МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ. ІГОРЯ СІКОРСЬКОГО”
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ

НЕЙРОННІ МЕРЕЖІ ДЛЯ РОЗПІЗНАВАННЯ ЕМОЦІЙ У ВІДЕО ТА АУДІО ПОТОКАХ

Роботу виконали:

Антон Щербина

Вадим Коршунов

Науковий керівник:

Юрій Тимошенко

АКТУАЛЬНІСТЬ

- Неперервна діагностика емоційного стану
- Додаткова інформація для маркетингових досліджень та рекомендаційних систем
- Адаптація інтерфейсів взаємодії з комп’ютером

МЕТА, ОБ'ЄКТ І ПРЕДМЕТ ДОСЛІДЖЕННЯ

- Мета. Дослідити методи глибокого навчання для вирішення задачі класифікації емоції з відео та аудіо в режимі реального часу, імплементувати та протестувати найкращі моделі.
- Об'єкт. Розпізнавання людських емоцій за відео та аудіо в режимі реального часу.
- Предмет. Методи глибокого навчання, які дозволяють вирішити проблему розпізнавання емоцій.

КЛАСИЧНІ МЕТОДИ

- Ручне виділення ознак
- Необхідне знання доменої області
- Низька узагальнююча здатність

ГЛИБОКЕ НАВЧАННЯ

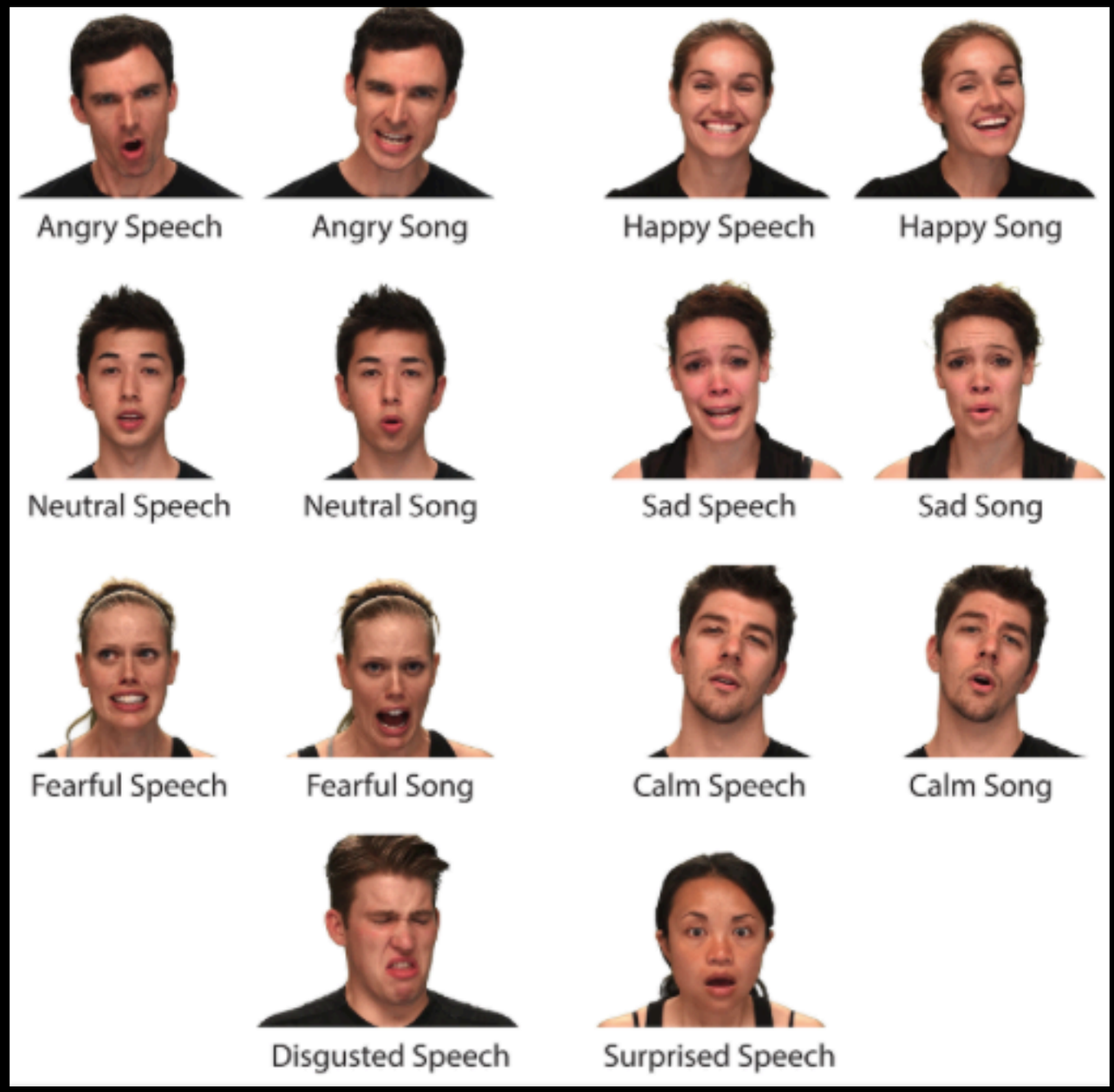
- Не потребують ручного виділення ознак
- Дозволяють працювати з даними великої розмірності
- Можуть навчатись end-to-end

ЕТАПИ РОБОТИ

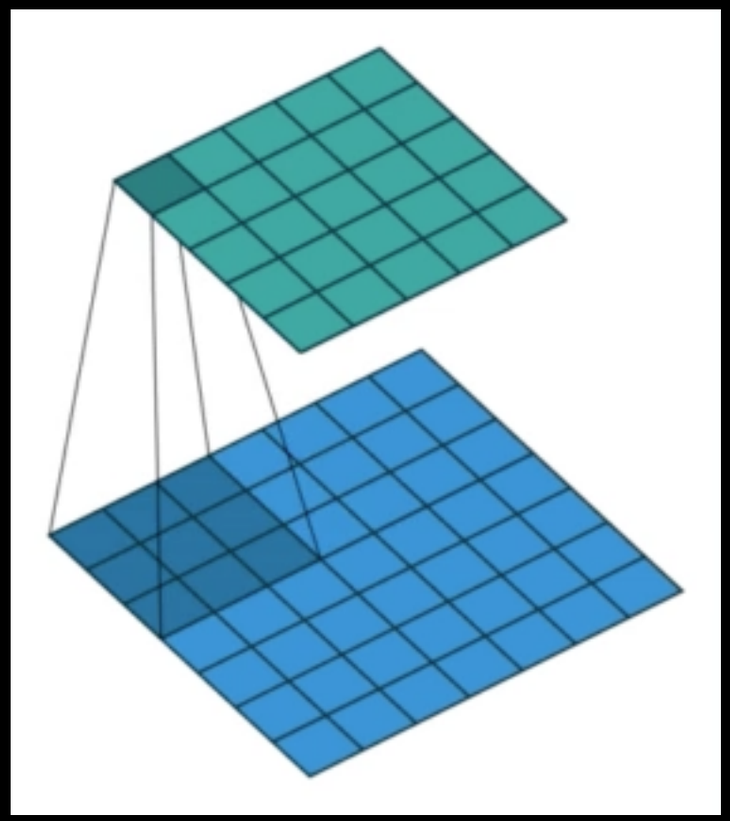
- Збір даних та їх передобробка
- Побудова архітектур моделей
- Оптимізація параметрів моделей
- Оцінка генералізуючої здатності кожної з моделей
- Ансамблювання моделей

ДАНІ

- 4900 високоякісних відео та аудіо файлів
- 24 актори
- 8 емоцій
- 2 типи інтенсивності
- 2 типи мовлення
- 2 типи повідомлення



