ChatGPT

Cấu trúc tổng thể của CAFace

Hình: Khung tổng quan của hệ thống CAFace (Cluster and Aggregate). Đầu vào là tập N ảnh probe, lần lượt qua ba thành phần chính – Style Input Maker (SIM), Cluster Network (CN) và Aggregation Network (AGN) – để sinh ra một vectơ đặc trưng khuôn mặt hợp nhất cuối cùng [1] [2].

CAFace là framework hai giai đoạn nhằm hợp nhất một tập lớn các ảnh probe thành một đặc trưng duy nhất cho nhận dạng khuôn mặt [1]. Tổng quan dữ liệu luồng như sau (Hình trên): - **Đầu vào (Probe):** Tập N ảnh (kích thước H×W×3). Mỗi ảnh được đưa qua mạng trích xuất đặc trưng (backbone) để thu được: vectơ đặc trưng nhận dạng **f_i** ∈ ℝ^D (ví dụ D=512) và ma trận đặc trưng trung gian **F_i** ∈ ℝ^{C×H′×W′}.
- **Style Input Maker (SIM):** Nhận mỗi ma trận F_i, SIM tính vectơ thông tin style **s_i** ∈ ℝ^d bằng cách lấy trung bình μ và độ lệch chuẩn σ theo từng kênh của F_i [3]. Vectơ s_i (chiều d) có thể được ghép thêm embedding của chuẩn (norm) của f_i [3]. Kết quả là N vectơ style **S = {s₁, ..., s_N}**.
- **Cluster Network (CN):** CN nhận cặp (f_i, s_i) của N ảnh. Nó dùng M truy vấn toàn cục (các tâm cụm học được) để tính ma trận phân cụm **A ∈ ℝ^{M×N}** (soft assignment) bằng cách so s_i với M tâm cụm rồi Softmax theo cột [1] [2]. Ma trận A cho biết mức độ mỗi ảnh được gán vào từng cụm. Sau đó CN tạo ra M vectơ đặc trưng cụm **C_j ∈ ℝ^D** và M vectơ style cụm **X_j ∈ ℝ^d**, trong đó mỗi C_j là tổng tuyến tính của các f_i thuộc cụm j, và mỗi X_j là tổng tuyến tính của các s_i [1].
- **Aggregation Network (AGN):** AGN nhận M cặp (C_j, X_j) từ CN và dùng một MLP-Mixer để học mối quan hệ giữa các cụm. AGN sinh ra vectơ trọng số **w ∈ ℝ^M** cho mỗi cụm [4]. Cuối cùng, vectơ đặc trưng hợp nhất **f ∈ ℝ^D** được tính bằng tổng có trọng số: f = Σ_{j=1..M} w_j·C_j [4]. Đầu ra của toàn hệ thống là vectơ f dùng cho nhận dạng.

Kết hợp lại, CAFace chia quy trình feature fusion thành ba thành phần riêng biệt. Luồng dữ liệu cụ thể có thể tóm tắt như sau:

- **Đầu vào:** Tập ảnh probe {I₁,...,I_N}. Mỗi ảnh đầu tiên qua mạng nhận dạng khuôn mặt (backbone) để lấy đặc trưng **f_i ∈ ℝ^D** và các đặc trưng trung gian **F_i ∈ ℝ^{C×H′×W′}**.
- **SIM:** Nhận F_i, tính vectơ style **s_i = W·[μ(F_i); σ(F_i)]** (cộng thêm embedding của ||f_i||) [3]. Đầu ra của SIM là tập vectơ style S = (s₁,...,s_N) (Mỗi s_i ∈ ℝ^d).
- **CN:** Nhận (f_i, s_i)_{i=1..N}, sử dụng M truy vấn Q_j để tính ma trận A ∈ ℝ^{M×N} (công thức attention sửa đổi như trong (2)-(3) [5] [6]). Từ A, CN tạo ra hai ma trận: **C ∈ ℝ^{M×D}** (tập M vectơ C_j) và **X ∈ ℝ^{M×d}** (M vectơ X_j), với C_j = (Σ_i A_{j,i} f_i)/Σ_iA_{j,i} và tương tự cho X_j.
- **AGN:** Nhận C, X (kích thước M×(D+d)), qua MLP-Mixer sinh trọng số **w ∈ ℝ^M** [4]. Tính vectơ f_{fusion} = Σ_j w_j C_j. Đầu ra là vectơ đặc trưng hợp nhất f_{fusion} ∈ ℝ^D.

