YOLOv12: Ориентированные на внимание детекторы объектов в режиме реального времени

Выполнил студент группы Б21-221
Денисов В.А.

YOLO

YOLO(You Only Look Once) - широко известная архитектура компьютерного зрения, первая версия вышла в 2016 году и решала только задачу детекции объектов на изображении, сейчас представляет из себя целую фундаментальную модель, которую можно использовать для классификации, детекции, сегментации , трекинга объектов на видео в реальном времени. Таким образом важным показателем является не только точность, но и скорость работы.

Ключевые улучшения

1. Зональный модуль внимания(area attention module)
2. R-ELAN для аггрегации
3. Модернизация архитектуры

Зональный модуль внимания

Механизм внимания - выявление глобальных зависимостей, но медленный.

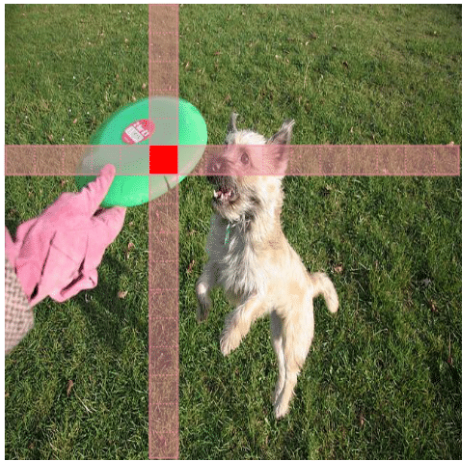
Свойства	CNN	Attention mechanism
Сложность	O(L)	$O(L^2) = 2L^2hd$
Доступ к памяти	структурированный	нерегулярный

L - длина последовательности.

Выход: зональный модуль внимания(area attention, A2).

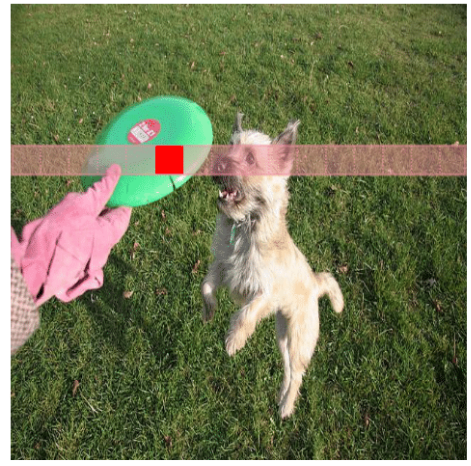
Сложность $\frac{1}{2}\ L^2hd, L \approx 640 \times 640 = const$.

Линейный механизм внимания с сложностью O(L) ухудшает глобальную зависимость, нестабильный.

