Uniwersytet Mikołaja Kopernika w Toruniu

Wydział Matematyki i Informatyki

Szymon Nowacki

nr albumu: 316259

informatyka (n1 – studia niestacjonarne)

Praca inżynierska

# Porównawcza analiza architektur głębokiego uczenia w rozpoznawaniu emocji na podstawie obrazów twarzy

Promotor

Dr Marta Burzańska

Toruń 2024

# Spis treści



# 1. Wstęp

## 1.1 Wprowadzenie

Sztuczna inteligencja w ciągu ostatnich lat zyskuje coraz większe znaczenie jako dziedzina nauki i technologii. Jej wpływ obejmuje coraz szersze aspekty życia, lecz na szczególną uwagę zasługują te z nich, które intuicyjnie wydawały się nieosiągalne dla algorytmów i komputerów. Jednym z takich zastosowań jest analiza ludzkich emocji na podstawie samych tylko obrazów twarzy. Innowacje osiągnięte w dziedzinie masowego i błyskawicznego wykrywania stanów emocjonalnych bez potrzeby ingerencji innego człowieka mogą doprowadzić do usprawnień w wielu dziedzinach życia opartych na osobistych odczuciach.

Obecnie, detekcja emocji znajduje zastosowanie w wielu dziedzinach. W marketingu pozwala na personalizację kampanii reklamowych, znacząco poszerzając zakres informacji zwrotnych jakie reklamodawca może otrzymać od odbiorców. W opiece zdrowotnej obecność takiej technologii może umożliwić dokładniejszą diagnozę i terapię zaburzeń emocjonalnych. Dostęp do tak ludzkiego aspektu życia ma możliwość poskutkować znacznym postępem w relacjach człowiek-komputer, dając asystentom wirtualnym dostęp do całkowicie nowego sposobu na rozumienie swoich użytkowników.

### 1.1.1 Podstawy teoretyczne

#### *Emocje i ich cechy charakterystyczne*

Emocje to reakcje psychofizyczne mające odzwierciedlać ludzkie przeżycia, są podstawą interakcji międzyludzkich. Każda z nich wywołuje specyficzne zmiany w ciele, zachowaniu oraz co najważniejsze w ekspresji mimicznej twarzy. Można je charakteryzować na podstawie intensywności, trwania oraz kontekstu pojawienia się (1). Te cechy umożliwiają klasyfikację emocji, istotny fakt zarówno w psychologii, jak i w podejściach technologicznych.

#### *Ekspresje emocji*

Ekspresje emocji to uniwersalny język, którym ludzie komunikują wewnętrzne stany emocjonalne. Paul Ekman zidentyfikował 6 podstawowych emocji, które są powszechne we wszystkich odkrytych kulturach świata. Każdą z nich można scharakteryzować przez wzorce ruchów mięśni twarzy (2):

- **Radość:** Wyrażana przez uniesienie kącików ust, często towarzyszy jej uśmiech oraz zmarszczenie skóry wokół oczu.
- **Smutek**: Przejawia się opuszczenie kącików ust i lekkim uniesieniem wewnętrznych krawędzi brwi. Oczy mogą być lekko zwężone albo zamglone.
- **Złość**: Objawia się ściągniętymi brwiami, co powoduje pojawienie się między nimi zmarszczek. Pojawia się napięcie wokół ust, mogą być zaciśnięte lub przypominać cienką linię.
- **Strach**: Rozpoznawany przez uniesienie brwi, szeroko otwarte oczy oraz otwarte usta.
- **Zaskoczenie**: Charakteryzuje się uniesionymi brwiami, które tworzą zmarszczki na czole, szeroko otwartymi oczami. Usta często tworzą kształt litery „O”.
- **Wstręt**: Wyrażany przez uniesienie górnej wargi, marszczeniem nosa i opuszczeniem brwi.

