Raport z Projektu

Few shot learning - analiza metod

Nauczanie na strzała

Wakacyjne Wyzwanie Solvro 2025

28 września 2025

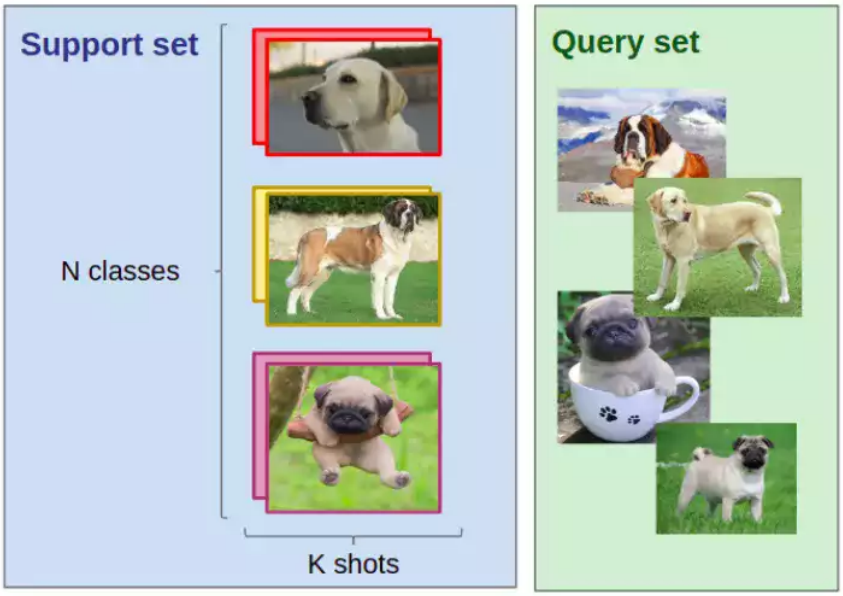
Zespół: Paweł Litwin, Bartłomiej Niedbała, Przemysław Mazurowski, Wiktor Oziewicz

Koordynator: Julia Farganus

Spis treści

Streszczenie

Abstrakt: Współczesne systemy rozpoznawania emocji z twarzy, oparte na głębokim uczeniu, osiągają imponującą skuteczność. Jednak ich fundamentalnym ograniczeniem jest uzależnienie od dużych, starannie oznaczonych zbiorów danych treningowych. Zebranie tysięcy próbek dla każdej kategorii emocji jest w wielu praktycznych zastosowaniach kosztowne, czasochłonne, a często wręcz niemożliwe. W odpowiedzi na to wyzwanie, niniejszy projekt koncentruje się na eksploracji metod **few-shot learning** (uczenia się z kilku przykładów) w kontekście rozpoznawania emocji. Głównym celem pracy jest sprawdzenie efektywności techniki **N-way K-shot**, która pozwala modelowi na naukę nowych klas emocji na podstawie jedynie kilku przykładów w tzw. *support set*. W podejściu tym wykorzystano transfer learning - modele wstępnie wytrenowane na dużych zbiorach obrazów służą do ekstrakcji cech twarzy, które następnie są adaptowane do nowych zadań klasyfikacji. Projekt obejmuje porównanie skuteczności metod few-shot learning z tradycyjnymi podejściami, w celu identyfikacji najbardziej obiecujących technik dla zastosowań praktycznych z ograniczoną liczbą danych..



Rysunek 1: Wizualizacja techniki N-way K-shot

Na Rysunku 1. przedstawiono przykład zadania typu few-shot: mając $K = 2$ przykłady dla każdej z $N = 3$ klas w zbiorze wsparcia, chcemy przypisać etykiety $Q = 4$ psom ze zbioru zapytań jako Labrador, Bernardyn lub Mops. Dla człowieka, nawet bez wcześniejszej znajomości tych ras, zadanie to byłoby intuicyjnie proste. Jednak w przypadku sztucznej inteligencji konieczne jest zastosowanie **meta-learningu**, czyli podejścia, w którym algorytm „uczy się uczyć", albowiem poprawia on swoje zdolności rozwiązywania nowych zadań dzięki doświadczeniu zdobytemu na wielu różnych problemach. W przeciwieństwie do tradycyjnego uczenia, meta-learning skupia się na szybkim dostosowaniu się do nowych klas na podstawie jedynie kilku przykładów.

1. Wprowadzenie

1.1. Cel projektu

Głównym celem projektu jest zbadanie możliwości wykorzystania technik **few-shot learningu** do rozpoznawania emocji z twarzy na podstawie zdjęć ze zbioru: Jaffe, CK+ i fer2013.

1.2. Zakres projektu

Projekt obejmuje:

- EDA na zbiorach danych: fer013, CK+, Jaffe,
- Ekstrakcję cech twarzy przy pomocy modeli pretrenowanych,
- Poszukanie sposobów na robienie few-shot learningu, jak to się ma do transferu stylu i uczenia nieadzorowanego,
- Porównanie accuracy różnych metod.

1.3. Opis użytych zbiorów danych

1.3.1. Zbiór CK+

Rysunek 2: Wizualizacja zbioru CK+

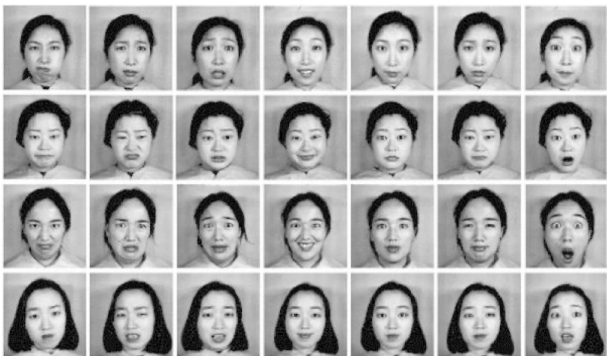
Rozszerzony zbiór danych Cohn-Kanade (CK+) zawiera 593 obrazy pochodzące od 123 różnych osób w wieku od 18 do 50 lat, różnej płci. Każdy obraz ma rozdzielczość 640x480 pikseli i jest zapisany głównie w skali szarości. 327 z tych obrazów zostało oznaczonych jednym z siedmiu emocjonalnych wyrażeń twarzy. Baza danych CK+ jest uznawana za najpowszechniej stosowany laboratoryjnie kontrolowany zbiór danych do klasyfikacji wyrazów twarzy i jest wykorzystywana w większości metod klasyfikacji ekspresji emocjonalnych. Przykładowe obrazy ze zbioru danych CK+ przedstawiono na Rysunku 2.