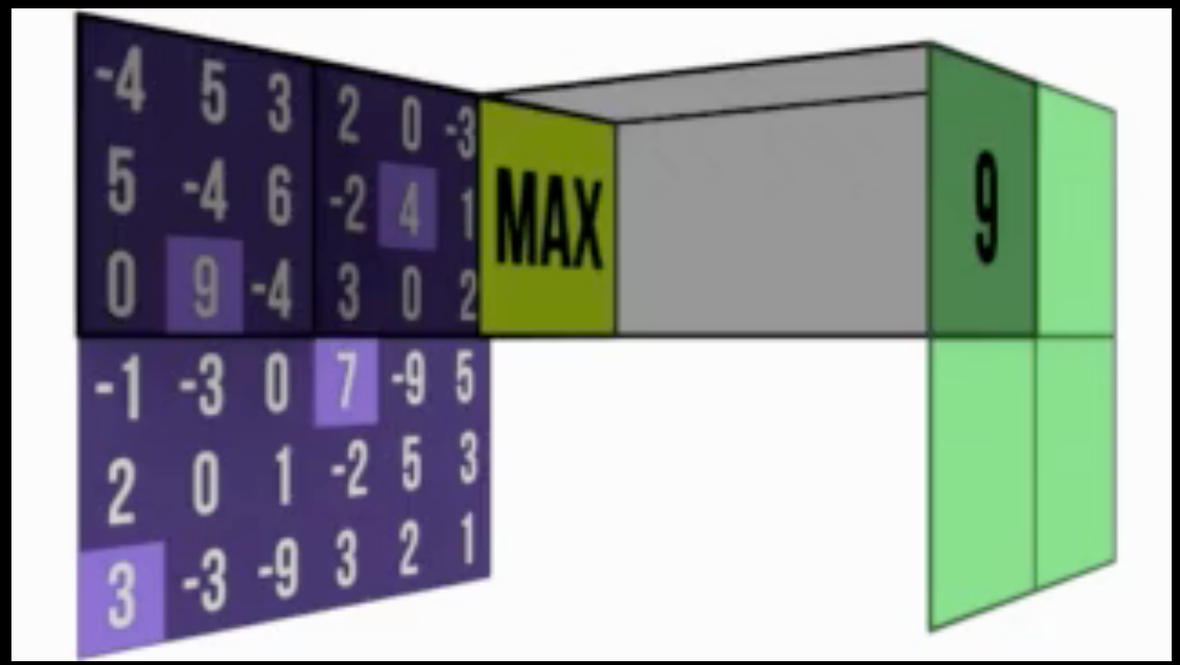
ЗГОРТКА

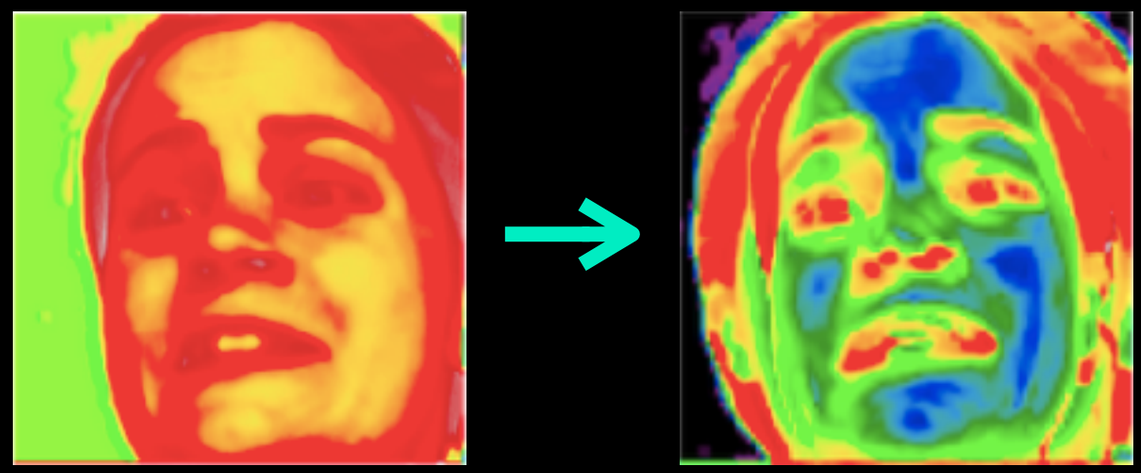


$$Output[i][j] = \sum_{k=1}^{m}\sum_{l=1}^{n} Image(i+k-1, j+l-1) * Kernel(k,l)$$

$$i = 1,...,M-m+1, j = 1...N-n+1$$

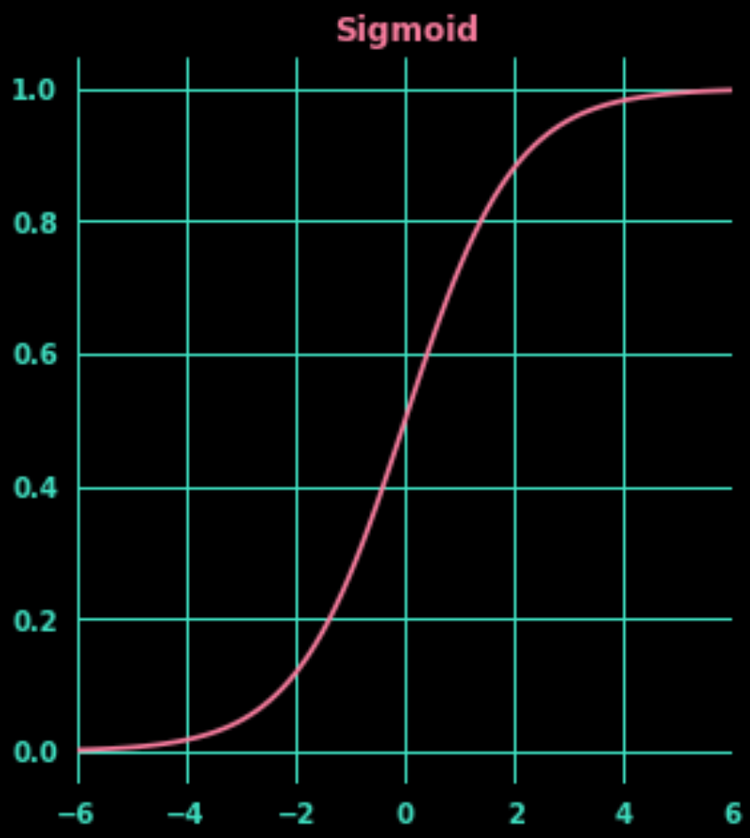
MAX POOLING

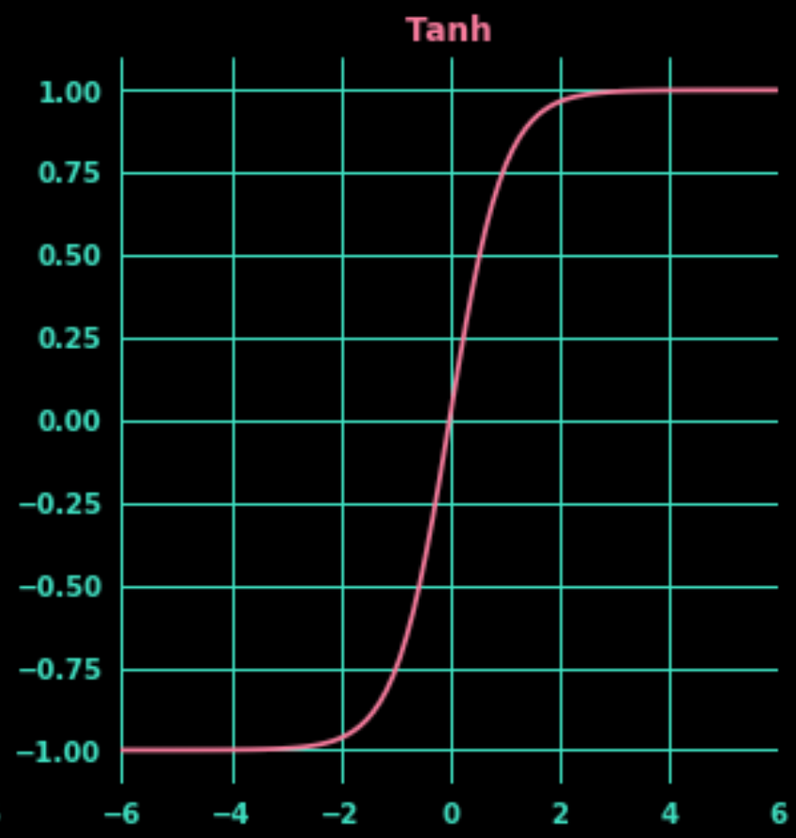


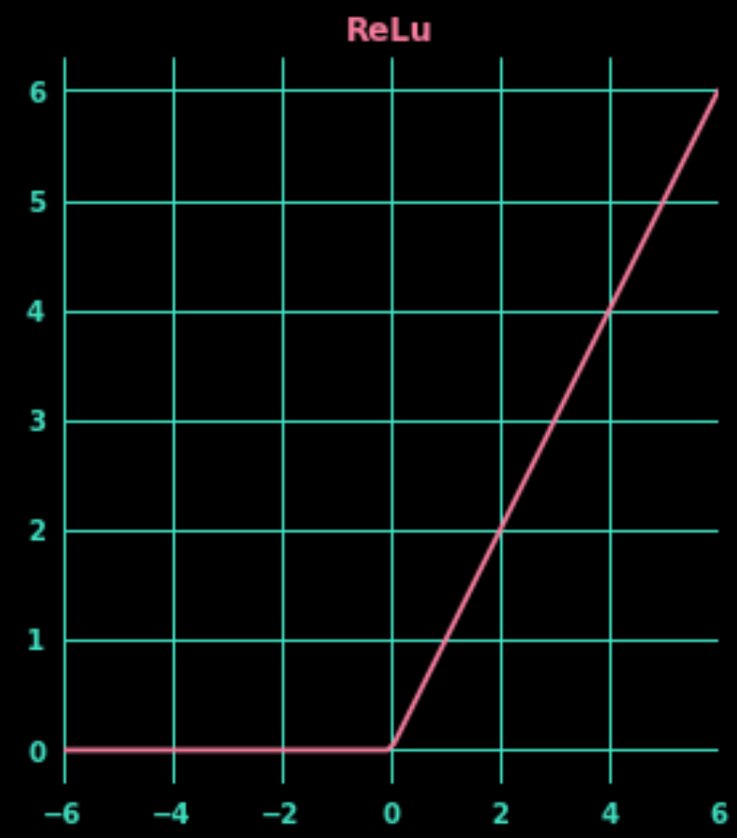


$$Output[i][j] = max\{Image(i+k-1, j+l-1), k = 1...m, l = 1...n, i = 1,...,M-m+1, j = 1...N-n+1\}$$

