Поиск эффективной архитектуры для решения задачи получения мультимодальных эмбеддингов для трех и более модальностей

Жаров Георгий

МФТИ

10 марта 2024 г.

Цель работы

- Сравнить существующие методы получения мультимодальных эмбеддингов методом CCA.
- Найти эффективную нейросетевую архитектуру для решения задачи получения мультимодальных эмбеддингов для трех и более модальностей.

CCA

Даны данные из двух модальностей $X_1 \in \mathbb{R}^{d_1}$ и $X_2 \in \mathbb{R}^{d_2}$, с ковариационными матрицами, Σ_{11} и Σ_{22}, соответственно, и кросс-ковариационной матрицей, Σ_{12}
Рассматривается задача поиска линейных преобразований (u_1^*, u_2^*), максимизирующих корреляцию между модальностями.

CCA

Итоговая задача отптимизации имеет вид:

$$(u_1^*, u_2^*) = \arg \max_{u_1 \in \mathbb{R}^{d_1}, u_2 \in \mathbb{R}^{d_2}} corr(u_1^\top X_1, u_2^\top X_2)$$

$$= \arg \max_{u_1 \in \mathbb{R}^{d_1}, u_2 \in \mathbb{R}^{d_2}} \frac{u_1^\top \Sigma_{12} u_2}{\sqrt{u_1^\top \Sigma_{11} u_1 u_2^\top \Sigma_{22} u_2}}$$

Иначе

$$(u_1^*, u_2^*) = \arg \max_{u_1^\top \Sigma_{11} u_1 = u_2^\top \Sigma_{22} u_2 = 1} u_1^\top \Sigma_{12} u_2$$

Deep CCA

Основное ограничение классического CCA – невозможность поиска нелинейного преобразования. Рассмотрим следующее расширение.
Пусть теперь $f_1(X_1)$ и $f_2(X_2)$ выходы двух нейросетей с весами W_1 и W_2 соответственно.
Тогда новая оптимизационная задача имеет вид:

$$(u_1^*, u_2^*, W_1^*, W_2^*) = \arg\max_{u_1, u_2} corr(u_1^\top f_1(X_1), u_2^\top f_2(X_2))$$

GCCA

Еще одно расширение классического CCA.
Ставится задача нахождения некоторого общего представления G для J различных модальностей, где N — число объектов выборки, d_j — размерность j-ой модальности, r — размерность обучаемого представления, и $X_j \in \mathbb{R}^{d_j \times N}$ — матрица данных для j-ой модальности.
Тогда итоговая задача имеет вид:

$$\begin{aligned} \underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{} & \sum_{j=1}^{J} \| G - U_j^\top X_j \|_F^2 \\ \text{subject to} \quad & GG^\top = I_r \end{aligned}$$

Deep GCCA

Пусть теперь f_i — сеть для j-ой модальности данных, а o_j — размерность ее выходного вектора. Тогда, по аналогии с GCCA можем поставить задачу для DGCCA:

$$\begin{aligned} \min_{U_j \in \mathbb{R}^{o_j \times r}, G \in \mathbb{R}^{r \times N}} & \sum_{j=1}^{J} \| G - U_j^\top f_j(X_j) \|_F^2 \\ \text{subject to} \quad & GG^\top = I_r \end{aligned}$$