1.3.2. Zbiór fer2013

Rysunek 3: Wizualizacja zbioru fer2013

Zbiór danych FER2013 wykorzystany do stworzenia modelu w niniejszym badaniu zawiera 35 887 obrazów twarzy. Obrazy te są zapisane w skali szarości i mają rozdzielczość 48x48 pikseli. Część zdjęć posiada przypisane etykiety emocji. Zbiór zawiera fotografie przedstawiające 7 różnych emocji - szczęście, smutek, zaskoczenie, wstręt, złość, strach oraz neutralną - opisanych w poprzednich sekcjach. Jest to zestaw danych o otwartym dostępie. Przykładowe obrazy wykorzystane w zbiorze FER2013 przedstawiono na Rysunku 3.

1.3.3. Zbiór Jaffe



Rysunek 4: Wizualizacja zbioru Jaffe

Baza danych Jaffe jest publicznie dostępnym zbiorem zawierającym 213 obrazów wyrażeń twarzy 10 japońskich kobiet, nazywana Japanese Female Facial Expression. Każda z badanych osób prezentuje 7 podstawowych emocji, przy czym każde wyrażenie emocjonalne jest reprezentowane przez 3-4 obrazy na uczestniczkę. Wszystkie obrazy są zapisane w skali szarości i mają rozdzielczość 256×256 pikseli. Fotografie twarzy zostały wykonane w podobnych, ściśle kontrolowanych warunkach oświetleniowych, bez wpływu czynników zakłócających takich jak włosy czy okulary. Przykładowe obrazy wykorzystane w zbiorze Jaffe przedstawiono na Rysunku 4.

2. Metody

W tej sekcji przedstawiono użyte przez nas modele backbone wykorzystane do transfer-learningu i podstawy oraz implementacyjne aspekty metod few-shot learning wykorzystanych w badaniu. Zaprezentowano klasyfikację metod few-shot learning, szczegółowy opis wybranych podejść, oraz specyfikację techniczną zastosowanych modeli.

2.1. Użyte modele backbone:

- Inception ResNet v1, trenowany na datasecie VGGface2

2.2. Rodzaje metod few-shot

W Tabeli 1. znajduje się klasyfikacja czterech głównych typów metod Few-Shot Learning. Każda kategoria reprezentuje odmienną filozofię rozwiązywania problemu uczenia się z ograniczoną liczbą przykładów, od metod opartych na metrykach po zaawansowane architektury modeli. Klasyfikacja ta służy jako punkt wyjścia do szczegółowego omówienia konkretnych metod i ich zastosowań w praktyce.

Typ	**Opis**
Metric	Uczy embeddingów i klasyfikuje przez porównywanie próbek w przestrzeni cech.
Model	Projektuje specjalną architekturę sieci (np. pamięć, attention), aby lepiej generalizowała do nowych zadań.
Optimization	Metody oparte na optymalizacji wykorzystują Learner i meta-learner.
Hybrid	Różne pomysły i połączenia technik – opisane szczegółowo w literaturze.

Tabela 1: Podstawowe typy metod Few-Shot Learning i ich krótka charakterystyka.

2.3. Metody few-shot

Tabela 2. przedstawia przegląd metod typu Metric, które koncentrują się na nauce przestrzeni embeddingu, gdzie podobne próbki są grupowane blisko siebie. Kluczową cechą tych metod jest klasyfikacja przez porównywanie odległości między próbkami query setu a przykładami z support setu. Podejścia takie jak Prototypical Networks czy Relation Networks różnią się głównie sposobem obliczania miar podobieństwa w przestrzeni cech.

2.3.1. Metody Metric

Metoda	Opis
CSN	Dwie identyczne CNN wyciągają cechy z par obrazów, a mała sieć na wyjściu przewiduje, czy należą do tej samej klasy.
Matching Networks	Każdy obraz support setu otrzymuje wagę uwagi na podstawie podobieństwa do obrazu query, a predykcja to ważona suma etykiet.
Prototypical Networks	Oblicza prototypy i klasyfikuje query na podstawie najbliższego prototypu.
Relation Networks	Używa sieci neuronowej do obliczania podobieństwa między cechami query i support setu, zamiast prostej metryki.
TADAM	Uczy się dopasowywać embeddingi do zadania poprzez warstwy FILM, które modyfikują cechy w zależności od support setu.
TapNet	Używa macierzy projekcji do dostosowania cech do zadania, zamiast warstw FILM, co czyni go lżejszym od TADAM.
CTM	Skupia się na ekstrakcji cech relevantnych dla zadania poprzez silny meta-learning, ale jest złożony i trudny do implementacji.

Tabela 2: Opis Metod typu Metric.

Tabela 3. przedstawia przegląd metod typu Model, które charakteryzują się projektowaniem specjalnych architektur sieci neuronowych. Wspólną cechą tych podejść jest wykorzystanie mechanizmów pamięciowych lub warunkowania, umożliwiających modelowi zapamiętywanie informacji z support setu. Metody te oferują zaawansowane możliwości adaptacji do nowych zadań, choć często wiążą się z większą złożonością obliczeniową.

2.3.2. Metody Model

Metoda	Opis
MANN	Przechowuje dane z support setu, co pozwala modelowi lepiej zapamiętywać i następnie klasyfikować query set. Niestety ciężkie obliczeniowo.
MM-Net	Przechowywuje cechy support setu w pamięci key-value i używa ich do klasyfikacji query setu. Meta-learning + ciężkie obliczeniowo.
MetaNets	Wykorzystuje fast weights przechowywane w pamięci, by szybko adaptować model do nowego zadania bez wielu iteracji.
CSN	Warunkowanie aktywacji sieci na podstawie specyficznych dla zadania danych przechowywanych w pamięci.
SNAIL	Konwolucje sekwencyjne + atencja sekwencyjna, umożliwiają modelowi „zapamiętanie” kontekstu zadania.

Tabela 3: Opis Metod typu Model.

Tabela 4. obejmuje metody Hybrid, które łączą w sobie elementy różnych podejść few-shot learning. Kombinacje te mogą obejmować na przykłąd połączenia metod metric-based z optimization-based.

2.3.3. Metody Hybrid

Metoda	Opis
Cross-Model FSL	Sprowadza się do basic fine-tuningu, ponieważ nie mamy innych danych.
Semi-Supervised FSL	Używa się, gdy większość danych jest unlabeled.
Generalized FSL	Zbyt złożone (overkill) dla 7 klas, używane przy dziesiątkach setkach klas.
Generative FSL	Augmentowanie i generowanie nowych syntetycznych danych na podstawie support setu.
Cross-domain FSL	Model pretrenowany na zadaniach z jednej domeny, testowany na zadaniach z innej domeny.
Transductive FSL	Fine-tuning na support secie z wykorzystaniem query setu przy tym procesie.
Unsupervised FSL	Brak użyteczności w tym kontekście.
Zero-Shot learning	Zero support setu, model korzysta z wiedzy ogólnej do klasyfikacji.

