# Architektura wielozadaniowego uczenia maszynowego w rozpoznawaniu afektu i generowaniu obrazów

## 1. Ramy teoretyczne analizy afektu

Ewolucja systemów analizy afektu dokonała gwałtownego przejścia od kontrolowanych warunków laboratoryjnych do paradygmatu „in-the-wild", gdzie modele muszą radzić sobie ze zmiennym oświetleniem, ekstremalnymi pozami głowy oraz okluzjami. W kontekście nowoczesnej interakcji człowiek-komputer (HCI), kluczowe jest zintegrowanie różnych reprezentacji emocji, co pozwala na pełniejsze mapowanie stanów psychologicznych użytkownika.

Współczesna analiza afektu operuje na trzech komplementarnych płaszczyznach:

1. **Podejście kategoryczne:** Klasyfikacja do siedmiu podstawowych kategorii (złość, wstręt, strach, szczęście, smutek, zaskoczenie oraz stan neutralny).

2. **Jednostki Działania (AUs):** Oparte na systemie Facial Action Coding System (FACS), opisują atomowe ruchy mięśni twarzy. W przeciwieństwie do kategorii, AUs nie wykluczają się wzajemnie, co pozwala na precyzyjny opis złożonych stanów mimicznych.

3. **Model wymiarowy (Valence-Arousal):** Mapuje emocje na ciągłą przestrzeń 2D (Emotion Wheel). Oś **Valence** (Walencja) określa stopień przyjemności (od „unpleasant" do „pleasant"), natomiast **Arousal** (Pobudzenie) mierzy intensywność (od stanów pasywnych „mild/sleepy" do aktywnych „intense/excited").

Poniższa tabela mapuje kluczowe jednostki działania wykorzystywane w analizie:

| Jednostka działania (AU) | Opis fizyczny (Angielski) | Opis fizyczny (Polski) |
|---|---|---|
| **AU 1** | Inner brow raiser | Uniesienie wewnętrznych części brwi |
| **AU 2** | Outer brow raiser | Uniesienie zewnętrznych części brwi |
| **AU 4** | Brow lowerer | Obniżenie brwi |
| **AU 6** | Cheek raiser | Uniesienie policzków |
| **AU 12** | Lip corner puller | Podciągnięcie kącików ust |
| **AU 15** | Lip corner depressor | Obniżenie kącików ust |
| **AU 20** | Lip stretcher | Rozciągnięcie ust |
| **AU 25** | Lips part | Rozchylenie warg |

**So What?:** Precyzyjne kodowanie AUs jest kluczowe, ponieważ kategoryczne modele nie są w stanie uchwycić mikroekspresji i płynnych przejść między stanami. Przykładowo, klatka o ujemnej walencji (-0.69) i wysokim pobudzeniu (0.92) zazwyczaj wiąże się z aktywacją AU 1, 4 i 25, co sygnalizuje strach – informacja ta jest niemożliwa do pełnego odczytania bez analizy wymiarowej i atomowej.

# 2. Metodologia adnotacji i charakterystyka zbioru danych Aff-Wild

Zbiór Aff-Wild to pierwszy wielkoskalowy zbiór typu in-the-wild (1,2 mln klatek). Dla celów uczenia wielozadaniowego (MTL) wyselekcjonowano podzbiory wymagające rygorystycznej weryfikacji.

## Charakterystyka ilościowa i jakość danych

Proces adnotacji angażował trzech ekspertów pracujących niezależnie. Aby zapewnić najwyższą czystość etykiet w procesie uczenia głębokiego, zastosowano filtr **100% zgodności** między wszystkimi koderami. Wymóg ten znacząco zredukował finalną liczbę klatek, podnosząc jednak ich wartość informacyjną:

- **Podzbiór AUs:** Z pierwotnych 234 000 klatek, tylko **139 298** zawierało co najmniej jedną aktywną jednostkę działania zatwierdzoną przez wszystkich ekspertów.
- **Podzbiór Ekspresji Podstawowych:** Z 288 000 klatek wyselekcjonowano **115 640** spełniających wymóg pełnej zgodności adnotacji.

