УДК 00.009.43

МЕХАНИЗМ ВНИМАНИЯ ДЛЯ СЛИЯНИЯ ПРИЗНАКОВ: СРАВНЕНИЕ МЕТОДОВ МУЛЬТИМОДАЛЬНОГО СЛИЯНИЯ

К.Н. Евсюков[1], В.Г. Литвинов[1]

[1]Самарский университет, г. Самара
e-mail: kirillevsukov125@gmail.com

Аннотация

В работе исследуются методы мультимодального слияния для локализации объекта. Каждый пример содержит изображение из 10 наложенных фигур и вектор их важностей. Архитектура состоит из свёрточного энкодера изображения и полносвязного энкодера скаляров. Сравниваются 8 методов слияния. Адаптивные методы показывают наилучшие результаты.

Ключевые слова: искусственный интеллект, мультимодальное слияние признаков, локализация объекта, механизм внимания, сверточные нейронные сети.

Введение

В современных задачах компьютерного зрения всё чаще требуется объединять информацию из различных источников для достижения высокой точности предсказаний. При решении таких задач пользуются мультимодальным обучением, когда модель одновременно обрабатывает визуальные данные (изображения) и семантические признаки (числовые векторы). Одной из ключевых проблем в таких моделях является выбор эффективного метода слияния признаков.

В данной работе исследуется задача локализации объекта на синтетически генерируемых изображениях.

1. Методы слияния

В работе исследуются восемь методов мультимодального слияния. Обозначим через f_{img} вектор признаков размерности 128, полученный из изображения свёрточным энкодером, а через f_{scal} вектор признаков размерности 128, полученный из скалярных данных полносвязным энкодером.

1.1. Объединение (Concat)

Метод заключается в конкатенации двух векторов признаков, который можно выразить:

$$f = [f_{img}; f_{scal}., \quad (1)$$

1.2. Сложение (Add)

Выполняется поэлементное сложение векторов, данное преобразование можно выразить:

$$f = f_{img} + f_{scal}. \quad (2)$$

1.3. Умножение (Multiply)

Скалярные признаки преобразуются в вектор через сигмоидную функцию, это преобразование представлено в виде следующих двух формул

$$v = \sigma(W_g \times f_{scal}), \quad (3)$$

$$f = f_{img} \circ v, \quad (4)$$

где σ – сигмоидная функция, W_g – обучаемая матрица, $\circ$ – поэлементное умножение. Преобразование линейно[1].

1.4. Послойная линейная модуляция (FiLM)

Метод реализует аффинное преобразование визуальных признаков под управлением скаляров, эти преобразования можно описать следующими формулами:

$$\gamma = W_\gamma \times f_{scal}, \quad (5)$$

$$\beta = W_\beta \times f_{scal}, \quad (6)$$

$$f = \gamma \circ f_{img} + \beta, \quad (7)$$

где γ и β – векторы масштабирования и сдвига, W_γ, W_β – обучаемые матрицы [2, 3].

1.5. Взвешенное управление (Gated Fusion)

Динамическое взвешивание двух модальностей, эти преобразования приведены в формулах 7–9:

$$\alpha = \frac{e^{h1}}{(e^{h1}+e^{h2})}, \quad (8)$$

$$\beta = \frac{e^{h2}}{(e^{h1}+e^{h2})}, \quad (9)$$

$$f = \alpha \times f_{img} + \beta \times f_{scal}, \quad (10)$$

где $h1$ и $h2$ – выходы полносвязного слоя, принимающего на вход объединённый вектор $[f_{img}; f_{scal}]$ [4].

1.6. Обучаемый вес (Learnable Weight)

Вводится скалярный обучаемый параметр w, преобразование представлено формуле 10:

$$f = [f_{img}; w \times f_{scal}]. \quad (11)$$

1.7. Сложение с проекцией (Simple Add)

Скалярные признаки предварительно проецируются в пространство визуальных признаков, эти преобразования представлены формулами 11 и 12:

$$f_{scal}' = W_p \times f_{scal}, \quad (12)$$

$$f = f_{img} + f_{scal}', \quad (13)$$

где W_p – обучаемая матрица проекции.

1.8. Механизм внимания (Attention Weighted)

Генерация весов внимания для каналов визуальных признаков, эти преобразования представлены в формулах 13 и 14:

$$a = \sigma(W_2 \times ReLU(W_1 \times f_{scal})), \quad (14)$$

$$f = f_{img} \circ a, \quad (15)$$

где a – вектор весов внимания размерности 128 [5].

Преобразование нелинейно.

2. Экспериментальная установка

Обучение проводилось на синтетическом датасете из 20 000 примеров, из которых 7000 для обучения, 1500 для валидации и 1500 для тестирования. Каждый объект содержал

изображение 100 на 100 пикселей, сформированное наложением десяти геометрических фигур, таких как эллипсы и многоугольники. Центры фигур выбирались со смещением до 25 пикселей от истинной цели. Большая полуось эллипсов составляла от 10 до 25 пикселей, количество вершин многоугольников от 4 до 7.