Tabela 4: Opis Metod typu Hybrid.

2.3.4. Metody Optimization

Metody oparte na optymalizacji wykorzystują Learner i Meta-Learner.

- **Learner** to model trenowany na support secie do klasyfikacji, podczas gdy
- **Meta-Learner** uczy się na dystrybucji podobnych tasków w celu wyuczenia parametrów początkowych i strategii aktualizacji algorytmu.

Jako iż w naszym przypadku mamy tylko jeden subtask, nie ma sensu stosować meta-learningu.

2.3.5. Metody wybrane przez nas jako najlepsze do przetestowania na dostępnych datasetach

Metoda	**Opis**
Matching Networks	Każdy obraz support setu otrzymuje wagę uwagi na podstawie podobieństwa do obrazu query, a predykcja to ważona suma etykiet.
Prototypical Networks	Oblicza prototypy i klasyfikuje query na podstawie najbliższego prototypu.
Cross-domain FSL	Model pretrenowany na zadaniach z jednej domeny, ale dotrenowujemy/testujemy na zadaniach z innej domeny.

Tabela 5: Opis Metod uznanych przez nas za najlepsze.

Powyższa tabela 5 zawiera opis wybranych przez nas metod, które uznaliśmy za właściwe w kontekście posiadanych zbiorów danych.

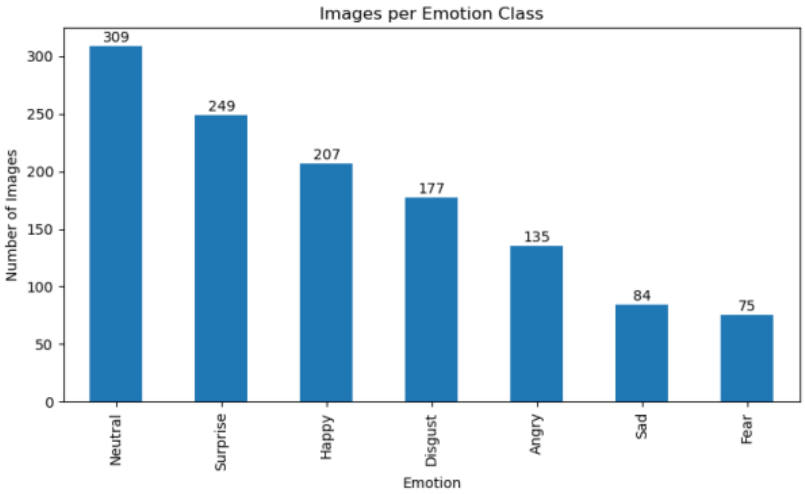
3. Wyniki i analiza

W tej sekcji przedstawiono szczegółowe wyniki przeprowadzonych eksperymentów oraz ich kompleksową analizę. Badania obejmują EDA dla trzech wykorzystanych zbiorów danych oraz ewaluację skuteczności metod few-shot learning w klasyfikacji emocji. Wyniki zostały zwizualizowane w formie tabel, a następnie poddane szczegółowej interpretacji pod kątem skuteczności, wydajności i praktycznego zastosowania.

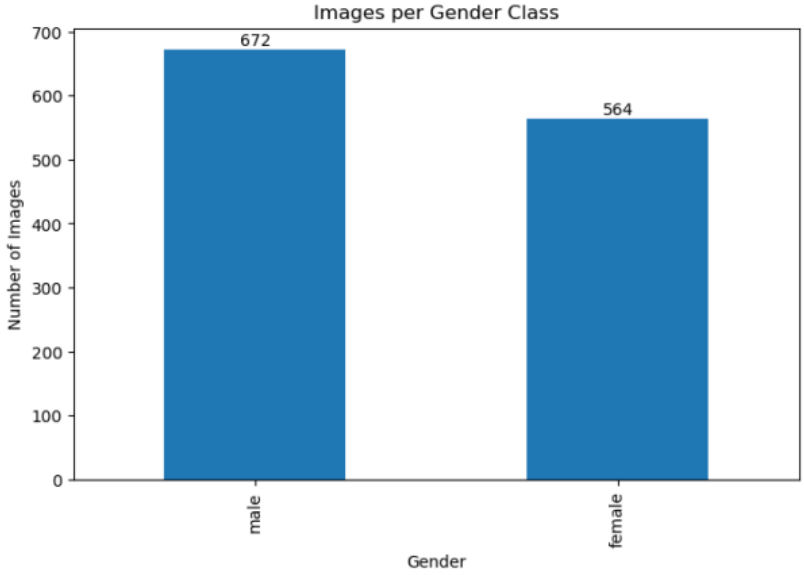
Z powodów ograniczonego dostępu do mocy obliczeniowej wymaganej do dotrenowania modelu backbone, wykorzystanego w technikach opartych na metrykach, dotrenowywanie modelu zostało jedynie wykonane przy korzystaniu z datasetu Jaffe.