### *Facial Emotion Recognition (FER)*

Rozpoznawanie emocji na podstawie mimiki twarzy to technologia, która potrafi klasyfikować emocje z pomocą uczenia maszynowego i wizji komputerowej (Computer Vision, CV). Proces obejmuje detekcję twarzy, ekstrakcję cech, klasyfikację i opcjonalnie interpretację wyniku.

### *Sieci Neuronowe*

Sieci neuronowe to struktury obliczeniowe inspirowane biologicznym mózgiem będące podstawą współczesnych algorytmów sztucznej inteligencji. Składają się z warstw sztucznych neuronów, na których strukturę i działanie składa się:

- **Wejście**: reprezentowanego jako wektor liczb
- **Wagi**: każde z wejść posiada swoją, jest to miara znaczenia tego połączenia dla podejmowanej decyzji. Wagi mogą być dodatnie, ujemne lub zerowe. Są dostosowywane podczas procesu uczenia, a ich modyfikacja jest środkiem do minimalizowania błędu pomiędzy oczekiwanym a rzeczywistym wyjściem neuronu.
- **Suma ważona**: neuron oblicza sumę ważoną swoich wejść. Wynik tej operacji nazywa się aktywacją neuronu.
- **Funkcje aktywacji**: Służy do przetworzenia wartości aktywacji neuronu, decyduje o tym jaki sygnał zostanie przekazany dalej.
- **Wyjście**: wartość, która jest rezultatem zastosowania funkcji aktywacji do sumy ważonej.

Każdy neuron jest połączony z innymi w warstwie poprzedniej i następnej z pomocą wag. Celem sieci jest zminimalizowanie różnicy między przewidywaniem modelu a rzeczywistym wynikiem. Sieci neuronowe adaptują różne architektury, od jedno lub dwuwarstwowych perceptronów do Konwolucyjnych sieci zaprojektowanych do przetwarzania obrazów, mowy czy pisma. Te struktury są szczególnie skuteczne w wyszukiwaniu wzorców w dużych zbiorach danych (3).

*Głębokie uczenie*

Głębokie uczenie, będące poddziedziną uczenia maszynowego, opiera się na użyciu wielowarstwowych sieci neuronowych. Głównym usprawnieniem w stosunku do tradycyjnych algorytmów jest automatyczna identyfikacja szukanych wzorców na różnych poziomach abstrakcji. W procesie propagacji wstecznej wagi należące do poszczególnych neuronów są aktualizowane tak, by minimalizować błąd pomiędzy oczekiwanym a rzeczywistym (dostarczonym przez dane treningowe) wynikiem.

## 1.2 Cel pracy

Celem niniejszej pracy jest zaprojektowanie modeli AI zdolnych do wykrywania emocji na podstawie analiz obrazów twarzy. Podczas przeglądu istniejących rozwiązań zostaną wyłonione rodzaje architektur i poszczególne implementacje sieci neuronowych zdolne do klasyfikacji obrazów. W ramach pracy zostaną zaprojektowane i zaimplementowane różne modele, a następnie przeprowadzone testy mające na celu porównanie ich skuteczności. Analiza wyników pozwoli na ocenę efektywności poszczególnych rozwiązań w kontekście rozpoznawania emocji na podstawie obrazów twarzy. Niezbędnym elementem pracy będzie także dobór i ewentualna augmentacja zbioru danych treningowych i testowych.

## 1.3 Zakres pracy

Niniejsza praca inżynierska koncentruje się na analizie, zaprojektowaniu, implementacji i ewaluacji modelu sztucznej inteligencji zdolnego do rozpoznania 6 podstawowych emocji: radości, smutku, strachu, złości, zaskoczenia i wstrętu (4) na podstawie obrazów twarzy. Projekt ten koncentruje się na:

1. **Przegląd literatury i standardów rynkowych:** W pierwszej kolejności zostanie przeprowadzony przegląd badań, analiz oraz rozwiązań w dziedzinie rozpoznawania emocji za pomocą sztucznej inteligencji. Ta część ma na celu identyfikację zalet i wad stosowanych obecnie podejść oraz wyłonienie architektur przystosowanych do analizy obrazów.
2. **Projektowanie modeli:** Z użyciem przyswojonych wcześniej podstaw teoretycznych zostaną zaprojektowane nowe modele. Dla każdego z rozwiązań na drugą część opracowania złożą się: wybór odpowiedniej architektury sieci neuronowej, definicji warstw, algorytmów optymalizacji i funkcji aktywacji. Kandydaci na docelowe rozwiązanie będą trenowani na małym zestawie danych w celu wyłonienia najbardziej obiecujących kombinacji cech modelu. Każdy znaczący wybór zostanie odpowiednio uzasadniony, a ciekawsze odrzucone przypadki odpowiednio udokumentowane. Na tym etapie nastąpi też selekcja i ewentualna modyfikacja zbioru danych służących do treningu i testowania sieci.
3. **Implementacja modeli:** Zaprojektowane modele zostaną stworzone, wytrenowane i wdrożone. Kluczowymi narzędziami, które pozwolą zaprojektować wytypowane rozwiązanie są: Python (główny język programowania), TensorFlow (główna biblioteka umożliwiająca tworzenie i trenowanie modeli głębokiego uczenia), Keras (wysokopoziomowe API do TensorFlow) i OpenCV (przetwarzanie i augmentacja obrazów, wykrywanie twarzy).

4. **Testowanie i analiza:** Zaimplementowane modele zostaną poddane szeregowi testów, które dostarczą informacje niezbędne do oceny ich skuteczności i wydajności. Podstawowymi miarami do oceny działania będą: dokładność (accuracy), precyzja (precision), czułość (recall), czas odpowiedzi i zużycie zasobów. W przypadku niewystarczalności analiz ilościowych (specyficzny wzorzec popełnianych błędów, trudne przypadki graniczne) dodatkowo zostaną przeprowadzone wszelkie potrzebne analizy jakościowe. Zwieńczeniem tego etapu prac będzie ustandaryzowany raport prezentujący wspomniane metryki.

## 1.4 Przegląd istniejących rozwiązań

### 1.4.1 Convolutional Neural Networks (CNN)

Najczęściej stosowana architektura w dziedzinie głębokiego uczenia, szczególnie w przetwarzaniu obrazów. Wyróżnia się zdolnością do automatycznego wyodrębniania wzorców i rozpoznawania wzorców przestrzennych częściowo odpornych na translację. Kluczowymi elementami w CNN są różnorodne warstwy, które odgrywają odmienne role w procesie analizy danych. Podstawową warstwą w CNN jest warstwa konwolucyjna, której zadaniem jest wykrywanie cech obrazu, takich jak krawędzie czy tekstury, poprzez przesuwanie filtrów (jąder) po obrazie. Operacja konwolucji dla jednego filtru $k$ opisuje się wzorem:

$$\mathrm{S(i,j)} = \sigma\left(\sum_m \sum_n I(i-m, j-n) \cdot K(m,n)\right)$$

Gdzie:

- $I$ to macierz pikseli wejściowego obrazu,
- $K$ to macierz filtru (jądra),
- $S(i,j)$ to wartość w mapie cech (feature map) w pozycji $(i,j)$
- σ to funkcja aktywacji, np. ReLU: $\sigma(\mathrm{x}) = \max(0, \mathrm{x})$

Wynikiem tego procesu jest tzw. mapa cech (feature map), która przedstawia, jak intensywnie dana cecha jest obecna w różnych częściach obrazu. Kolejnym istotnym elementem jest warstwa łącząca (pooling layer), która zazwyczaj następuje po warstwie konwolucyjnej. Ta warstwa nie posiada wag, a jej głównym zadaniem jest redukcja liczby parametrów wejściowych, co zmniejsza złożoność obliczeniową modelu oraz ryzyko przeuczenia (overfitting). Redukcja ta jest realizowana poprzez operacje takie jak maksymalne i średnie klastrowanie (max and average pooling), co prowadzi do celowej utraty części informacji, ale jednocześnie poprawia efektywność modelu (3)**Error! Bookmark not defined.**. Dla przykładu, maksymalne klastrowanie dla regionu $R$ opisuje się wzorem:

$$P(i,j) = \max_{(m,n)\in R} S(m,n)$$

Gdzie:

- $S(m,n)$ to wartości mapy cech w regionie $R$,

- $P(i,j)$ to wartość po poolingu w pozycji $(i,j)$.