НЕЛІНІЙНОСТІ





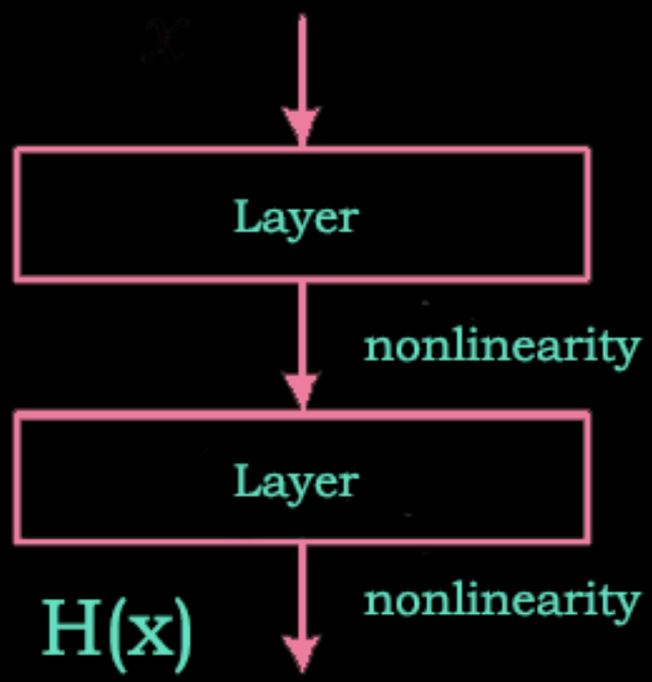


$$f(x) = \frac{1}{1 + e^{-x}}$$

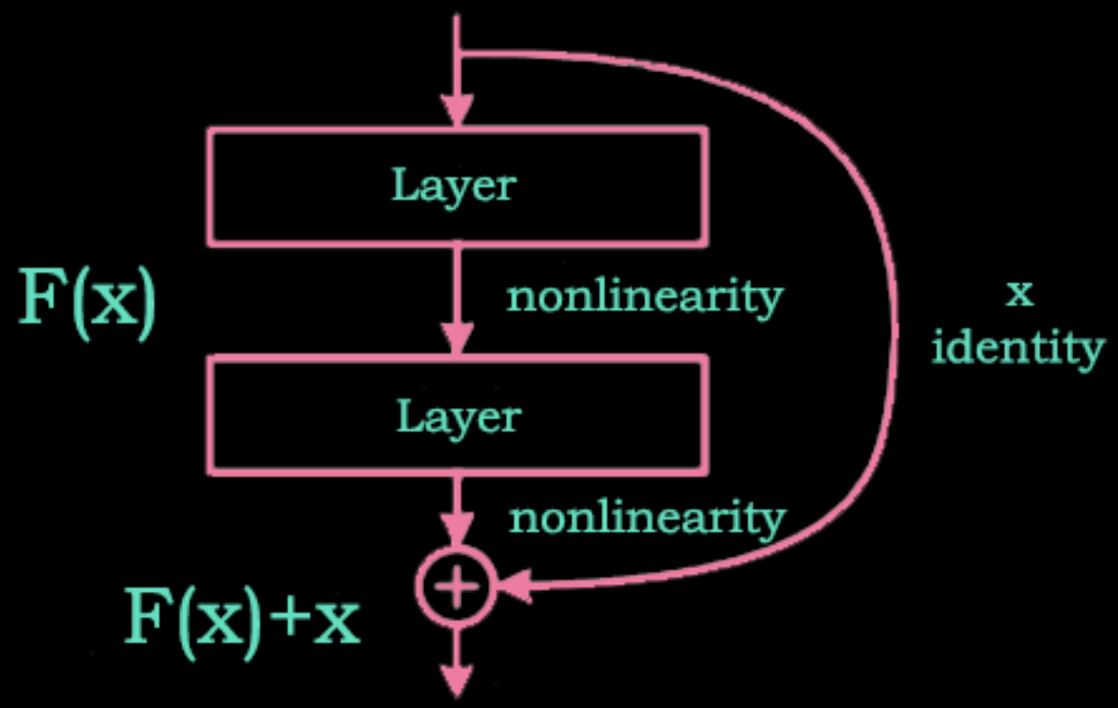
$$f(x) = \frac{2}{1 + e^{-2x}} - 1$$

$$f(x) = max(0,x)$$

ЗАЛИШКОВІ МЕРЕЖІ

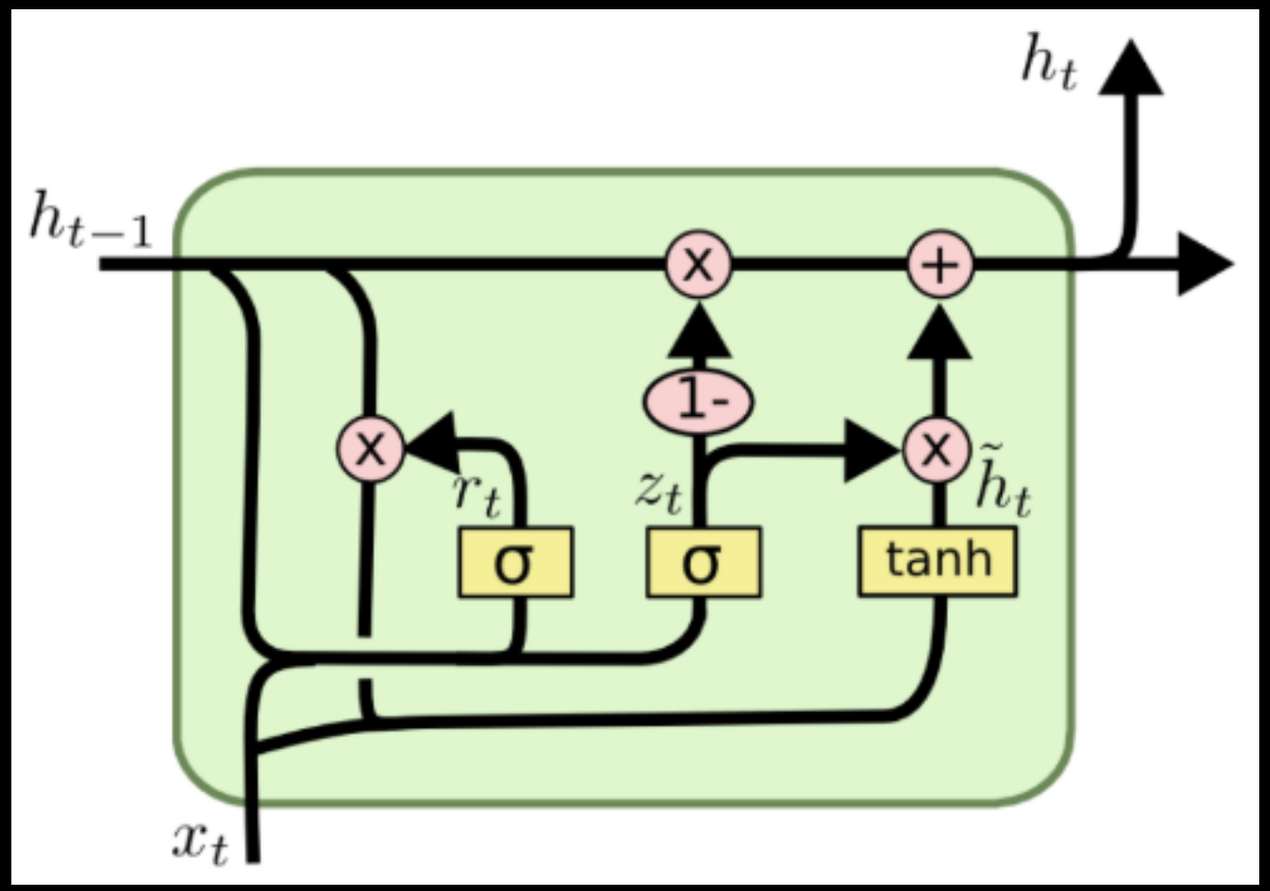


Звичайний блок



Залишковий блок

GATED RECURRENT UNIT



$$z_t = \sigma(W_z[h_{t-1}, x_t])$$

$$r_t = \sigma(W_r[h_{t-1}, x_t])$$

$$\hat{h}_t = tanh(W[r_t * h_{t-1,x_t}]$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t$$