3.1. Wyniki EDA

3.1.1. EDA dla datasetu CK+



(a) Rozkład emocji

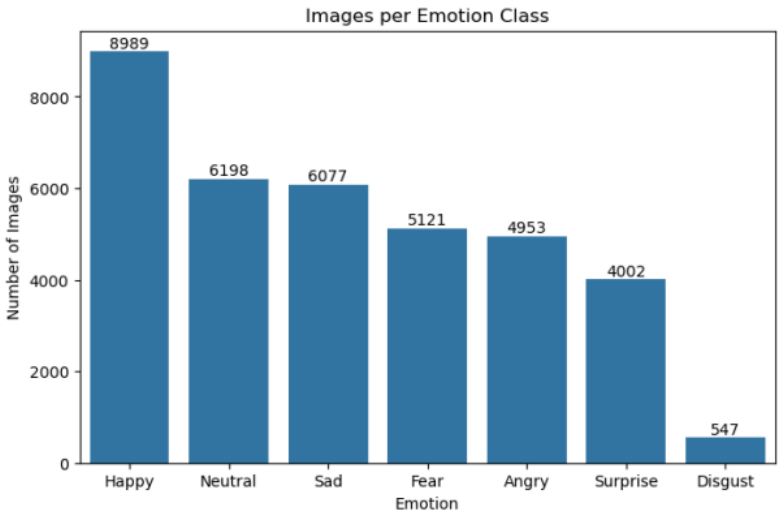


(b) Rozkład płci

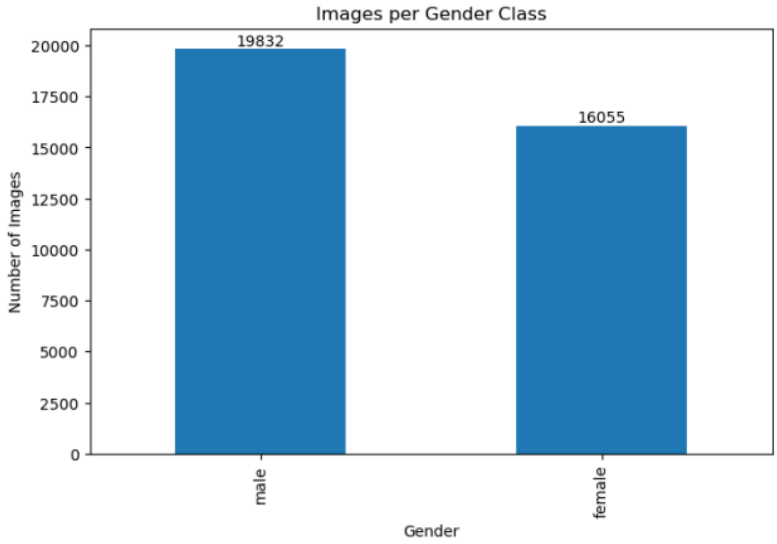
Rysunek 5: Analiza zbioru CK+ - rozkład emocji i płci

Podsumowanie: Zbiór CK+ wyróżnia się sekwencyjnym charakterem danych (3-5 klatek na osobę), co pozwala na przechwycenie dynamicznych zmian mimiki twarzy. Występuje umiarkowana nierównowaga klasowa między emocjami. Różnice w rozkładzie płci są nieznaczne. Mieszany format obrazów (część RGB, część grayscale) oraz wyższa rozdzielczość (640×480) zapewniają więcej informacji, ale wymagają ujednolicenia przed przetwarzaniem.

3.1.2. EDA dla datasetu fer2013



(a) Rozkład emocji dla FER2013

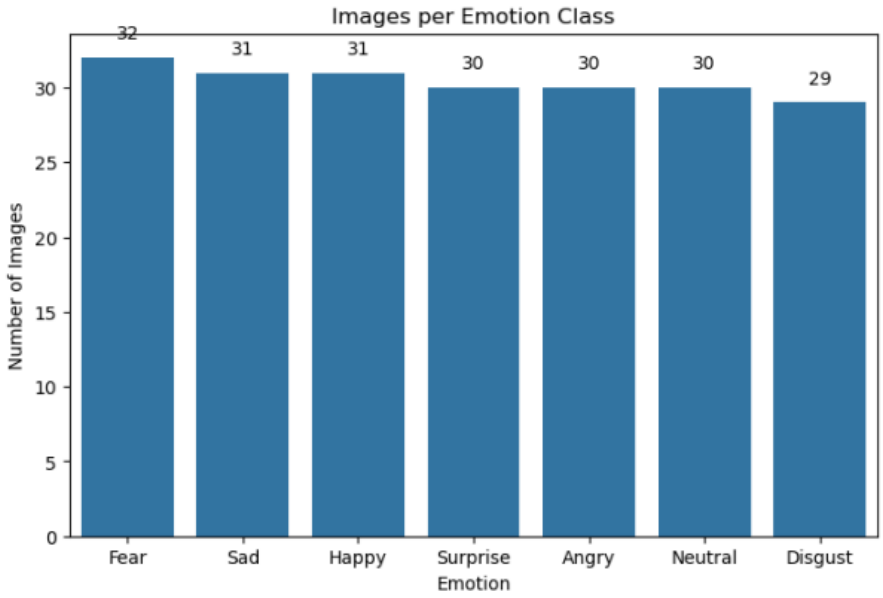


(b) Rozkład płci dla FER2013

Rysunek 6: Analiza zbioru FER2013 - rozkład emocji i płci

Podsumowanie: Zbiór FER-2013 zawiera obrazy w skali szarości o stałej niskiej rozdzielczości 48×48 pikseli, co znacząco ogranicza ilość informacji dostępnej dla modelu. Występuje ekstremalna nierównowaga klasowa - różnica między największą a najmniejszą klasą wynosi około 1650%. Konieczne może być zastosowanie technik balansowania klas (np. oversampling, undersampling, weighting). Różnice w rozkładzie płci są nieznaczne.

3.1.3. EDA dla datasetu Jafffe



Rysunek 7: Rozkład emocji dla Jaffe

Podsumowanie: Zbiór Jaffe zawiera zbalansowane liczebnie klasy emocji (29-31 obrazów na emocję) ale charakteryzuje się silnym biasem płciowym - wszystkie zdjęcia przedstawiają wyłącznie kobiety. Obrazy mają wyższą rozdzielczość (256×256 pikseli) w porównaniu do FER-2013. Format TIFF zapewnia lepszą jakość obrazu, jednak mały rozmiar zbioru (213 obrazów) może ograniczać możliwości treningowe.

3.2. Analiza wyników dla użytych metod few-shot

3.2.1. Analiza wyników dla metody Matching networks

Wstęp do metody:

Opis metody: Matching Networks to architektura sieci neuronowej, stworzona do klasyfikacji few-shot. Działanie tego modelu opiera się na porównywaniu ze sobą embeddingu obrazu z query, z embeddingami wszystkich obrazów z support setu. Dzięki zastosowaniu mechanizmu atencji, odpowiedzą modelu na zapytanie jest klasa z support setu, do której próbek obraz z zapytania był najbardziej podobny.

Rozkład prawdopodobieństwa przynależności próbki z query do każdej z klas z support setu obliczany jest z poniższego wzoru:

$$\hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i \tag{1}$$

gdzie:
a - funkcja atencji
$\hat{x}$ - probka z query set
x_i, y_i - probki i ich etykiety z support set $S = \{(x_i, y_i)\}_{i=1}^{k}$