Na końcu architektury CNN znajduje się warstwa w pełni połączona (fully connected layer). Jest to ostatnia warstwa, która integruje wyekstrahowane cechy w celu dokonania ostatecznej klasyfikacji. Warstwa ta łączy wszystkie neurony z poprzedniej warstwy z każdym neuronem w tej warstwie, co umożliwia podejmowanie decyzji na podstawie całościowego zestawu cech (5). W nowoczesnych architekturach CNN, takich jak ResNet (6) czy EfficientNet (7), coraz częściej stosuje się różne podejścia do zmniejszenia liczby parametrów przed końcową warstwą klasyfikacyjną, w tym globalny pooling. Zamiast używać tradycyjnej warstwy w pełni połączonej, która wiąże się z dużą liczbą parametrów, globalny pooling pozwala na wyciągnięcie kluczowych cech z ostatnich map cech, co redukuje złożoność modelu i poprawia jego ogólną wydajność. W miarę przechodzenia przez kolejne warstwy sieci, model staje się coraz bardziej skomplikowany i jest w stanie rozpoznawać coraz bardziej złożone wzorce i struktury. Warstwy CNN są zorganizowane hierarchicznie, gdzie każda kolejna warstwa jest odpowiedzialna za rozpoznawanie cech o wyższym poziomie abstrakcji. Pierwsze warstwy skupiają się na prostych cechach, takich jak krawędzie czy tekstury, podczas gdy kolejne mogą rozpoznawać bardziej skomplikowane struktury (7), aż do złożonych wzorców reprezentujących konkretne emocje na twarzy. Do wiodących implementacji opartych na CNN należą: VGG-Net (8) i Feature Decomposition and Reconstruction Learning (FDRL) (9).

### 1.4.2 Graph Neural Networks (GNN)

Grafowe Sieci Neuronowe (GNN) to klasa modeli zaprojektowanych do przetwarzania danych o strukturze grafu. W kontekście rozpoznawania emocji na podstawie twarzy (FER), GNN są szczególnie użyteczne do analizy przestrzennych relacji między punktami charakterystycznymi twarzy. Twarz może być modelowana jako graf, gdzie węzły reprezentują punkty charakterystyczne, takie jak kąciki ust, oczu, i brwi, a krawędzie odzwierciedlają relacje przestrzenne między nimi. GNN wykorzystują mechanizm propagacji informacji między węzłami, co pozwala na uchwycenie złożonych relacji i zależności między różnymi częściami twarzy (10).

Głównym mechanizmem działania GNN jest Message Passing, który opiera się na dwóch kluczowych etapach: agregacji informacji od sąsiadów i aktualizacji stanu węzła. Każdy węzeł wysyła informacje o swoim stanie do sąsiadujących węzłów. W każdej warstwie sieci, dla danego węzła $v$, najpierw agregowane są informacje z sąsiadujących węzłów $N(v)$ przy użyciu funkcji agregacji (10):

$$\mathrm{m}_v{}^{(k)} = \mathrm{AGGREGATE}^{(k)}\left(\left\{h_u{}^{(k-1)} : u \in N(v)\right\}\right)$$

