od CNN-RNN do Transformerów

Do momentu spopularyzowania architektury Transformer (ok. 2020 roku), domyślnym standardem w zadaniach Image Captioning były architektury hybrydowe, łączące splotowe sieci neuronowe (CNN) z rekurencyjnymi sieciami neuronowymi (RNN). Podejście to opierało się na paradygmacie Encoder-Decoder, w którym dwie odrębne sieci współpracowały ze sobą w sekwencyjnym potoku przetwarzania.

**CNN jako enkoder obrazu**

**RNN jako dekoder tekstu**

# 3. Metodyka badań i narzędzia

# 4. Eksperymenty i analiza wyników

# 5. Projekt i implementacja aplikacji

# 6. Dyskusja i Podsumowanie

# Bibliografia

[Dil00] A. Diller. *LaTeX wiersz po wierszu*. Wydawnictwo Helion, Gliwice, 2000.

[Lam92] L. Lamport. *LaTeX system przygotowywania dokumentów*. Wydawnictwo Ariel, Krakow, 1992.