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

РОЗПОДІЛ ЙМОВІРНОСТЕЙ

$$softmax(z) = \begin{pmatrix} \frac{e^{z1}}{\sum_{i=1}^{n} e^{z_i}} \\ \cdots \\ \frac{e^{zn}}{\sum_{i=1}^{n} e^{z_i}} \end{pmatrix} \quad \begin{pmatrix} 1.2 \\ 0.9 \\ 0.4 \end{pmatrix} \xrightarrow{softmax} \begin{pmatrix} 0.46 \\ 0.34 \\ 0.20 \end{pmatrix}$$

НАБЛИЖЕННЯ РОЗПОДІЛІВ

- Нехай є $< x_1, l_1 > , \ldots , < x_N, l_N >$ - навчальні приклади, де x - вхідний елемент, l - його відповідний клас
- Відомий модельований розподіл класів для кожного вхідного елементу та його реальний розподіл класів:

$$\forall x_i : Model(x_i; \theta) = (p_1(x_i; \theta), \ldots , p_n(x_i; \theta))$$
$$\forall x_i : Real(x_i; \theta) = (0_1, \ldots , 1_{l_i}, \ldots , 0_n)$$

- Як наблизити розподіл (q) іншим модельованим (p)?

ПЕРЕХРЕСНА ЕНТРОПІЯ

- Нехай є модельований розподіл p та справжній розподіл q
- Введемо міру “відстані” між розподілами - перехресну ентропію:

$$H(p,q) = -\sum_{i=1}^{n} q_i log(p_i) = H(q) + D_{KL}(q||p)$$

- На основі перехресної ентропії вводиться функція втрат для багатокласової класифікації на всьому наборі даних:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} q_j(x_i) log(p_j(x_i;\theta)) + \alpha \parallel \theta \parallel^2$$

ОПТИМІЗАЦІЯ

- Нехай ξ - дискретна випадкова величина, яка дорівнює рівномірному розподілу навчальних прикладів:

$$\mathbb{P}[\xi = x_i] = \frac{1}{N}$$

- Береться стохастична апроксимація функції втрат:

$$L^*(\theta) = \frac{1}{M}\sum_{i=1}^{M} H(p(\xi_i;\theta), q(\xi_i)) + \alpha||\theta||^2$$

- Апроксимація функції втрат та її градієнт - незміщені статистичні оцінки:

$$\mathbb{E}L^*(\theta) = L(\theta) \qquad \mathbb{E}\frac{\overrightarrow{\partial}}{\partial\theta}L^*(\theta) = \frac{\overrightarrow{\partial}}{\partial\theta}L(\theta)$$

ADAM

$$grad_t \leftarrow grad_\theta L_t^*(\theta_{t-1})$$

$$M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1) grad_t$$

$$V_t \leftarrow \beta_2 V_{t-1} + (1-\beta_2) grad_t^2$$

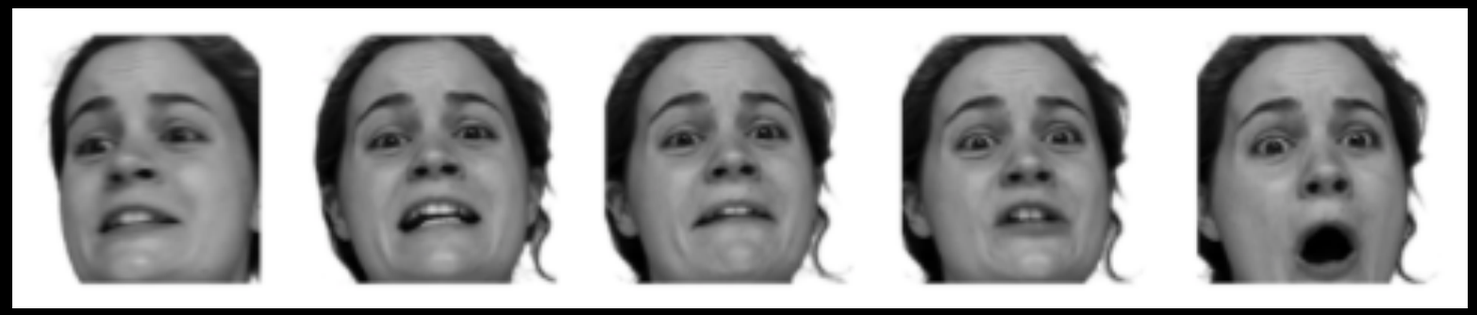
$$\tilde{M}_t \leftarrow \frac{M_t}{1-\beta_1^t}$$

$$\tilde{V}_t \leftarrow \frac{V_t}{1-\beta_2^t}$$

$$\theta_t \leftarrow \theta_t - \alpha \cdot \frac{\tilde{M}_t}{\tilde{V}_t^{\frac{1}{2}} + \epsilon}$$

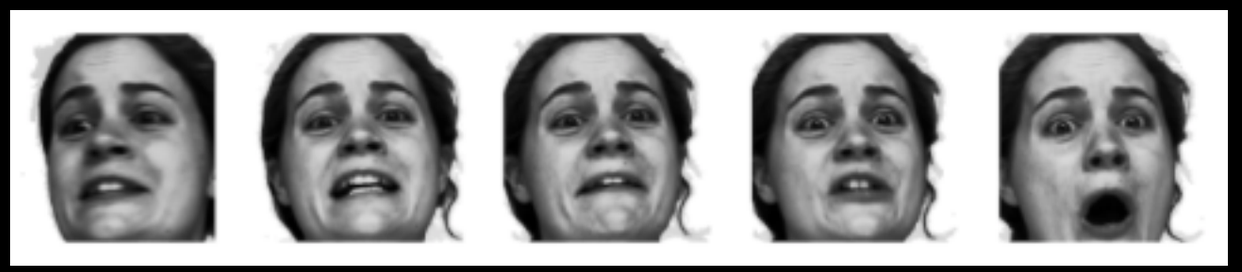
ОБРОБКА ВІДЕО

- Передискретизація вхідного відео
- Перетворення кольорових кадрів на чорнобілі
- Виділення облич людей з зображень
- Масштабування облич до однакового розміру