Następnie stan węzła $v$ jest aktualizowany na podstawie zebranych informacji $m_v^{(k)}$ i poprzedniego stanu $h_v^{(k-1)}$ za pomocą funkcji aktualizacji **Error! Bookmark not defined.**:

$$\mathrm{h}_\mathrm{v}^{(\mathrm{k})} = \mathrm{UPDATE}^{(\mathrm{k})}\left(\mathrm{h}_\mathrm{v}^{(\mathrm{k}-1)}, \mathrm{m}_\mathrm{v}^{\mathrm{k}}\right)$$

Jednym z najbardziej popularnych podejść w GNN jest Graph Convolutional Network (GCN). W przeciwieństwie do standardowych sieci grafowych, w tej architekturze propagowane są całe wektory cech sąsiednich węzłów, co umożliwia trenowanie wielowymiarowych powiązań. Jedna warstwa w Graph Convolutional Network jest definiowana w następujący sposób (11):

$$\mathrm{H}^{(\mathrm{k})} = \sigma\left(\widetilde{D}^{-0.5}\widetilde{\mathrm{A}}\,\widetilde{D}^{-0.5}H^{(k-1)}W^{(k)}\right)$$

Gdzie:

- $\mathrm{H}^{(\mathrm{k})}$: Macierz reprezentacji węzłów po k-tej warstwie sieci.
- $W^k$: Macierz wag dla k-tej warstwy sieci.
- $\tilde{A}$: Zmodyfikowana macierz sąsiedztwa grafu, do której dodano połączenia własne dla każdego węzła. Definiowana jako $\tilde{A} = A + I$, gdzie $A$ to macierz sąsiedztwa a $I$ to macierz jednostkowa.
- $\widetilde{D}$: Diagonalna macierz stopni węzłów
- $\sigma$: Funkcja aktywacji

Do wiodących implementacji opartych na GNN należą: Graph Attention Networks (GAT) (12), Superpixel-based Graph Convolutional Network (SP-GCN) (13)

### 1.4.3 Transformer Networks

Architektura, która pierwotnie zrewolucjonizowała dziedzinę przetwarzania języka naturalnego (NLP), znalazła również zastosowanie w rozpoznawaniu emocji na podstawie twarzy (FER). Transformery, dzięki swojej zdolności do modelowania długoterminowych zależności i kontekstu, oferują znaczące korzyści w porównaniu z tradycyjnymi sieciami neuronowymi, takimi jak CNN, które są ograniczone przez swoje lokalne receptory. Jednym z najważniejszych elementów sieci Transformer jest mechanizm uwagi (attention). Jego znaczenie polega na zdolności do dynamicznego skupiania się na różnych częściach danych wejściowych, co umożliwia modelowi efektywne wychwytywanie istotnych informacji niezależnie od ich pozycji w sekwencji. W tradycyjnych modelach, takich jak sieci rekurencyjne (RNN), informacje przetwarzane są sekwencyjnie, co może prowadzić do problemów z uchwyceniem długozasięgowych zależności. Mechanizm uwagi wprowadza nową jakość, umożliwiając równoczesne rozważanie wszystkich elementów sekwencji wejściowej (14).

Mechanizm uwagi w Transformerach jest dodatkowo wzbogacony przez mechanizm wielogłowowej uwagi (Multi-Head Attention). Pozwala on na równoczesne uchwycenie różnych kontekstów poprzez zastosowanie wielu zestawów zapytań (Queries), kluczy (Keys) i wartości (Values). Dla danego wektora wejściowego $X$, operacja uwagi jest definiowana wzorem (14):

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}(\frac{\mathrm{QK}^T}{\sqrt{d_k}})\mathrm{V}$$

Gdzie:

- $Q$ (Query), $K$ (Key) i $V$ (Value) to zlinearyzowane reprezentacje wektora wejściowego,
- $d_k$ to wymiar przestrzeni kluczowej,
- $softmax()$ to funkcja normalizująca wyniki, przekształcająca wartości w prawdopodobieństwa.

Mechanizm Multi-Head Attention rozszerza tę koncepcję, pozwalając na wielokrotne zastosowanie mechanizmu uwagi skupiając się na innych aspektach danych. Jest to szczególnie istotne w złożonych zadaniach, gdzie różne konteksty mogą dostarczać różnych, ale komplementarnych informacji (14).