Style Input Maker (SIM)

SIM là mạng phụ trích xuất thông tin kiểu ảnh (style) từ mỗi ảnh mà không liên quan đến nhận dạng. Đầu vào của SIM là ma trận đặc trưng trung gian **F_i ∈ ℝ^{C×H′×W′}** thu được từ backbone. SIM tính toán giá trị trung bình μ và độ lệch chuẩn σ của F_i theo từng kênh, rồi dùng ma trận trọng số học được để biến đổi thành vectơ style **s_i** [3]. Kết quả bao gồm thông tin về độ sáng, độ tương phản, chất lượng ảnh, độ xoay, v.v., vốn đã bị loại bỏ khỏi vectơ nhận dạng f_i do mạng FR được huấn luyện nhận dạng chỉ quan tâm đến ID. Ngoài ra, SIM còn nhúng chuẩn ||f_i|| (theo dạng sinusoidal) vào s_i để đánh giá độ tin cậy của ảnh [3]. - **Đầu vào:** Ma trận F_i (kiểu float) kích thước C×H′×W′.
- **Đầu ra:** Vectơ style **s_i ∈ ℝ^d** (độ dài d).

Cluster Network (CN)

CN dùng cơ chế attention sửa đổi để gom nhóm N vectơ đầu vào vào M cụm cố định. Đầu vào của CN là bộ N vectơ nhận dạng **f_i ∈ ℝ^D** và N vectơ style **s_i ∈ ℝ^d**. Bên trong, CN có M truy vấn toàn cục Q_j ∈ ℝ^d (học được chung cho mọi probe) [7]. CN tính ma trận affinities Q·S^T rồi áp dụng Softmax theo cột để thu ma trận phân cụm **A ∈ ℝ^{M×N}** [7] [1]. Từ A, CN tạo ra M vectơ cụm **C_j ∈ ℝ^D** và M vectơ style cụm **X_j ∈ ℝ^d** bằng tổng cộng tuyến tính: mỗi C_j = (Σ_i A_{j,i} f_i)/ Σ_iA_{j,i}, tương tự với X_j từ s_i [1]. Do đó, số lượng đầu ra của CN luôn cố định M cụm, bất kể N thay đổi [2]. - **Đầu vào:** N vectơ f_i ∈ ℝ^D và N vectơ s_i ∈ ℝ^d.
- **Đầu ra:** Ma trận phân cụm A ∈ ℝ^{M×N}, M vectơ **C_j ∈ ℝ^D** và M vectơ **X_j ∈ ℝ^d** (j=1..M) [1] [2].

Aggregation Network (AGN)

AGN kết hợp M vectơ cụm thành một vectơ duy nhất. Đầu vào của AGN là tập M cặp (C_j, X_j) (mảng kích thước M×(D+d)). AGN sử dụng cấu trúc MLP-Mixer để lan truyền thông tin giữa các cụm. Đầu tiên, nó biểu diễn M vectơ này thành một ma trận rồi xuất ra vectơ **w ∈ ℝ^M** biểu diễn "tầm quan trọng" của mỗi cụm [4]. Cuối cùng, vectơ đầu ra **f ∈ ℝ^D** được tính bằng tổng có trọng số: **f = Σ_{j=1..M} w_j·C_j**. Giá trị w được xác định sao cho Σ_j w_j=1 (qua Softmax) để f là trung bình có trọng số của các cụm. Kết quả f là vectơ đặc trưng cuối cùng phục vụ tác vụ nhận dạng khuôn mặt.
- **Đầu vào:** M vectơ **C_j ∈ ℝ^D** và M vectơ **X_j ∈ ℝ^d**.
- **Đầu ra:** Vectơ hợp nhất **f ∈ ℝ^D** và vectơ trọng số cụm **w ∈ ℝ^M** [4].

Mỗi thành phần trên đóng vai trò rõ ràng trong pipeline: SIM trích xuất thông tin style (độc lập với ID), CN gom nhóm đầu vào thành M đặc trưng trung gian cố định, và AGN tổng hợp các đặc trưng đó thành một biểu diễn cuối cùng. Cách thiết kế này cho phép CAFace xử lý hiệu quả các tập probe rất lớn và duy trì bất biến với thứ tự đầu vào [8] [9].

Tài liệu tham khảo: Các thông tin trên được trích từ bài báo gốc CAFace [1] [2] [3] [4].

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [2210.10864] Cluster and Aggregate: Face Recognition with Large Probe Set
https://ar5iv.labs.arxiv.org/html/2210.10864