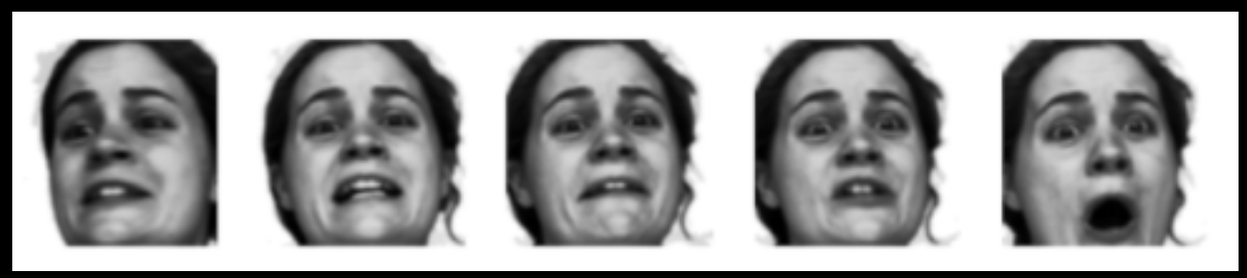


ОБРОБКА ВІДЕО

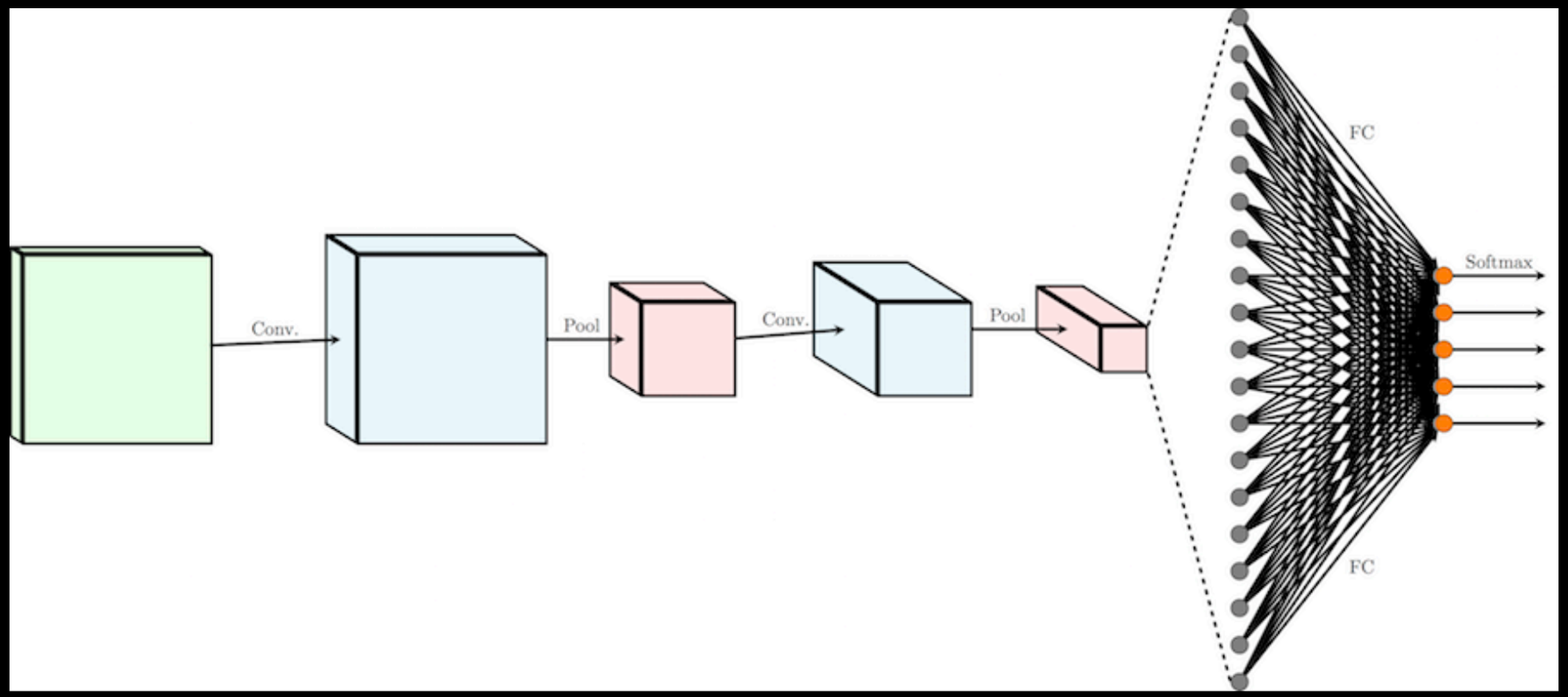
- Вирівнювання гістограми яскравостей



- Гауссове згладжування

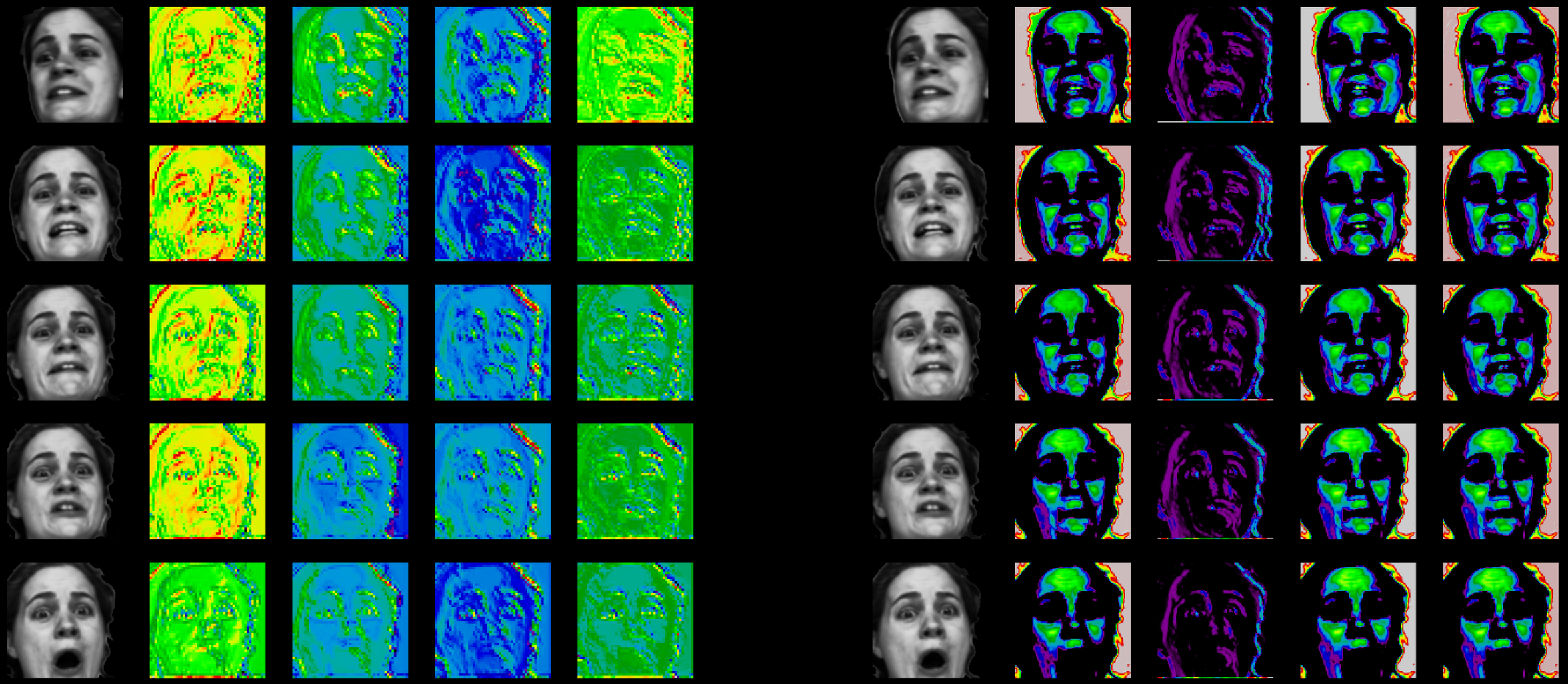


АРХІТЕКТУРА ДЛЯ ВІДЕО

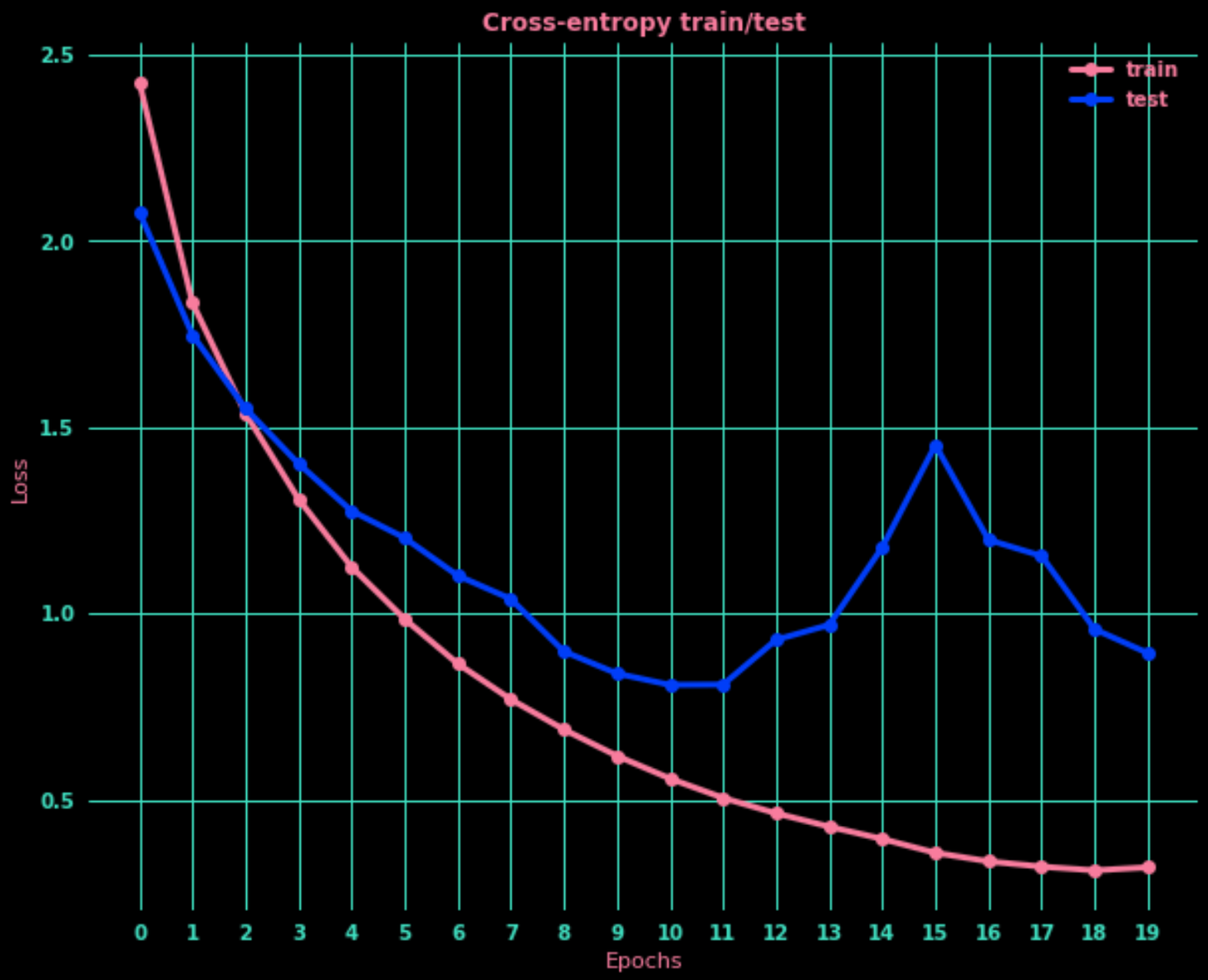


- Між блоками є також шари нелінійностей (ReLU) та технічні шари (BatchNormalization, Dropout)
- Замість звичайних блоків використовуються залишкові блоки