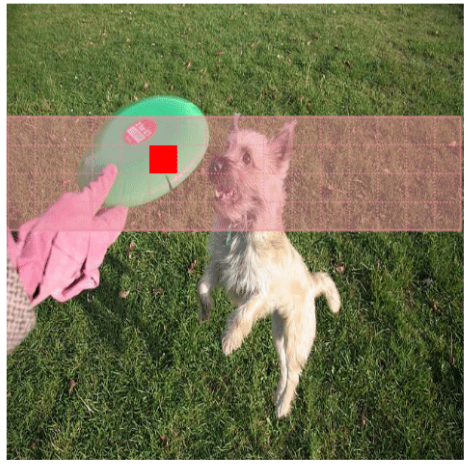


Criss-cross attention

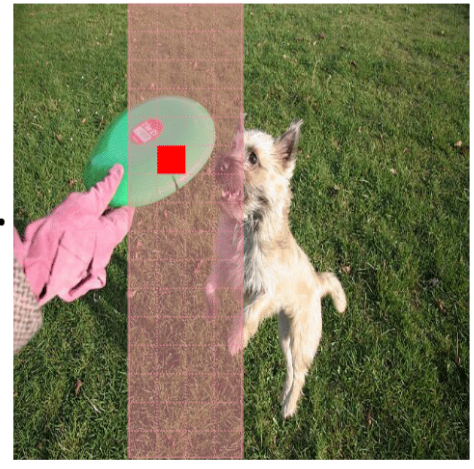
Window attention



Axial attention



or

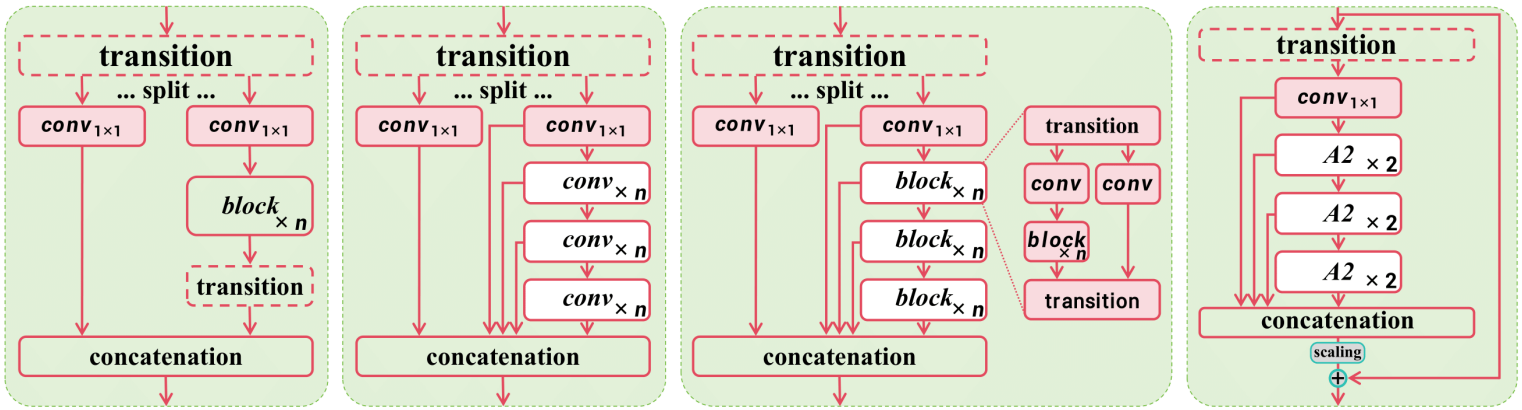


Area attention (Ours)

R-ELAN

ELAN(afficient layer aggregation networks) - блок для аггрегации. Нестабильный, затухание градиентов и отсутствие остаточных соединений между входом и выходом.

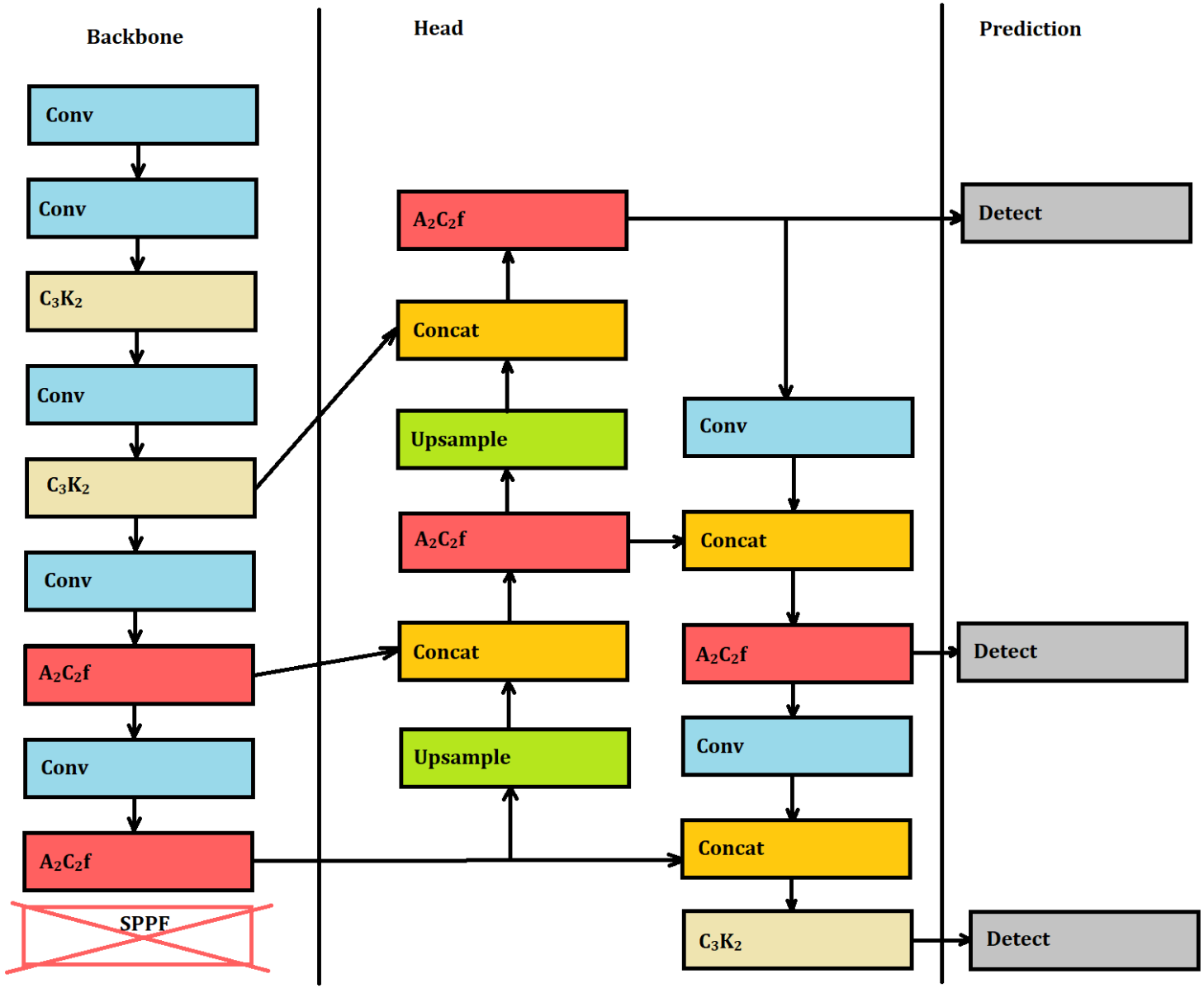
Выход: R-ELAN(residual) - новый подход к аггрегации, снижение вычислительных затрат, сохранение соединений I/O.