Vision Transformer (ViT) to przykład modelu, który adaptuje architekturę Transformerów do zadań wizualnych. ViT dzieli obraz na mniejsze fragmenty i traktuje je jako tokeny sekwencji, podobnie jak słowa w NLP. Mechanizm uwagi pozwala następnie na analizę całego obrazu, biorąc pod uwagę globalne zależności między różnymi częściami obrazu. W przypadku FER, to pochodne Vision Transformers stanowią wiodące implementacje na rynku (15). Należą do nich: Shifted Window Transformer (16), Data-efficient Image Transformers (17), Transformer in Transformer (18)

## Odwołania


1. **Plutchik, Robert.** *Emotion: A Psychoevolutionary Synthesis.* s.l. : Harper & Row, 1980.

2. **Ekman, Paul.** *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life.* 2nd ed. s.l. : Henry Holt and Co., 2007.

3. **Goodfellow, Ian, Yoshua Bengio, and Aaron Courville.** *Deep Learning.* s.l. : MIT Press, 2016.

4. *"An Argument for Basic Emotions".* **Ekman, Paul.** 3-4, 1992, Cognition & Emotion, Vol. 6, pp. 169-200.

5. *"ImageNet Classification with Deep Convolutional Neural Networks".* **Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E.** s.l. : Curran Associates Inc., 2012. Advances in Neural Information Processing Systems. Vol. 25, pp. 1097-1105.

6. *"Deep Residual Learning for Image Recognition".* **He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.** 2015. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778.

7. *"EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks".* **Tan, Mingxing, and Quoc V. Le.** 2019. Proceedings of the 36th International Conference on Machine Learning (ICML). pp. 6105-6114.

8. **Simonyan, Karen, and Andrew Zisserman.** "Very Deep Convolutional Networks for Large-Scale Image Recognition". *arXiv preprint.* [Online] 2014. https://arxiv.org/abs/1409.1556.

9. *"A Robust Framework for Deep Learning Approaches to Facial Emotion Recognition and Evaluation".* **Siddiqui, Nyle, N. Shakeeb Khan, Mahreen Lakhani, and Omar Ghori.** 2022, arXiv.

10. *"The Graph Neural Network Model".* **Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini.** 1, 2009, IEEE Transactions on Neural Networks, Vol. 20, pp. 61-80.

11. **Kipf, Thomas N., and Max Welling.** arXiv. *"Semi-Supervised Classification with Graph Convolutional Networks".* [Online] 2017. https://arxiv.org/abs/1609.02907.

12. *"Graph Attention Networks".* **Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio.** 2018. International Conference on Learning Representations (ICLR).

13. *"Superpixel Image Classification with Graph Convolutional Neural Networks Based on Learnable Positional Embedding".* **Lin, Heng, Su Yang, and Debesh Jha.** 2022, Applied Sciences, Vol. 12, p. 9176.

14. *Attention is all you need.* **Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N. and Kaiser, Łukasz and Polosukhin, Illia.** 2017. Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010.

15. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* **al., Alexey Dosovitskiy et.** s.l. : CoRR, 2020, Vol. abs/2010.11929.

16. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.* **Liu, Ze and Lin, Yutong and Cao, Yue and Hu, Han and Wei, Yixuan and Zhang, Zheng and Lin, Stephen and Guo, Baining.** s.l. : 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. pp. 9992-10002.

17. *Training data-efficient image transformers & distillation through attention.* **Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Herve Jegou.** s.l. : Proceedings of the 38th International Conference on Machine Learning, 2021.

18. *Transformer in Transformer.* **Wang, Kai Han and An Xiao and Enhua Wu and Jianyuan Guo and Chunjing Xu and Yunhe.** 2021.

19. *"Deep Learning".* **LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton.** 7553, 2015, Nature, Vol. 521, pp. 436-444.