КАРТИ ОЗНАК

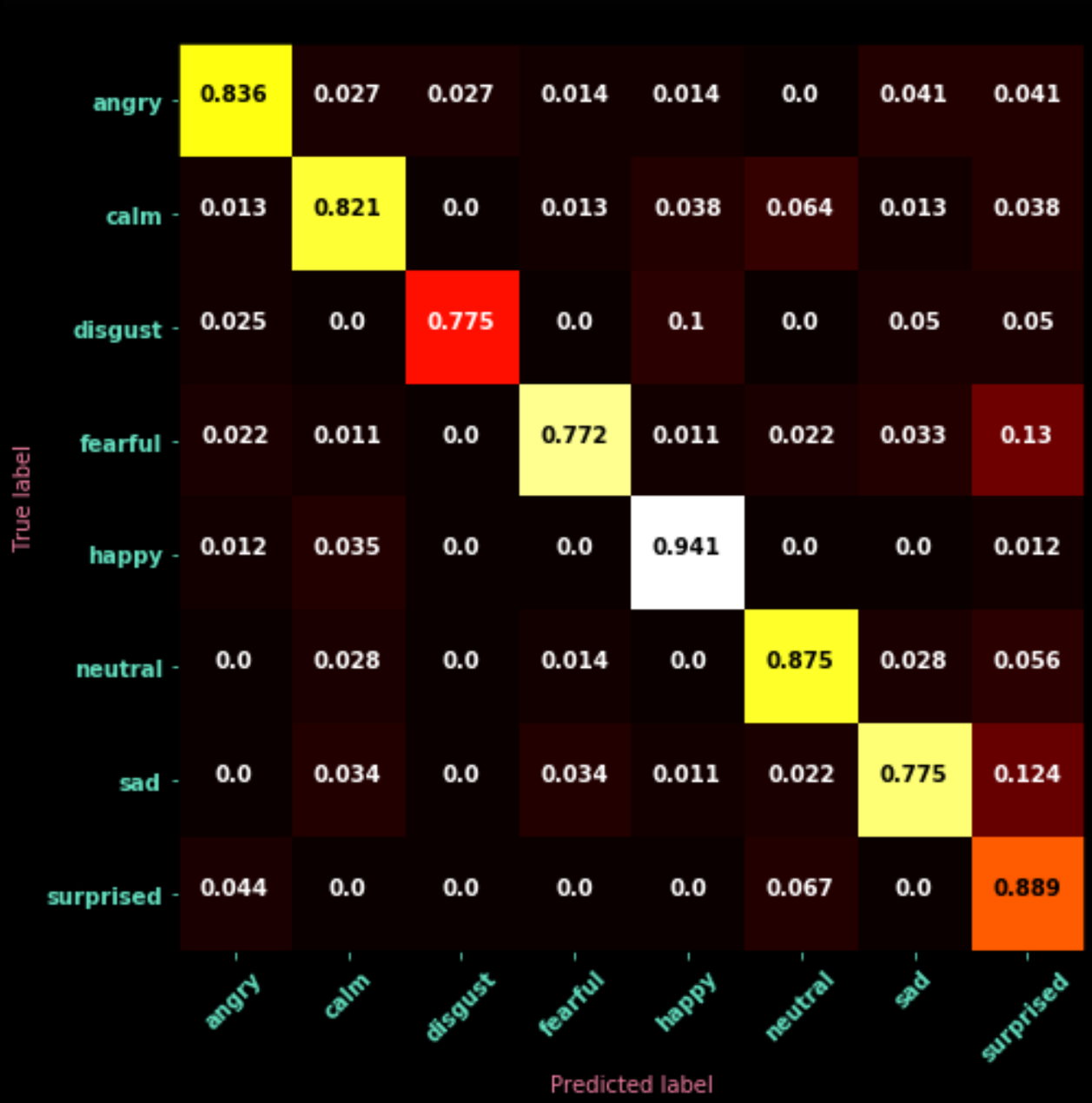


РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ



- Тренувальна вибірка - 0.95
- Тестова вибірка - 0.8

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ



$$C_{ij} = \mathbb{P}[l^* = j \mid l = i]$$

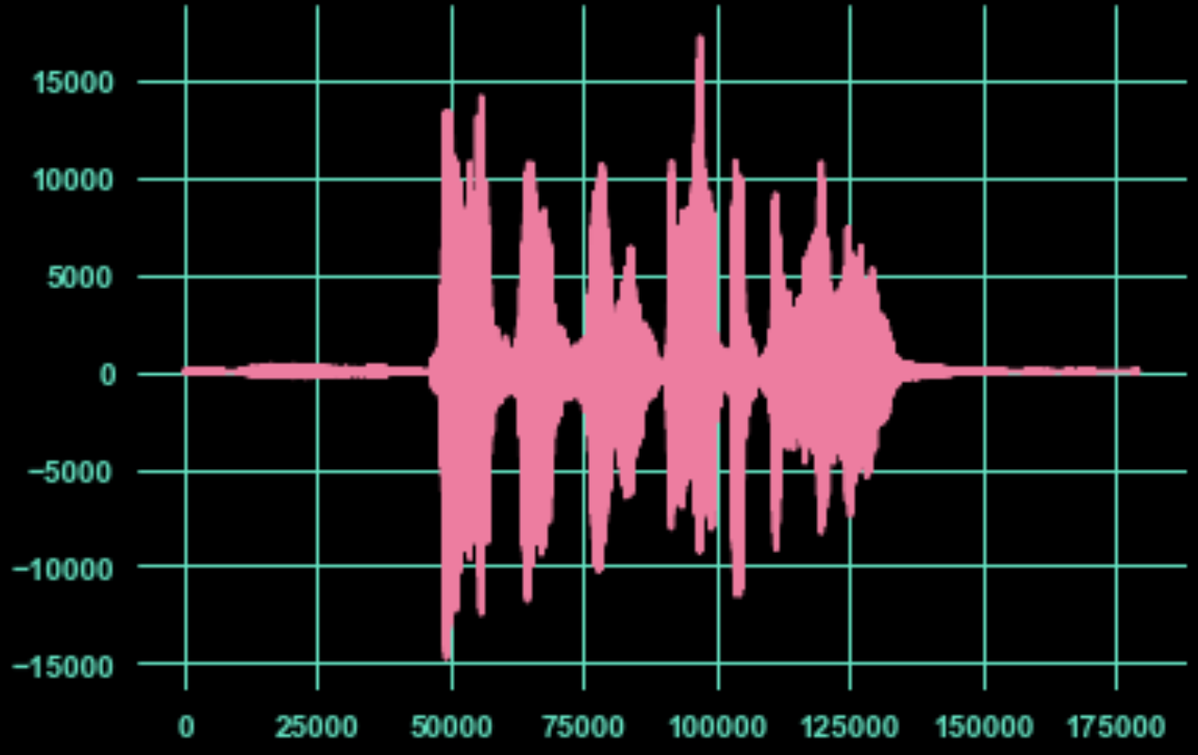
$$C^*_{ij} = \frac{C_{ij}}{\sum_{j=1}^{n} C_{ij}}$$

$$Accuracy = \frac{tr(C)}{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}}$$

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

ЕКСТРАКЦІЯ ОЗНАК

- Просодичні характеристики (енергія, f0)
- Спектральні характеристики (MFCC, спектрограми)



ЕКСТРАКЦІЯ ОЗНАК

- Фреймування вхідного аудіо:

$$\{x_t\}, t = 0,...,T$$

- Зглажування фрейму вікном Хеммінга:

$$x_t(n) = x_t(n) * w(n), n = 0,...,N-1$$

$$w(n) = 0.53836 - 0.46164cos(\frac{2\pi n}{N-1})$$

- Перетворення Фур'є

$$X(k) = \sum_{n=0}^{N-1} x_t(n)e^{-\frac{2\pi i}{N-1}nk}, k = 0,...,N-1, t = 0,...,T$$