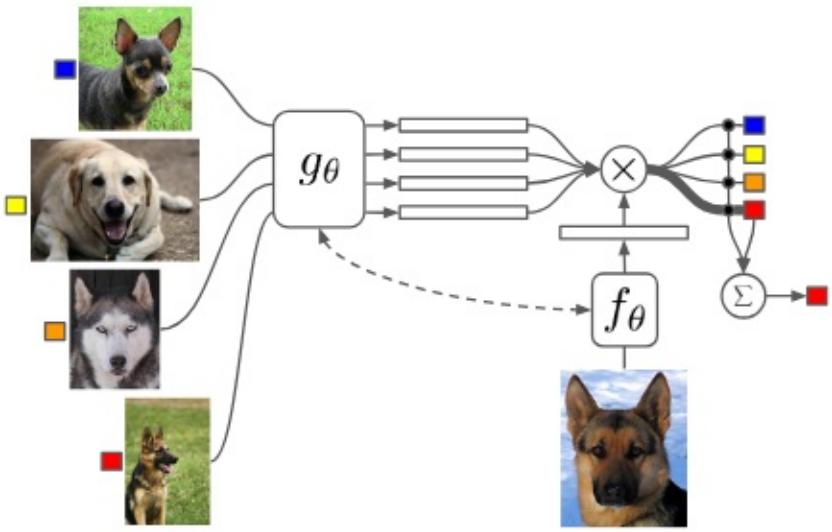
Ze względu na to, że etykiety są kodowane one-hot ($\hat{y}$ - też jest wektorem), wybranie klasy sprowadza się do znalezienia która kategoria otrzymała najwyższą ważoną sumę z mechanizmu atencji. Zastosowana funkcja atencji:

$$a(\hat{x}, x_i) = \frac{e^{c(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{k} e^{c(f(\hat{x}), g(x_j))}} \tag{2}$$

gdzie:
f - funkcja embeddingu dla query set
g - funkcja embeddingu dla support set
c - funkcja odległości cosinusowej

Zarówno fukncja f, jak i g w praktyce implementowane są jako odpowiednie sieci neuronowe (stąd nazwa - $MatchingNetworks$). Model podczas procesu meta-uczenia, uczy się jak najlepiej, na podstawie własnych parametrów i otrzymanego w danym tasku support setu, stworzyć dla nowo spotkanego zadania najlepsze funkcje - próbki z tej samej klasy muszą być przekształcone do postaci jak najbardziej do siebie podobnych i jednocześnie różnych od próbek z innych klas. Jeden z wariantów omawianej architektury zakłada, że $f = g$.

Dobrą wizualizację działania omówionej architektury stanowi zamieszczony niżej Rysnunek 7. Warto dodać, że w praktycznych implementacjach, zamiast obrazów, do sieci wprowadzane są embeddingi z pre-trenowanych do używanych danych modeli.



Rysunek 8: Architektura *MatchingNetworks*

Opis trenowania i testowania: Testy przeprowadzono przy użyciu implementacji modelu Matching Networks z biblioteki **easyfsl**. Do stworzenia wstępnych embeddingów obrazu użyto modelu **InceptionResnetV1** z biblioteki **facenet-pytorch**, pretrenowanego na zbiorze vggface2.

Pierwsze 3 testy - Tabele: 6, 7, 8 przeprowadzono na modelu bez zastosowania meta-learningu, zadania były wybierane z całego zbioru - 7 klas. Na test składają się zadania klasyfikacji do 2, 3 i 5 klas, na podstawie 1, 3 i 5 próbek. Zadań testowych dla każdych ustawień jest 50, a użyta metryka to accuracy. W zbiorze testowym z każdej klasy było 15 próbek (query set).

Ostatnie dwa testy - Tabele: 9, 10 - wykonano z podziałem na 4 klasy przeznaczone do treningu i 3 do testu. Pierwszy test został wykonany przed, a drugi po wytrenowaniu modelu. Trening przeprowadzono na klasach: angry, fear, sad, surprise - a test na klasach: disgust, happy, neutral. Model wytrenowano na zadaniach 3-way, 5-shot w ciągu 100 epok, z 30 zadaniami na epokę. Użyty optymizator to Adam z learning rate 0.001, a fukcja straty to CrossEntropyLoss. Testy przeprowadzono na zadaniach klasyfikacji do 2 i 3 klas z ustawieniami jak wyżej.

Uzyskane wyniki:

Accuracy	1-shot	3-shot	5-shot
2-way	55%	60%	61%
3-way	40%	44%	43%
5-way	24%	26%	30%

Tabela 6: Jaffe dataset

Accuracy	1-shot	3-shot	5-shot
2-way	55%	56%	59%
3-way	36%	40%	40%
5-way	22%	25%	26%

Tabela 7: Fer2013 dataset

Accuracy	1-shot	3-shot	5-shot
2-way	59%	63%	66%
3-way	43%	48%	52%
5-way	29%	32%	35%

Tabela 8: CK+ dataset

Accuracy	1-shot	3-shot	5-shot
2-way	53%	57%	55%
3-way	37%	39%	41%

Tabela 9: Fer2013 train baseline

Accuracy	1-shot	3-shot	5-shot
2-way	50%	50%	52%
3-way	34%	35%	35%

Tabela 10: Fer2013 dataset + train

Wnioski z otrzymanych wyników dla metody Matching networks: Model w wersji bez treningu działa lepiej niż losowy wybór klas i jego skuteczność rośnie wraz z liczbą obrazków w support secie, więc model działa zgodnie z oczekiwaniami.

Meta-uczenie nie przyniosło dobrych rezultatów - pogorszyło wyniki modelu. Najprawdopodobniej jest to spowodowane zbyt mała liczbą klas w treningu. Zamiast oczekiwanego efektu przystosowania się modelu do nowych zadań, nastąpiło dostosowanie się do niewielkiej liczby klas w treningu i obniżenie skuteczności dla nowych klas. Zbadana powyżej metoda nie jest więc skutecznym rozwiązaniem probelmu few-shot learningu dla analizy sentymentu, ze względu na małą liczbe klas w data setach.

3.2.2. Analiza wyników dla metody Prototypical Networks

Wstęp do metody:

Opis metody: Sieci prototypowe to kolejna architektura stworzona do zadań few-shot. Ich działanie polega na wyznaczaniu **prototypów** - centroidów tych samych klas z support set jako reprezentantów. Najpierw dane wejściowe są przekształcane za pomocą sieci neuronowej w swoje reprezentacje wektorowe w przestrzeni cech. Używana sieć neuronowe będzie pełniła funkcję ekstraktora cech dla wszystkich klas. Dla każdej klasy z support set obliczany jest prototyp.