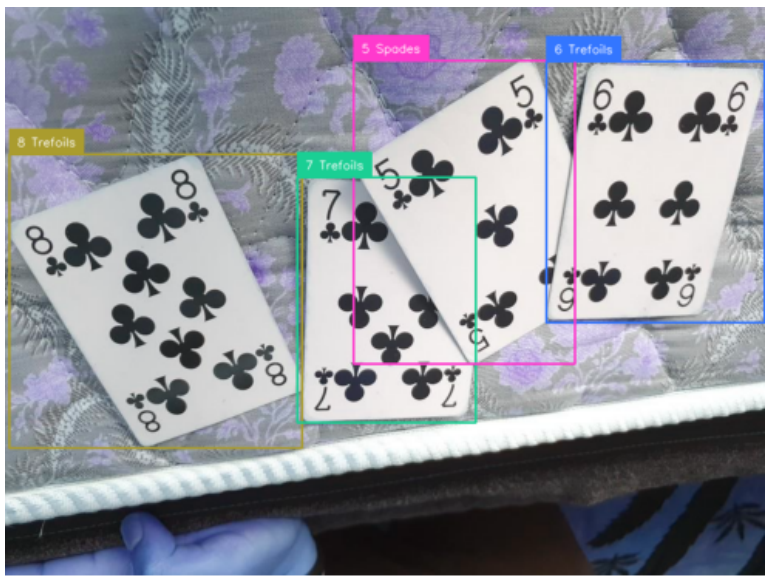
Модернизация архитектуры

- Flash Attention убирает позиционное кодирование для быстрого доступа к памяти в блоках attention
- соотношение MLP с 4 до 1.2 (увеличение размерности выхода) для балансировки attention и FFN(feed forward network) для повышения производительности
- замена последнего этапа пуллинга на R-ELAN

Блок-схема YOLOv12



Практическая часть, картинки



Практическая часть, метрики

Модель	$mAP_{50:90}$	mAP_{50}	mAP_{75}	Speed(ms)
YOLO 12n	0.35	0.42	0.41	16.1
YOLO 11	0.37	0.44	0.43	9.0
YOLO 10n	0.19	0.22	0.22	10.3
YOLO 9m	0.88	0.97	0.97	20.1
YOLO 8n	0.67	0.78	0.77	6.3

Модель	$mAP^{Mask}_{50:90}$	$mAP^{Box}_{50:90}$	Speed(ms)
YOLO 12n	0.46	0.52	23.1
YOLO 11	0.49	0.56	11.2
YOLO 9c	0.56	0.60	38.0
YOLO 8n	0.54	0.56	7.67

Источники

1. YOLOv12: Attention-Centric Real-Time Object Detectors Yunjie Tian, Qixiang Ye, David Doermann, URL: https://arxiv.org/abs/2502.12524