ЕКСТРАКЦІЯ ОЗНАК

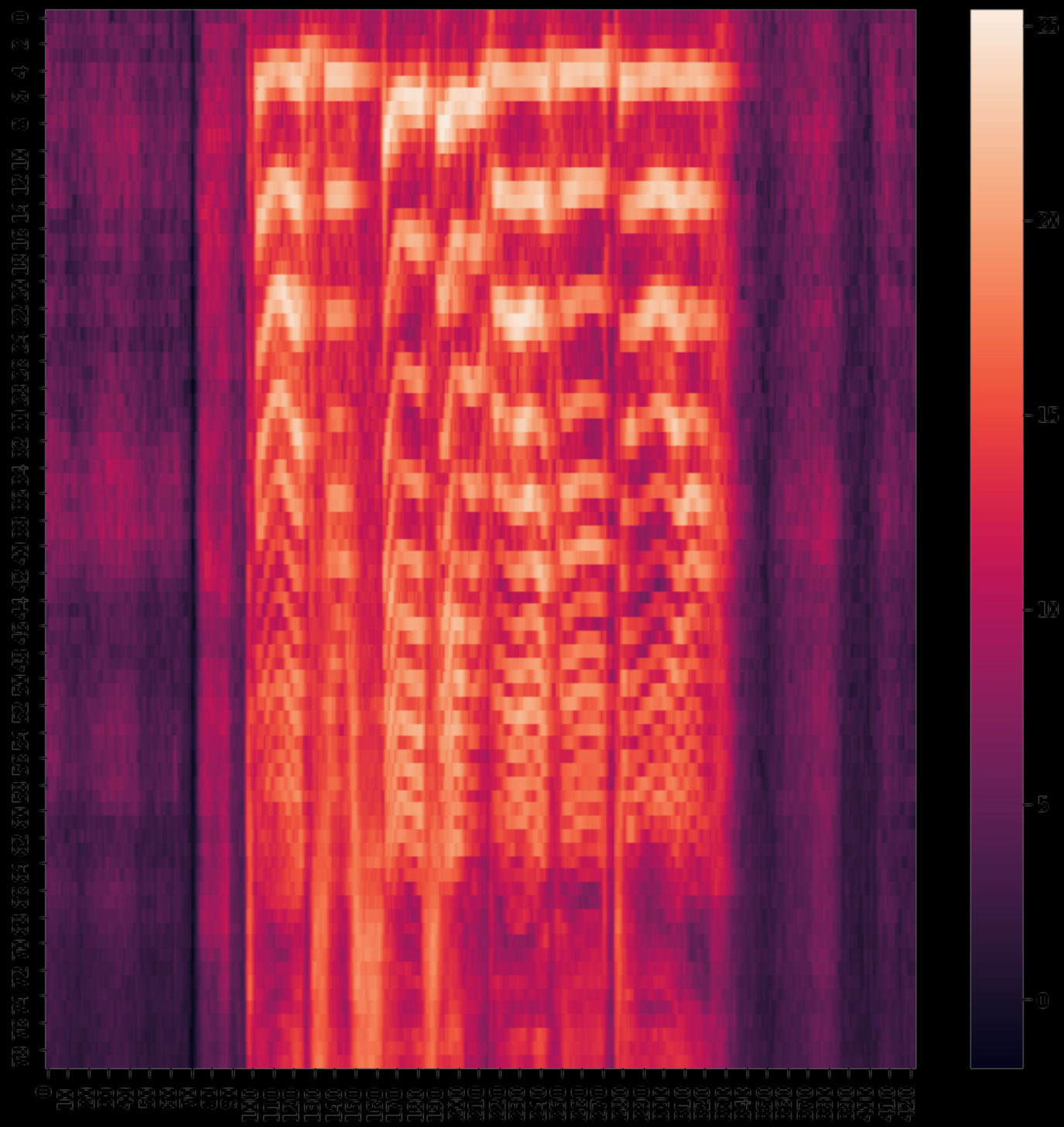
- Створюємо фільтри:

$$H_m = \begin{cases} 0, k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \le k < f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, f(m) \le k \le f(m+1) \\ 0, k > f(m+1) \end{cases}$$

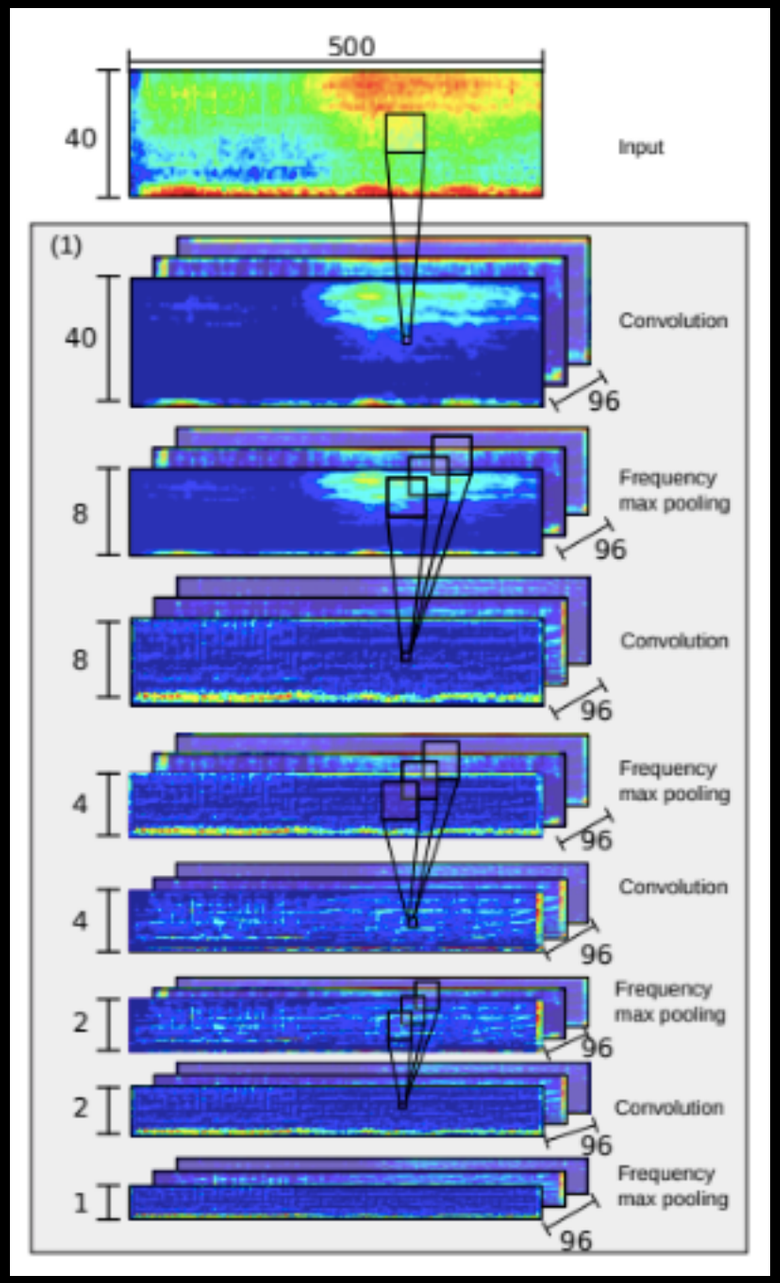
- Отримуємо енергії для кожного вікна:

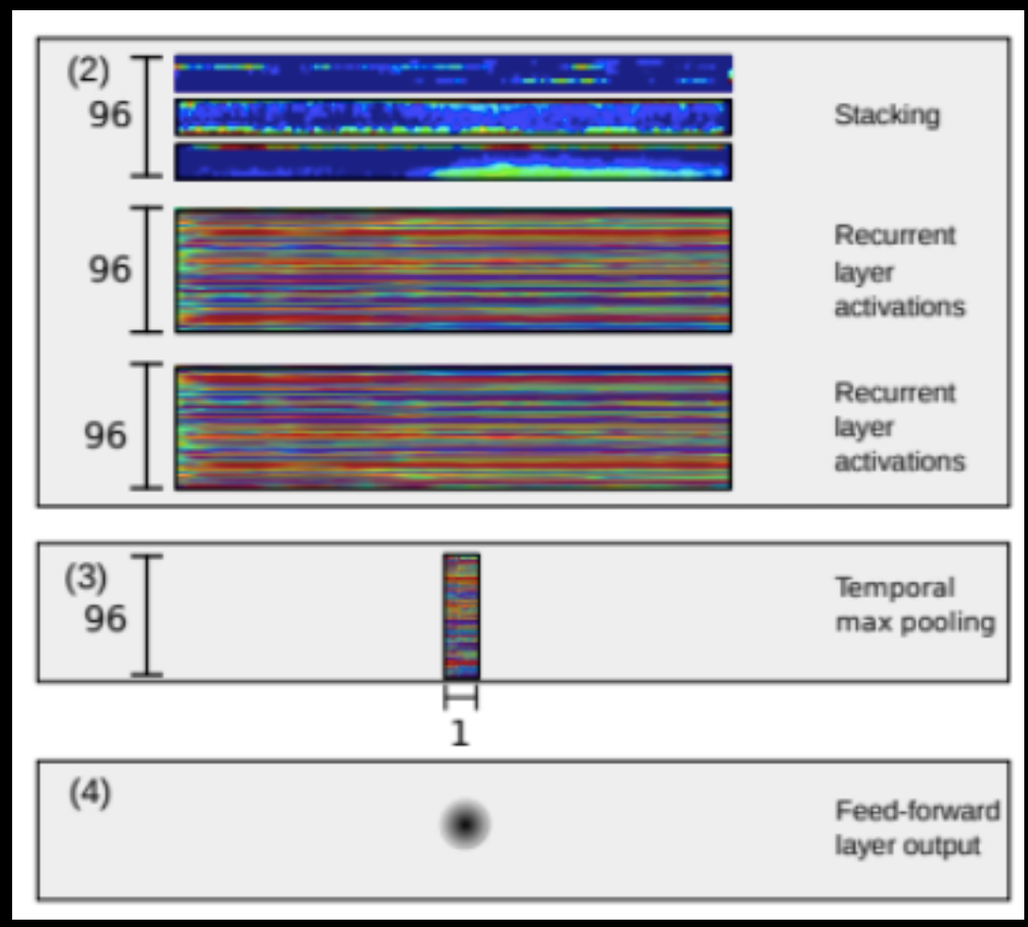
$$S_t(m) = ln(\sum_{k=0}^{N-1} \mid X_t(k) \mid^2 H_m(k)), m = 0,...,M$$