Analiza rozkładu danych potwierdza występowanie zjawiska *imbalance* – dominują stany „Neutral” oraz „Sadness”, podczas gdy ekspresje o wysokiej intensywności, jak „Fear”, są rzadsze, co odzwierciedla naturalne zachowania ludzkie w rzeczywistych nagraniach.

**So What?:** Tak rygorystyczna selekcja danych eliminuje szum informacyjny, co jest fundamentem stabilności dla modeli MTL. Różnorodność warunków (oświetlenie, pozy) w bazie Aff-Wild wymusza na modelach naukę cech odpornych na zakłócenia, co jest niemożliwe w warunkach laboratoryjnych.

# 3. Architektury GAN w klasyfikacji i generowaniu obrazów

Zastosowano Generative Adversarial Networks (GAN) w trybie półnadzorowanym, gdzie dyskryminator pełni potrójną rolę: klasyfikatora AUs, regresora VA oraz detektora autentyczności obrazu.

### Specyfikacja techniczna konfiguracji

- **Konfiguracja 1 (SSGAN):** Generuje obrazy **32x32x3**. Wykorzystuje cztery warstwy conv2d transpose z normalizacją wsadową i aktywacjami ReLU/LeakyReLU.
- **Konfiguracja 2 (DCGAN):** Rozdzielczość wyjściowa **96x96x3**. Wykorzystuje głębszą architekturę z filtrami 5x5 lub 7x7, co jest niezbędne do uchwycenia subtelnych detali mimicznych (np. zmarszczek przy AU 6) w warunkach in-the-wild.

Dyskryminator posiada **11 wyjść**: 2 liniowe dla regresji VA, 8 z funkcją sigmoid dla klasyfikacji wieloetykietowej AUs oraz 1 wyjście dla klasy „fake". Generator przetwarza 100-wymiarowy szum Z w realistyczne wizerunki, zdolne do modyfikacji atrybutów takich jak zarost, okulary czy płeć.

**So What?:** Uczenie półnadzorowane rozwiązuje problem deficytu etykiet w warunkach in-the-wild. Syntetyczne próbki generowane przez sieć służą jako zaawansowana augmentacja danych, która zachowuje specyficzny rozkład cech rzeczywistych, stabilizując tym samym proces klasyfikacji AUs.

# 4. Model CNN-RNN z mechanizmem uwagi (attention layer)

Do analizy sekwencyjnej wykorzystano architekturę CNN-RNN opartą na modelu AffWildNet (VGGFACE-GRU), wzbogaconą o warstwę uwagi.

Model przetwarza sekwencje klatek (Frame 0 do Frame t), gdzie cechy ekstrahowane przez CNN trafiają do dwuwarstwowej sieci GRU. Kluczowym innowacją jest **warstwa uwagi (Attention Layer)** nałożona na stany ukryte RNN dla sekwencji **80 kolejnych klatek**. Mechanizm ten pozwala modelowi selektywnie ważyć istotność klatek, skupiając się na momentach szczytowych ekspresji (*peak*) i ignorując klatki przejściowe o niskim ładunku afektywnym.

Wyjście modelu stanowi 9 kanałów: 2 liniowe (VA) oraz 7 z funkcją softmax dla kategorii ekspresji.

**So What?:** Emocje są procesami dynamicznymi. Modelowanie zależności czasowych z mechanizmem uwagi pozwala na poprawną interpretację kontekstu – ten sam ruch mięśni może oznaczać co innego w fazie narastania (*onset*), niż w fazie wygasania (*offset*) emocji.

# 5. Funkcje strat i strategie optymalizacji wielozadaniowej

Kluczem do sukcesu MTL jest matematyczne zbalansowanie celów regresji i klasyfikacji.