Deep GCCA

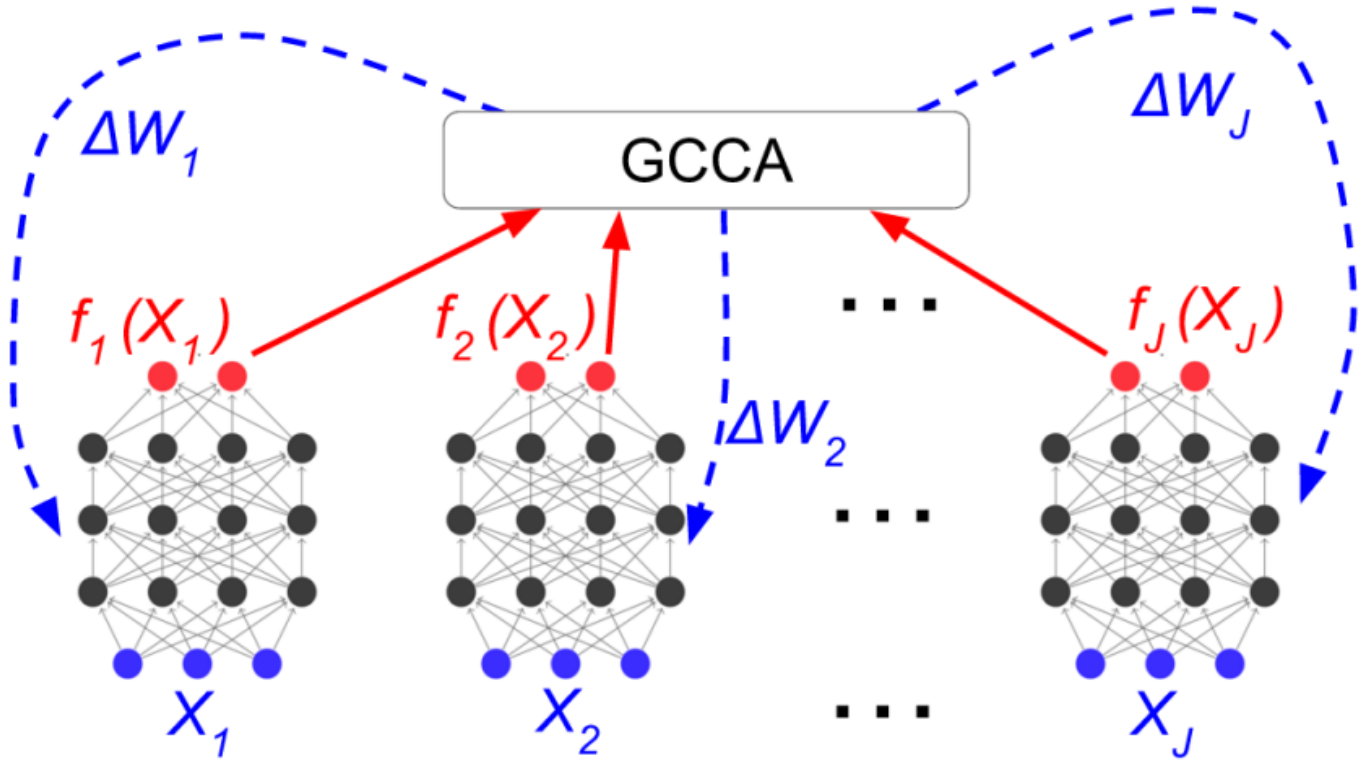
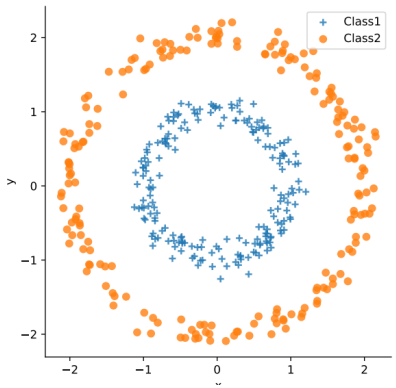
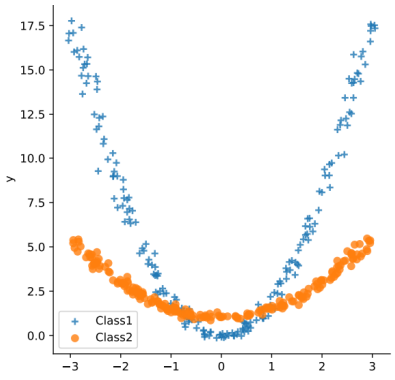