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\theta(x_i) \tag{3}$$

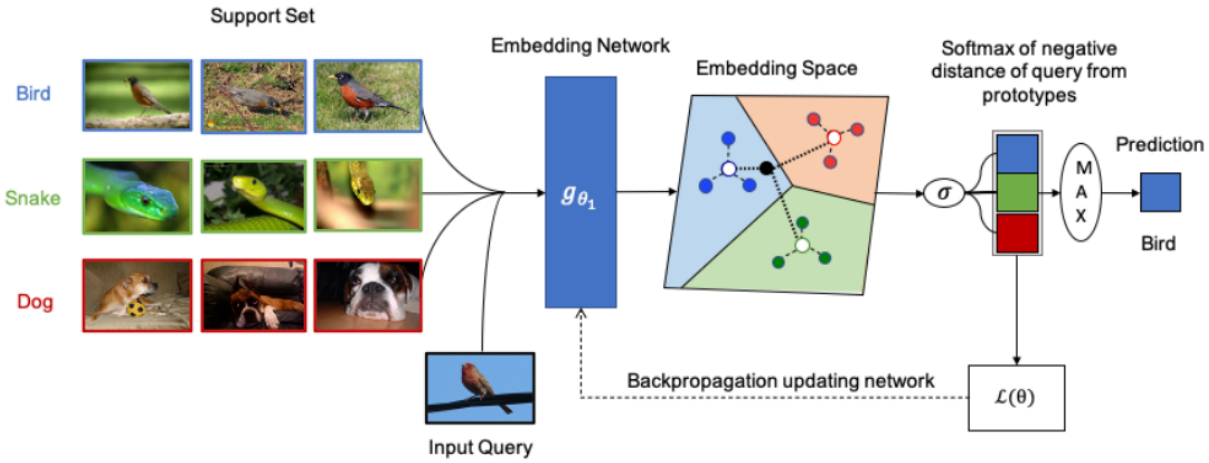
gdzie:
c_k - prototyp klasy k
S_k - support set klasy k
f_θ - funkcja ekstrakcji

Następnie nowe próbki (query set) są przekształcane w wektory w tej samej przestrzeni cech za pomocą f_θ. Ich odległości do prototypów są liczone za pomocą danej funkcji odległości, a rozkład prawdopodobieństwa klas jest obliczany za pomocą funkcji softmax.

$$p(y = k|x) = softmax(-d(f_\theta(x), c_k)) = \frac{e^{-d(f_\theta(x), c_k)}}{\sum_{c'_k \in C} e^{-d(f_\theta(x), c'_k)}} \tag{4}$$

gdzie:
x - wektor cech próbki
p - funkcja rozkładu prawdopodobieństwa
C - zbiór prototypów

Wynikiem klasyfikacji jest klasa o najwyższym prawdopodobieństwie, co odpowiada klasie prototypu z najmniejszą odległością w przestrzeni cech od próbki.



Rysunek 9: Architektura metody protypical networks

Opis trenowania i testowania: Testy przeprowadzono w dwóch wariantach: bez trenowania i z trenowaniem modelu. Użyto modelu bazowego **InceptionResnetV1** pretrenowanego na zbiorze **VGGFace2**.

Faza testów pre-trenowanego modelu: Architektórę modelu bazowego zmodyfikowano poprzez usunięcie końcowej warstwy klasyfikującej, adaptując ją do roli ekstraktora cech. Do testowania modelu używano klas: disgust, happy, neutral. We wszystkich eksperymentach używano następującej konfiguracji:

- Wielkość zbioru query: 15
- Metryki: dokładność na zbiorze query
- Wielkości n-way zostały przetestowane od 2 do 3
- Wielkości k-shot zostały przetestowane od 1 do 5
- Liczba epizodów na epokę: 50

Tej fazie odpowiadają wyniki w tabelach 11, 12 i 13.

Faza meta-learningu: Ten sam model bazowy trenowano na podzbiorze używając klas: angry, fear, sad, surprise. Hiperparametry, takie jak ilość epok, liczba epizodów na epokę i współczynnik uczenia, zmieniano podczas treningu. Z uwagi na małą ilość klas w zbiorach, wyniki przedstawiono dla najlepszej znalezionej konfiguracji:

- Ilość epok: 30
- Optimizer: Adam (learning rate = 0.001)
- Funkcja straty: CrossEntropyLoss

Tej fazie odpowiadają wyniki w tabelach 14, 15 i 16.

Uzyskane wyniki: W tabelach przedstawiono szczegółowe wyniki przeprowadzonych eksperymentów.

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	51.20%	55.20%	57.40%	53.80%	55.07%
3-way	38.27%	38.71%	38.98%	39.07%	40.80%

Tabela 11: CK+ dataset

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	51.07%	50.73%	50.47%	48.80%	50.67%
3-way	33.51%	35.38%	34.53%	34.44%	34.36%

Tabela 12: FER2013 dataset

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	67.07%	69.33%	66.87%	67.60%	69.87%
3-way	52.13%	54.44%	54.40%	53.07%	54.76%

Tabela 13: JAFFE dataset

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	51.20%	53.40%	53.47%	56.67%	55.27%
3-way	35.20%	37.16%	37.91%	38.58%	37.87%

Tabela 14: CK+ dataset + train

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	49.07%	48.87%	51.00%	52.47%	50.60%
3-way	33.64%	33.51%	35.16%	34.71%	33.51%

Tabela 15: FER2013 dataset + train

Accuracy	1-shot	2-shot	3-shot	4-shot	5-shot
2-way	62.47%	67.93%	70.80%	74.20%	68.67%
3-way	50.62%	50.58%	53.42%	54.13%	56.09%

Tabela 16: JAFFE dataset + train

Wyniki i analiza metody Prototypical Networks Eksperymenty wykazały, że wpływ dodatkowego trenowania modelu na zbiorach docelowych nie był jednoznaczny. W niektórych przypadkach model korzystający jedynie z wstępnie wytrenowanego embeddera osiągał minimalnie lepsze wyniki — przykładem jest zbiór FER2013 (tabele **12** i **15**), gdzie dodatkowy trening nie poprawił jakości klasyfikacji, a dokładności oscylowały wokół wartości zbliżonych do losowych. Z drugiej strony, w zbiorach bardziej jednorodnych i mniej zaszumionych, takich jak JAFFE (tabele **13** i **16**) czy CK+ (tabele **11** i **14**), dodatkowe trenowanie pozwoliło uzyskać wyższe wyniki, zwłaszcza w konfiguracjach 2-way i wyższych k-shot.
Można stwierdzić, że skuteczność podejścia zależała od charakterystyki zbioru danych: dla zbiorów czystszych i bardziej spójnych trenowanie na docelowych klasach poprawiało reprezentacje prototypów, natomiast dla zbiorów trudniejszych dodatkowe dostrajanie mogło prowadzić do przeuczenia i spadku dokładności.

Korelacja z rozmiarem zbioru wsparcia (k-shot) Wyniki potwierdzają oczekiwaną zależność: w większości konfiguracji dokładność klasyfikacji rosła wraz ze zwiększaniem liczby przykładów na klasę (k-shot) w zbiorze wsparcia. Korelacja ta jest zgodna z założeniami metody Prototypical Networks, ponieważ większa liczba przykładów pozwala na estymację bardziej reprezentatywnego i odpornego na szumy prototypu każdej klasy.

