Specyfikacja techniczna architektury VGG-Transformer w analizie behawioralnej

1. Wstęp i Strategiczne Założenia Systemu

Współczesna diagnostyka kliniczna stoi przed wyzwaniem obiektywizacji oceny stanów psychofizycznych w obliczu globalnego kryzysu zdrowia psychicznego, pogłębionego przez pandemię COVID-19. Zaburzenia te, dotykające blisko 450 milionów ludzi, naruszają biologiczną homeostazę organizmu, objawiając się w subtelnych sygnałach behawioralnych. Prezentowana architektura hybrydowa stanowi odpowiedź na potrzebę precyzyjnej, nieinwazyjnej detekcji wzorców mowy ciała i mimiki, łącząc rygor konwolucyjnych sieci neuronowych (CNN) z elastycznością modeli Transformer.

Analiza transformacji ("So What?"): Integracja warstw Attention z modelem VGG-16 jest kluczowa dla modelowania długodystansowych zależności w danych wideo. Podczas gdy standardowe warstwy CNN (wykorzystujące małe filtry konwolucyjne – milowy krok w architekturze VGG) doskonale radzą sobie z lokalną ekstrakcją cech wizualnych, mechanizm Multi-Head Attention pozwala systemowi na zrozumienie kontekstu czasowego i "wyostrzenie" uwagi na kluczowych markerach diagnostycznych. Zapewnia to stabilność wyników, eliminując szum tła i skupiając moc obliczeniową na mikroruchach twarzy i gestach

---

2. Inżynieria danych i potok przetwarzania wstępnego (RAF-DB)

Fundamentem systemu jest zbiór **RAF-DB**, ceniony za realistyczne przedstawienie emocji z uwzględnieniem głębi i koloru. Potok przetwarzania został zaprojektowany tak, aby zapewnić najwyższą czystość sygnału wejściowego dla architektury hybrydowej.

Parametry potoku (Preprocessing Pipeline):

- **Skalowanie i normalizacja:** Piksele skalowane są do zakresu [0, 1] przy użyciu mnożnika `1./255`.
- **Resizing:** Obrazy są standaryzowane do formatu **100x100 px**.
- **Augmentacja (tylko zbiór treningowy):** Wprowadzenie zoomu (0.3) oraz poziomego przerzucania obrazu (horizontal flip) w celu zwiększenia generalizacji.
- **Zbiór testowy:** Poddawany identycznej normalizacji i resizingowi, lecz bez augmentacji, co zapewnia rygorystyczną ocenę kliniczną.
- **Kodowanie:** Etykiety dla 7 klas emocji (Surprise, Fear, Disgust, Happy, Sad, Angry, Neutral) przetwarzane są metodą **one-hot encoding**.

Zastosowanie funkcji `keras.applications.vgg16.preprocess_input` umożliwia pełną synergię ze wstępnie wyuczonymi wagami modelu bazowego przy zachowaniu wielkości pakietu (batch size) na poziomie **64**.

---

3. Architektura Hybrydowa: integracja VGG-16 z mechanizmami Attention

System wykorzystuje model **VGG-16 jako ekstraktor cech**, korzystając z zamrożonych wag (non-trainable) wyuczonych na zbiorze ImageNet. Pozwala to na zachowanie potężnych reprezentacji wizualnych przy jednoczesnym skupieniu treningu na warstwach hybrydowych.

Specyfikacja TransformerBlock:

Parametr	Wartość
Wymiar osadzenia (Embedding Dimension)	512
Liczba głowic uwagi (Attention Heads)	4
Wymiar sieci Feed-Forward	512
Dropout Rate	0.1
Funkcja Aktywacji	ReLU

Analiza strukturalna ("So What?"): Po bloku transformera wdrożono dodatkową warstwę **Multi-Head Attention** (`key_dim=512`), która wyostrza detekcję najbardziej finezyjnych wzorców mimicznych. Istotnym elementem architektury jest warstwa **Global Average Pooling (GAP)**, która przekształca wyrafinowane cechy w wektor o stałym rozmiarze. Następnie sygnał przechodzi przez **dwie w pełni połączone warstwy (Dense)**, każda po **4096 jednostek**, z zastosowaniem dropoutu 0.5 w celu przeciwdziałania przeuczeniu. Proces kończy warstwa Softmax, zwracająca prawdopodobieństwo dla 7 klas docelowych. Takie podejście łączy "pamięć" wizualną VGG z dynamiczną hierarchią ważności cech oferowaną przez Transformery.

---

4. Specjalistyczne moduły detekcji: DeepFace i MediaPipe

Synergia modelu bazowego z bibliotekami zewnętrznymi umożliwia wielowymiarowe mapowanie behawioralne w czasie rzeczywistym (klatkowanie co 30 klatek dla wideo 30 fps).

DeepFace: analiza emocjonalna

Wykorzystuje wagowany model VGG do precyzyjnej klasyfikacji 7 stanów emocjonalnych. Funkcja `DeepFace.analyze()` zwraca sumę prawdopodobieństw, co pozwala na ilościową ocenę intensywności afektu (np. identyfikacja "płaskiego afektu" w depresji).

MediaPipe: biomarkery kliniczne

Moduł Hand Tracking śledzi **21 punktów charakterystycznych dłoni** (precyzja >95%). Kluczowe wskaźniki obliczane przez system to:

- **EAR (Eye Aspect Ratio):** Marker lęku (częstotliwość mrugania >20/min) lub depresji i zmęczenia (EAR < 0.2 przez ponad 0.5 s).
- **MAR (Mouth Aspect Ratio):** Wykrywanie ziewania (zmęczenie) lub nadmiernego otwarcia ust jako biomarkera hiperwygilancji w PTSD.
- **HPE (Head Pose Estimation):** Estymacja kątów rotacji (Yaw, Pitch, Roll) do mierzenia zaangażowania społecznego w terapii ASD.
- **Wykrywanie RRP (Restricted and Repetitive Patterns):** Automatyczna identyfikacja ruchów stereotypowych, takich jak "arm flapping" (machanie rękami) czy kołysanie się, poprzez analizę współrzędnych X-Y nadgarstków i stawów.

---

5. Parametry treningowe i metryki wydajności

Proces uczenia wykorzystuje optymalizator **Adam** (LR=0.0001) oraz funkcję straty **categorical cross-entropy**. Zastosowano rygorystyczne mechanizmy kontrolne: `EarlyStopping` (zatrzymanie po 3 epokach bez poprawy i **przywrócenie najlepszych wag**) oraz `ReduceLROnPlateau` (redukcja LR o współczynnik 0.2).

Zestawienie wyników:

- **Training Accuracy:** 94.33%
- **Validation Accuracy: 95.76%**

Analiza porównawcza (Accuracy):

- **Model Hybrydowy VGG-Transformer (RAF-DB): 95.76%**
- VGG16 (ImageNet): 92.7% (**Top-5 Accuracy**)
- SE-Resnet (RAF-DB): 83.37%
- Transformers w YOLOv8: 81%
- MediaPipe (ruch w czasie rzeczywistym): 97.67%

Wysokie wyniki walidacyjne potwierdzają, że model nie tylko uczy się wzorców, ale wykazuje wysoką zdolność generalizacji, co jest niezbędne w klinicznym wspieraniu decyzji.