ЕКСТРАКЦІЯ ОЗНАК



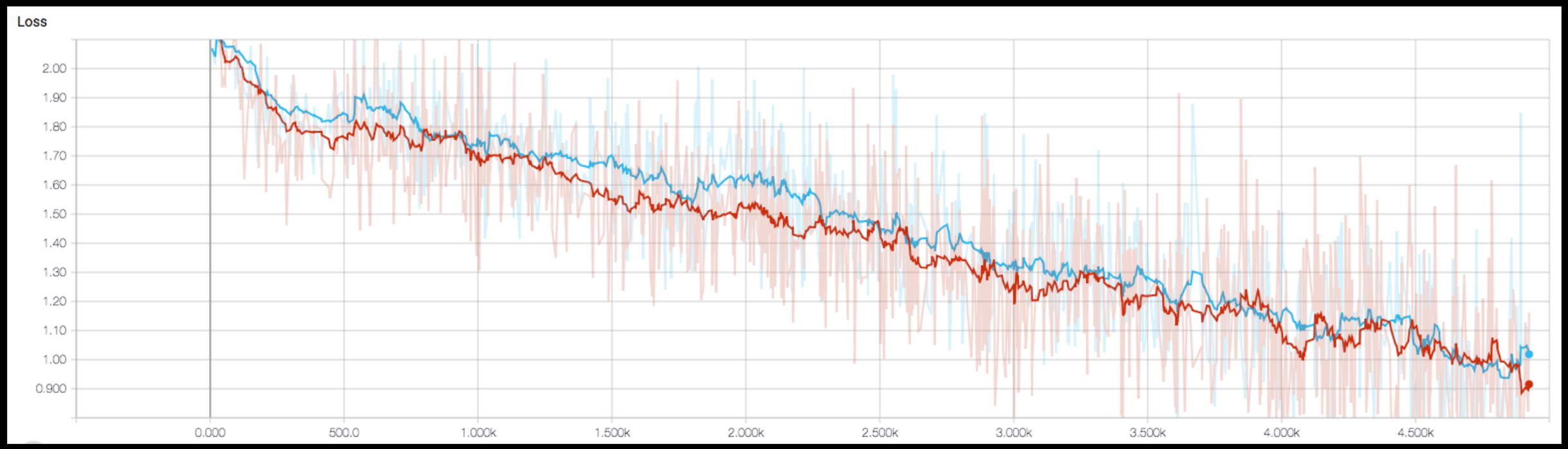
МОДЕЛЬ CRNN





https://ieeexplore.ieee.org/document/8081508/

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ



- Тренувальна вибірка - 0.81
- Тестова вибірка - 0.75

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

happy.wav

- Prob 0.00906, angry
- Prob 0.00505, calm
- Prob 0.00111, disgust
- Prob 0.00062, fearful
- Prob 0.75538, happy
- Prob 0.20468, neutral
- Prob 0.02045, sad
- Prob 0.00364, surprised

sad.wav

- Prob 0.00739, angry
- Prob 0.09296, calm
- Prob 0.04937, disgust
- Prob 0.00393, fearful
- Prob 0.13521, happy
- Prob 0.11970, neutral
- Prob 0.48490, sad
- Prob 0.10654, surprised

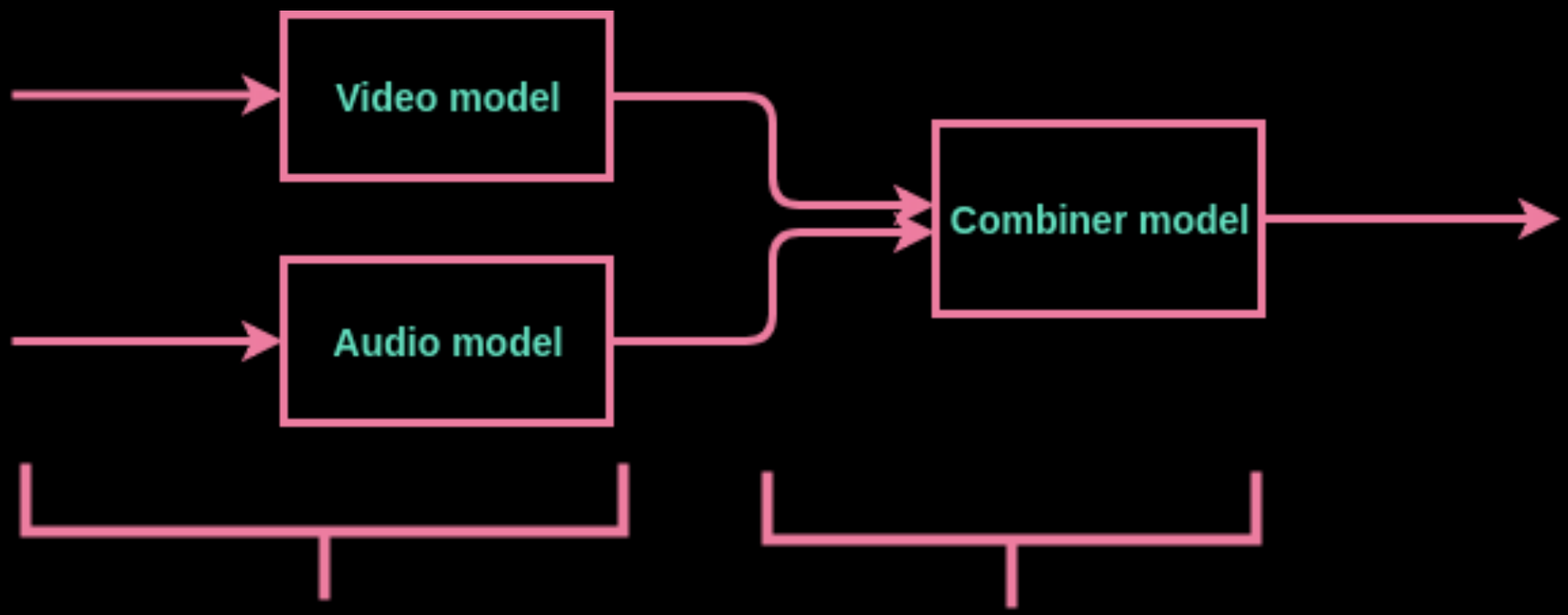
fear.wav

- Prob 0.00908, angry
- Prob 0.00434, calm
- Prob 0.00232, disgust
- Prob 0.80547, fearful
- Prob 0.11106, happy
- Prob 0.00208, neutral
- Prob 0.02724, sad
- Prob 0.03841, surprised

angry.wav

- Prob 0.44661, angry
- Prob 0.00974, calm
- Prob 0.02614, disgust
- Prob 0.04990, fearful
- Prob 0.21963, happy
- Prob 0.04308, neutral
- Prob 0.19184, sad
- Prob 0.01306, surprised

АНСАМБЛЮВАННЯ



РЕЗУЛЬТАТИ АНСАМБЛЮВАННЯ

- Тренувальна вибірка - 0.96
- Тестова вибірка - 0.86

ВИСНОВКИ

- Досліджено основні принципи роботи нейронних мереж
- Розглянуті найефективнші методи глибокого навчання для класифікації відео та аудіо
- Побудовані нейронні мережі для класифікації емоцій з відео та аудіо
- Побудований ансамбль моделей, який допоміг значно покращити результат
- Побудовані моделі адаптовані до зйомки з фронтальної відеокамери та мікрофону ноутбуків

ДЯКУЄМО ЗА УВАГУ