Ogólna ocena zastosowania few-shot learningu Zastosowanie metody Prototypical Networks do zadania rozpoznawania emocji na zdjęciach nie przyniosło zadowalających rezultatów. Uzyskane średnie poziomy dokładności były jedynie nieznacznie wyższe od wartości oczekiwanej dla klasyfikacji losowej. Technika okazała się przy tym bardzo wrażliwa na jakość embeddera, a jej skuteczność znacząco spadała w przypadku obrazów o niskiej jakości lub większej złożoności, takiej jak w fer2013. Wyniki sugerują, że w przypadku tak wymagających zadań klasyczne podejście few-shot learningu może być niewystarczające.

3.2.3. Analiza wyników dla metody Cross-domain FSL

Wstęp do metody:

Opis metody: Cross-Domain Few-Shot Learning (CD-FSL) stanowi zaawansowane podejście w dziedzinie uczenia maszynowego, które łączy wyzwania związane z transferem wiedzy między domenami (cross-domain) z uczeniem się na małej liczbie przykładów (few-shot learning). Głównym celem CD-FSL jest opracowanie modeli zdolnych do szybkiej adaptacji do nowych zadań w warunkach, gdy:

- Dostępna jest jedynie niewielka liczba przykładów treningowych (few-shot)
- Domena danych treningowych różni się od domeny danych testowych (cross-domain)

Opis trenowania i testowania:

Faza fine-tuningu: Model bazowy InceptionResnetV1 został poddany procesowi dostrajania na wybranym podzbiorze klas ze zbioru CK+. Wykorzystano następujące parametry treningowe:

- Optimizer: Adam (learning rate = 0.0005, weight decay = 1e-5)
- Funkcja loss: CrossEntropyLoss
- Liczba epok: 12
- Podział danych: 90% treningowe, 10% walidacyjne

Faza few-shot learning: Po etapie fine-tuningu przeprowadzono eksperymenty few-shot learning z wykorzystaniem protokołu n-way k-shot:

- Konfiguracja: 5-shot, 10-query
- Liczba zadań na epokę: 10
- Liczba epok: 10
- Metryki: dokładność na zbiorze support i query

Uzyskane wyniki: W tabelach przedstawiono szczegółowe wyniki przeprowadzonych eksperymentów dla różnych konfiguracji liczby klas.

Tabela 17: CK+ dataset

Liczba klas FT	**FT Accuracy [%]**	**Liczba klas FSL**	**Support Acc [%]**	**Query Acc [%]**
2	83,33	5	81,07	98,87
3	74,97	4	75,00	93,33
4	65,77	3	33,55	41,86
5	68,10	2	46,67	61,50

Tabela 18: Fer2013 dataset

Liczba klas FT	**FT Accuracy [%]**	**Liczba klas FSL**	**Support Acc [%]**	**Query Acc [%]**
2	85,34	5	26,00	40,53
3	55,84	4	41,50	33,25
4	47,51	3	30,67	40,67
5	49,25	2	50,00	54,50

Tabela 19: Jaffe dataset

Liczba klas FT	**FT Accuracy [%]**	**Liczba klas FSL**	**Support Acc [%]**	**Query Acc [%]**
2	85,71	5	–	90,67
3	92,96	4	–	97,25
4	75,00	3	–	45,78
5	72,22	2	–	58,33

Wnioski z otrzymanych wyników dla metody Cross-domain Na podstawie przeprowadzonych eksperymentów można sformułować następujące wnioski:

Optymalna konfiguracja parametrów Konfiguracja z 2-3 klasami w fazie fine-tuningu okazała się najbardziej efektywna. W przypadku większej liczby klas (4-5) zaobserwowano znaczący spadek skuteczności. Zjawisko to najprawdopodobniej wynika z nadmiernej specjalizacji modelu na zbyt dużej liczbie klas podczas finetuningu, co ogranicza jego zdolność adaptacji do nowych zadań w fazie FSL. Problem ten jest szczególnie widoczny w przypadku datasetu Fer2013, gdzie nawet dla konfiguracji 2 klas fine-tuningu i 5 klas FSL model osiąga dokładność na poziomie zaledwie 40.53% na zbiorze query i 26% na zbiorze support. Zatem ilość danych zabija możliwośc przyszłej nauki modelu.

Negatywna korelacja skuteczności Wyniki wykazują wyraźną negatywną korelację pomiędzy liczbą klas wykorzystywanych w fine-tuningu a finalną skutecznością modelu w warunkach few-shot learning. Potwierdza to hipotezę, że nadmierna specjalizacja modelu na etapie pre-treningu ogranicza jego zdolności adaptacyjne do nowych domen.

Zdolności generalizacyjne We wszystkich konfiguracjach zaobserwowano wyższą dokładność na zbiorze query w porównaniu ze zbiorem support, co świadczy o dobrych zdolnościach generalizacyjnych wytrenowanych modeli. Model nie ulega nadmiernemu dopasowaniu do danych treningowych, lecz skutecznie przenosi nabytą wiedzę na nowe przykłady.

Długość treningu : We wszystkich konfiguracjach eksperymentów datasety Jaffe i CK+ trenowały się (faza fine-tuning + faza FSL), poniżej 20 minut, gdzie jedna klasa FER2013 w samym finetuningu mieliła się ok. 23 minuty. Przy takim rozkładzie danych, widać ogromną różnicę w wykonywanej pracy modelu.

3.2.4. Zbiorcze podsumowanie wyników

Podsumowanie wyników metod: Analizowane metody few-shot learningu – **Matching Networks**, **Prototypical Networks** oraz podejście **Cross-domain** – ujawniły zarówno mocne strony, jak i istotne ograniczenia w kontekście klasyfikacji emocji na podstawie obrazów twarzy.

Matching Networks. Model w wersji bez dodatkowego treningu działał lepiej niż losowy klasyfikator, a jego skuteczność rosła wraz z liczbą przykładów w zbiorze wsparcia (*support set*). Meta-learning nie poprawił jednak wyników - przeciwnie, spowodowało ich spadek. Najbardziej prawdopodobnym powodem była zbyt mała liczba klas dostępnych podczas treningu, co ograniczyło zdolności adaptacyjne sieci i prowadziło do nadmiernego dopasowania do wąskiego zakresu danych.

Prototypical Networks. Eksperymenty wykazały, że kluczowym elementem tej metody jest jakość wykorzystanego embeddera. Sieć korzystająca z wstępnie wytrenowanego modelu (np. na VGGFace2) osiągała porównywalną dokładność co sieć dotrenowywana na ograniczonej liczbie klas. Dodatkowy trening na małym zbiorze docelowym nie przyniósł korzyści i prowadził do gorszej separacji reprezentacji. Wyniki potwierdziły zależność między liczbą przykładów na klasę a skutecznością – większe *k-shot* umożliwiały tworzenie bardziej stabilnych i reprezentatywnych prototypów. Mimo to, średnia dokładność pozostawała tylko nieznacznie powyżej losowej klasyfikacji, szczególnie w przypadku trudniejszych zbiorów, takich jak FER2013.