Интенсивность каждой фигуры определялась её важностью, вычисляемой как обратная функция расстояния от центра фигуры до цели. Чем ближе центр, тем выше важность и ярче фигура. Десять фигур накладывались суммированием масок, умноженных на важности. Добавлялся гауссов шум с нулевым средним и стандартным отклонением 0,02. Изображение нормализовалось к диапазону от 0 до 1. Вектор скалярных признаков представлял собой набор из десяти значений важности.

Архитектура включала свёрточный энкодер изображения из трёх блоков: свёртка с 32, 64 и 128 фильтрами, пакетная нормализация и максимальный пулинг 2×2. Завершали энкодер глобальное усреднение и полносвязный слой на 128 нейронов с ReLU. Полносвязный энкодер скаляров состоял из двух слоёв на 64 и 128 нейронов с ReLU и dropout 0,3.

После слияния признаков одним из восьми методов итоговый вектор поступал в общую часть: два полносвязных слоя на 256 и 128 нейронов с пакетной нормализацией, dropout 0,3 и ReLU. На выходе получались две координаты цели.

Модели обучались 250 эпох с оптимизатором Adam и функцией потерь MSE. Качество оценивалось по средней евклидовой ошибке локализации в пикселях.

В результате экспериментов методы внимания и обучаемого веса дают предсказания, наиболее близкие к истинной точке, в то время как метод FiLM существенно ошибается.

3. Результаты экспериментов

Результаты сравнительного анализа на одной и той же тестовой выборке восьми методов слияния представлены в таблице 1. Наименьшую среднюю ошибку локализации в 3,35 пикселя показал метод внимания, что на 0,03 пикселя лучше метода с обучаемым весом (3,38) и на 0,12 пикселя лучше объединения (3,47). Медианная ошибка для этих методов составила 3,07, 3,10 и 3,21 пикселя соответственно.

Таблица 1 – Результаты сравнения методов

Метод	Средняя ошибка, пиксели	Медианная ошибка
Внимание	3,35	3,07
Обучаемый вес	3,38	3,10
Объединение	3,47	3,21
Сложение	3,64	3,36
Взвешенное управление	3,78	3,48
Сложение с проекцией	3,80	3,54
Умножение	4,28	3,91
Послойная линейная модуляция	5,33	4,79

Наихудший результат показал метод FiLM со средней ошибкой 5,33 пикселя, что на 59% хуже лучшего метода.

Заключение

В работе проведён сравнительный анализ восьми методов мультимодального слияния для задачи локализации объекта. Эксперименты на синтетическом показали, что выбор стратегии слияния существенно влияет на точность предсказаний.

Наилучшие результаты продемонстрировали адаптивные методы: механизм внимания (3,35 пикселя), обучаемый вес (3,38 пикселя) и объединение (3,47 пикселя). Метод FiLM показал низкую эффективность (5,33 пикселя), что может объясняться избыточной сложностью модели при ограниченном объёме данных. Простые методы (сложение, умножение) уступили адаптивным, подтверждая необходимость гибкой модуляции признаков.

Полученные результаты могут быть использованы при разработке систем локализации объектов, например при детектировании БПЛА, а также при проектировании мультимодальных архитектур для задач компьютерного зрения.

Список использованных источников

1. Kim, J. Hadamard Product for Low-rank Bilinear Pooling / J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, B. T. Zhang // arXiv preprint arXiv:1610.04325. – 2016. – P. 1–10.

2. Perez, E. FiLM: Visual reasoning with a general conditioning layer / E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – V. 32, № 1. – P. 3942–3951.

3. Dumoulin, V. Feature-wise Transformations / V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. De Vries, A. Courville, Y. Bengio // Distill. – 2018. – V. 3, № 6. – P. e11.

4. Arevalo, J. Gated multimodal units for information fusion / J. Arevalo, T. Solorio, M. Montes-y-Gomez // Proceedings of the International Conference on Learning Representations. – 2017. – P. 1–11.

5. Hu, J. Squeeze-and-Excitation Networks / J. Hu, L. Shen, G. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 7132–7141.

ATTENTION MECHANISM FOR FEATURE FUSION: COMPARISON OF MULTIMODAL FUSION METHODS

K.N. Evsukov[1], V.G. Litvinov[1]

[1]Samara University, Samara
e-mail: kirillevsukov125@gmail.com

Abstract

This work investigates multimodal fusion methods for object localization. Each example contains an image consisting of 10 overlapping shapes and a vector of their importances. The architecture consists of a convolutional image encoder and a fully connected scalar encoder. Eight fusion methods are compared. Adaptive methods show the best results.

Keywords: artificial intelligence, multimodal feature fusion, object localization, attention mechanism, convolutional neural networks.