Рис.: Архитектура DGCCA

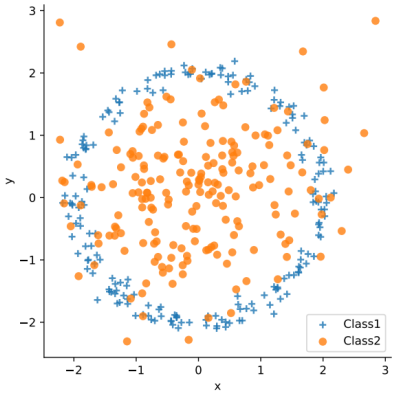
Применение DGCCA



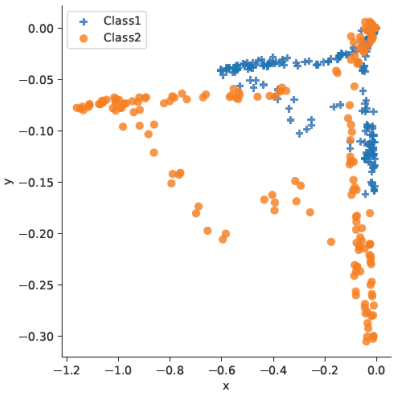
Mod 1 input



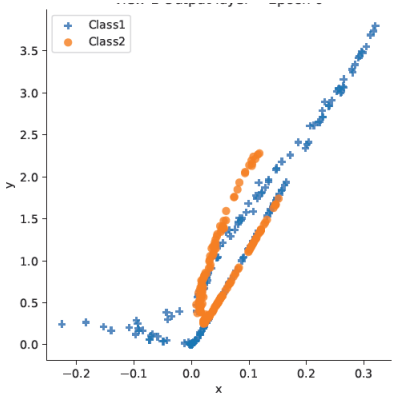
Mod 2 input



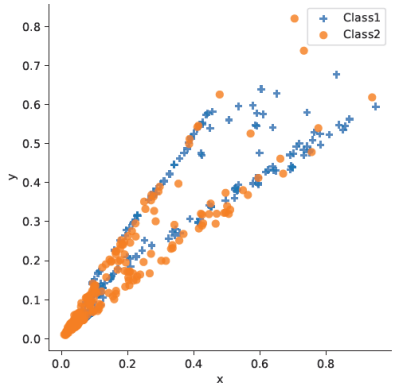
Mod 3 input



Mod 1 output

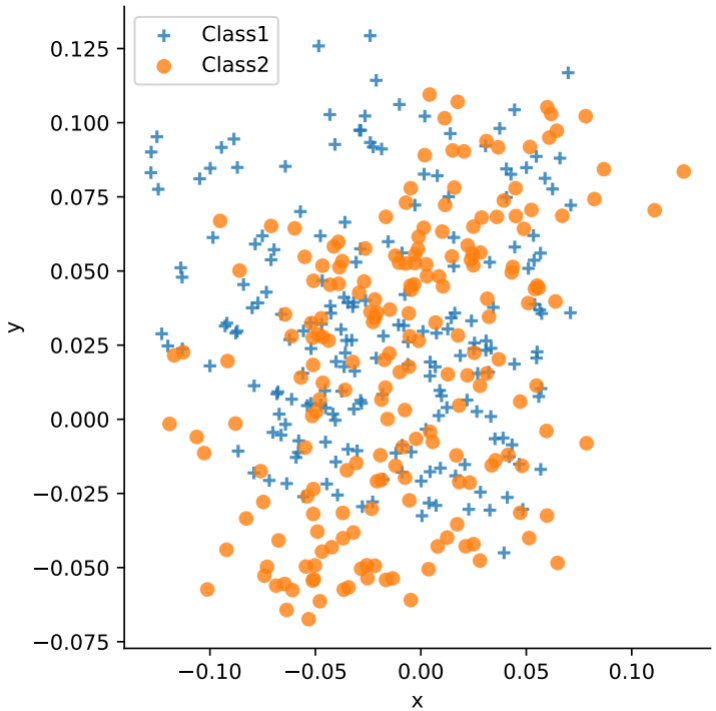


Mod 2 output

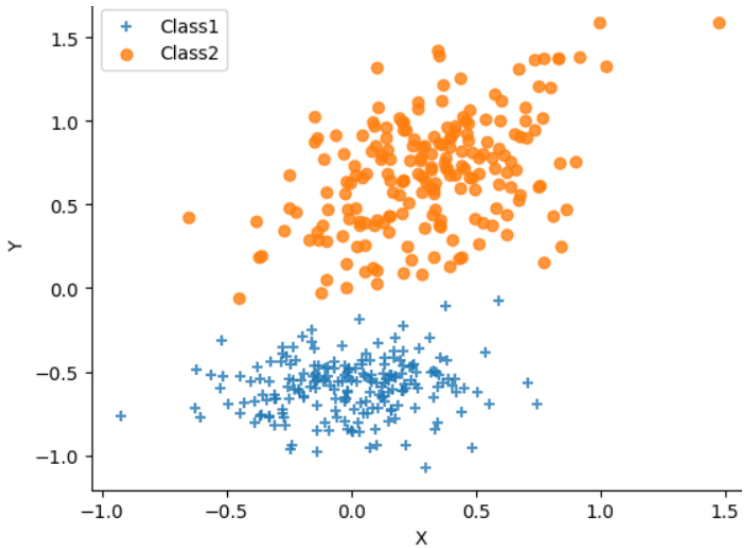


Mod 3 output

Применение DGCCA



GCCA



DGCCA

Рис.: Итоговые мультимодальные эмбединги для GCCA и DGCCA

Результаты

- Проведено исследование имеющихся методов получения мультимодальных эмбеддингов.
- Проведены вычислительные эксперименты с нейросетевой архитектурой GCCA.
- Написан код для проведения вычислительного эксперимента на синтетических данных.