Cross-domain. Ta metoda wykazała największą wrażliwość na konfigurację fine-tuningu. Najlepsze wyniki uzyskiwano przy wyborze 2–3 klas do fazy dostrajania, natomiast trenowanie na 4–5 klasach powodowało spadek skuteczności. Obserwowano wyraźną negatywną korelację pomiędzy liczbą klas w fine-tuningu a finalną skutecznością modelu. Warto podkreślić, że prawie we wszystkich przypadkach dokładność na zbiorze zapytań (*query set*) była wyższa niż na zbiorze wsparcia, co wskazuje na dobre zdolności generalizacyjne modeli i brak nadmiernego dopasowania do danych treningowych.

Wnioski ogólne. Zastosowane metody few-shot learningu pozwoliły potwierdzić znane w literaturze zależności: skuteczność klasyfikacji rośnie wraz z liczbą przykładów w zbiorze wsparcia oraz z jakością wstępnych reprezentacji cech. Jednocześnie ograniczona liczba klas w zbiorach użytych w projekcie znacząco obniżyła efektywność metod opartych na meta-uczeniu i fine-tuningu. W szczególności Matching Networks i Prototypical Networks osiągały wyniki niewiele lepsze od losowych, natomiast metoda Cross-domain, mimo problemów ze skalowalnością, wykazała stosunkowo największy potencjał dzięki utrzymywaniu zdolności generalizacji.

Podsumowując, w obecnych warunkach zbiory danych stanowiły główne ograniczenie efektywności badanych metod. Ich skuteczniejsze wykorzystanie wymagałoby bogatszych, bardziej zróżnicowanych klas oraz lepszej jakości danych wejściowych.

4. Dyskusja

4.1. Napotkane problemy i ograniczenia

- **Ograniczona moc obliczeniowa** - Ze względu na ograniczony dostęp do mocy obliczeniowej wymaganej do dotrenowania modeli backbone, pełne przetestowanie wszystkich metod few-shot learning było niemożliwe. Dotrenowywanie modelu w przypadku metody Matching Networks zostało wykonane jedynie na datasetcie FER2013.
- **Nierównowaga klasowa w danych** - Zbiór FER2013 charakteryzował się ekstremalną nierównowagą klasową (różnica około 1650% między największą a najmniejszą klasą), co wymagałoby zastosowania dodatkowych technik balansowania danych.
- **Ograniczenia datasetów** - Zbiór Jaffe miał silny bias płciowy (tylko kobiety) i mały rozmiar (213 obrazów), podczas gdy FER2013 miał bardzo niską rozdzielczość (48×48 pikseli), co ograniczało ilość dostępnych informacji. Ponadto bardzo ciekawy był rozkład danych w datasetach, Fer2013 miał w jednej klasie tyle próbek co CK+ i Jaffe razem wzięte, w Cross-Domain FSL był to jeden z największych problemów, ponieważ model dopasowując się do klas w fazie fine-tuningu, nie był już w stanie się douczyć w fazie FSL.

4.2. Dalszy rozwój

- **Większa ilość klas** - Z obserwacji wynika, że zbyt mała liczba klas negatywnie wpływa na wyniki eksperymentów. Rozwiązaniem może być ponowne przetestowanie użytych technik na zbiorach z większą liczbą klas, np. MiniImageNet.
- **Przyszłość przyniesie nowe techniki** - Klasyczne metody: Matching Networks i Prototypical Networks wykazały się niską skutecznością w porównaniu z techniką Cross-domain. Dobrym kierunkiem dalszych badań byłoby zatem przetestowanie nowszych, bardziej zaawansowanych metod few-shot learningu, które prawdopodobnie pozwoliłyby osiągnąć lepsze wyniki.
- **Inne zastosowania** - Próba zastosowania w praktyce, np. w problemie klasyfikacji rzadkich chorób na podstawie zdjęć rentgenowskich.
- **Równouprawnienie danych** - Podział danych w datasetach, aby był w grubszej mierze równy per class per dataset co umożliwi wykonanie eksperymentu który zbada działanie wielkości obrazu i jego jakości do działania modelu, co jest nadal ciężkie bo z 9000 próbek Fer2013 zejść nagle do 31 Jafee, to jak praktycznie zgubić cały dataset.

4.3. Podsumowanie

- **Few-shot learning pozostaje wciąż obszarem niedostatecznie zbadanym** - Uzyskane wyniki wskazują, że pomimo dużego potencjału, obecne metody FSL wciąż mogą nie zapewniać dostatecznej skuteczności w tak złożonych zadaniach jak rozpoznawanie emocji.
- **Rola jakości danych.** Wyniki eksperymentów pokazały, że jakość i spójność danych wejściowych mają kluczowe znaczenie dla skuteczności metod few-shot learningu. Niska rozdzielczość obrazów, silny bias płciowy czy skrajna nierównowaga klasowa istotnie ograniczały możliwości modeli.
- **Baseline z artykułów naukowych był najbardziej skuteczny** - Cross-Domain FSL uważany za baseline w każdym artkule naukowym związanym z FSl, tutaj również się bardzo dobrze spisał, problemem w kwestii jego "różnego accuracy"jest rozkład danych i zły dobór hiperparametrów i ilości danych/klas, na których jest uczony w fazie fine-tuningu. Za dużo danych zabija możliwość przyszłej nauki modelu.
- **Jakość embeddera.** Skuteczność metod takich jak Matching Networks i Prototypical Networks w dużej mierze zależy od jakości zastosowanego embeddera. Trenowanie go od zera jest szczególnie wymagające obliczeniowo, a w przypadku Matching Networks staje się dodatkowo problematyczne ze względu na złożoność metalearningu.

Literatura

[1] Archit Parnami and Minwoo Lee, Learning from Few Examples: A Summary of Approaches to Few-Shot Learning, 2022

[2] Etienne Bennequin, Few-Shot Image Classification with Meta-Learning, 2022

[3] Guneet S. Dhillon, Pratik Chaudhari, Avinash Ravichandran, Stefano Soatto,A Baseline For Few-Shot Image Classification, 2020

[4] Github library: easy-few-shot-learning

[5] Thomas Kopalidis, Vassilios Solachidis, Nicholas Vretos, Petros Daras, Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets, 2024

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra, Matching Networks for One Shot Learning, 2016