### Analiza matematyczna funkcji straty

- **Regresja VA:** Wykorzystano **Concordance Correlation Coefficient (CCC)**. Jest to metryka nadrzędna względem MSE, ponieważ łączy współczynnik korelacji Persona (relacja liniowa) z błędem średniokwadratowym (przesunięcie średniej od linii 45 stopni). Pozwala to na uniknięcie błędów przesunięcia, które są częste w tradycyjnym MSE.
- **Klasyfikacja AUs:** Standardowa strata Cross-Entropy.
- **Generator GAN:** Łączy stratę logarytmiczną wyjścia dyskryminatora ze stratą rekonstrukcji **Huber Loss** przy progu \delta = 1, co stabilizuje wczesną fazę uczenia.

Całkowita funkcja straty definiowana jest jako: L_{total} = \alpha \cdot L'_{VA} + \beta \cdot L_{basic_expressions} Wagi \alpha i \beta (optymalnie po 0.5) pozwalają uniknąć dominacji jednego zadania nad drugim.

**So What?:** Wykorzystanie CCC zamiast MSE w MTL pozwala modelowi na lepsze „zrozumienie" trendów w intensywności afektu, co w połączeniu z odpowiednim ważeniem zadań prowadzi do wzajemnego wzmacniania się estymatorów.

# 6. Ewaluacja wyników i korzyści z interakcji zadań

Eksperymenty dowodzą, że współzależność zadań (np. korelacja wysokiego pobudzenia z aktywacją AU 4 przy złości) realnie podnosi metryki modelu.

### Porównanie wydajności

Wyniki testów na zbiorze Aff-Wild wykazują znaczącą przewagę uczenia wielozadaniowego:

- Model **Single Task** (tylko VA, \alpha=1, \beta=0) osiągnął wynik CCC dla walencji na poziomie **0.579**.
- Model **Multi-Task** (\alpha=0.5, \beta=0.5) uzyskał wynik **0.616** CCC dla walencji oraz najwyższe wskaźniki F1-score.

W przypadku dyskryminatora GAN, funkcja straty oparta na MSE okazała się paradoksalnie nieco skuteczniejsza w surowej estymacji VA (CCC 0.484 dla walencji), jednak to model MTL z CCC wykazał lepszą stabilność przy jednoczesnej klasyfikacji AUs. Jakościowa analiza obrazów potwierdza, że generator skutecznie transferuje cechy „wild" (np. dodanie okularów lub zarostu), co świadczy o robustności przestrzeni latentnej modelu.

**So What?:** Udowodniono, że zadania rozpoznawania afektu nie są niezależne. Informacje o atomowych ruchach mięśni (AUs) dostarczają sieci cennych wskazówek do precyzyjnego

określenia położenia w przestrzeni Valence-Arousal, co potwierdza przewagę architektury MTL.

# 7. Wnioski i kierunki rozwoju

Raport potwierdza, że zaawansowane architektury MTL, łączące GAN i CNN-RNN z mechanizmem uwagi, stanowią nowy standard w analizie afektu in-the-wild.

**Główne wnioski:**

1. **Synergia MTL:** Jednoczesne uczenie VA i ekspresji/AUs zapewnia lepszą generalizację i wyższą dokładność (skok CCC z 0.579 do 0.616).
2. **Rola Uwagi:** Mechanizm uwagi w sekwencjach 80 klatek pozwala na skuteczne filtrowanie szumu informacyjnego w dynamicznych wideo.
3. **Potencjał GAN:** Półnadzorowane uczenie zwiększa odporność dyskryminatora na zróżnicowane warunki środowiskowe.

**Kierunki rozwoju:** Przyszłe prace powinny koncentrować się na dalszej optymalizacji mechanizmów uwagi oraz wykorzystaniu generatorów GAN do tworzenia syntetycznych, rzadkich przypadków emocjonalnych (np. ekstremalny strach), co pozwoli na jeszcze lepsze zrównoważenie zbiorów treningowych.

**So What?:** Integracja generowania i klasyfikacji otwiera drogę do budowy empatycznych systemów AI, które nie tylko pasywnie rozpoznają afekt, ale potrafią go aktywnie i realistycznie symulować w interakcji z człowiekiem.