УДК 004.853 URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1077.html

КОМПЛЕКСНАЯ ОПТИМИЗАЦИЯ DEEPLABV3 ДЛЯ МОБИЛЬНЫХ УСТРОЙСТВ С ПРИМЕНЕНИЕМ ОБЛЕГЧЕННЫХ АРХИТЕКТУР

П.В. Малышев pasha_malyshev01@mail.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Глубокие сверточные нейронные сети (CNN) стали основой для решения задач сегментации изображений, что делает их незаменимыми для различных приложений, включая автономные системы и мобильные устройства. Однако высокие требования к вычислительным ресурсам делают использование моделей, таких как Deeplabv3, затруднительным на устройствах с ограниченными ресурсами. В данной статье исследованы возможности оптимизации модели Deeplabv3 с помощью замены бэкбона ResNet-50 более легкими архитектурами, такими как ResNet-18 и MobileNetV2. Исследовано снижение потребления ресурсов без значительных потерь точности модели при использовании облегченных архитектур.

Ключевые слова: сверточные нейронные сети, сегментация, цифровая обработка изображений, глубокое обучение, оптимизация

Введение. Сегментация изображений представляет собой одну из сложных задач в области компьютерного зрения, требующую анализа на уровне пикселей. Модель Deeplabv3, основанная на глубинных нейронных сетях и пространственной пирамидальной свертке (Atrous Spatial Pyramid Pooling, ASPP), показывает хорошие результаты в сегментации, однако ее вычислительная сложность ограничивает возможности применения на устройствах с ограниченными ресурсами.

В данной работе исследованы методы оптимизации моделей сегментации и выполнено сравнение точности и производительности модели Deeplabv3 со стандартным бэкбоном[1] ResNet-50 с моделями, в которых был применен метод дистилляции знаний путем обучения стандартной модели-учителя, и моделей-учеников, в которых была проведена замена бэкбона более компактными архитектурами, такими как ResNet-18 и MobileNetV2.

Оптимизация процесса обучения. Для повышения скорости обучения модели Deeplabv3 и снижения нагрузки на аппаратные ресурсы можно ис-

[1] Бэкбон (англ. *backbone*) — основная (базовая) сеть, служащая для извлечения признаков из поступающего на вход изображения.

 ISSN 2541-8009

пользовать несколько подходов к настройке обучения. Эти методы включают изменение размера входных изображений, аугментацию данных и выбор оптимального размера батча[2]. Каждый из этих подходов позволяет сократить время обучения и эффективно использовать доступные ресурсы.

Изменение размера входных изображений — один их методов для снижения вычислительных затрат и потребления памяти. Благодаря уменьшению числа пикселей для обработки достигаются ускорение вычислений и снижение требований к оперативной памяти. При этом потенциально возможно снижение точности из-за потери деталей на изображениях с низким разрешением.

Помимо размера изображений на объем памяти, необходимый для вычислений, влияет размер батча. Меньший размер батча позволяет снизить нагрузку на память графического процессора, делая процесс обучения более стабильным даже на устройствах с малым объемом памяти. И наоборот, чем больше батч, тем больше требуется оперативной памяти, но в то же время повышается стабильность градиентов, что может способствовать более плавному обучению [1]. Выбор батчей небольшого размера может привести к нестабильности градиентов, что усложняет обучение и делает его более шумным. Чтобы компенсировать этот эффект, можно применять технику накопления градиентов. При использовании оборудования с ограниченной памятью всегда необходимо подбирать размер батча с учетом возможностей памяти.

Выбор оптимизаторов и методов регуляризации. Подходы к выбору оптимизатора и методов регуляризации[3] помогают ускорить сходимость модели, избежать переобучения и обеспечить ее более стабильное и эффективное обучение.

В рассматриваемых примерах использовали оптимизатор AdamW (Adam с весовым затуханием). Adam и AdamW — популярные алгоритмы оптимизации, но AdamW имеет ряд преимуществ в контексте глубоких нейронных сетей. В нем используется L2-регуляризация, что позволяет модели лучше обобщать данные и предотвращать переобучение [2].

Затухание весов в оптимизаторе AdamW действует на веса модели отдельно от градиентов, улучшая регуляризацию. В Adam L2-регуляризация

[2] Батч (англ. *batch*) — подмножество обучающего набора данных (изображений), которое подается на вход нейронной сети одновременно для обработки и вычисления градиентов во время обучения модели.

[3] Регуляризация (англ. *regularization*) — техника обучения нейронных сетей, предотвращающая переобучение путем добавления штрафов к функции потерь, что контролирует сложность модели и улучшает обобщение на новых данных.

также используется, но она применяется как часть градиента, что иногда вызывает чрезмерное обновление весов. AdamW более устойчив к изменению скорости обучения и улучшает способность модели обобщать данные. Это делает его подходящим для крупных моделей с большим числом параметров, таких как Deeplabv3.

Помимо оптимизатора AdamW была использована регуляризация Dropout. Регуляризация Dropout — один из методов регуляризации, который помогает уменьшить риск переобучения, исключая случайные нейроны на каждом шаге обучения. Это позволяет модели лучше обобщать данные и справляться с шумом, который может присутствовать в обучающем наборе. На каждом шаге обучения случайно отключается определенный процент нейронов в каждом слое. Это заставляет модель обучаться на каждом шаге с неполной информацией, что улучшает ее способность выявлять общие закономерности, а не запоминать конкретные данные. Регуляризация Dropout повышает устойчивость к переобучению и улучшает способность модели обобщать данные. Это особенно важно при обучении на небольших наборах данных, где велик риск переобучения.

В дополнение ко всему был применен метод Batch Normalization, который стабилизирует и ускоряет обучение, нормализуя активации на каждом слое нейронной сети [3]. Это позволяет использовать более высокие скорости обучения и избежать проблем с сходимостью, которые часто встречаются в глубоких сетях.

С помощью метода Batch Normalization можно нормализовать входы для каждого слоя, вычитая среднее значение и деля на стандартное отклонение, а затем добавляя обучаемые параметры смещения и масштаба. Это снижает зависимость от начальных параметров и делает сеть менее чувствительной к изменениям скорости обучения. Метод Batch Normalization помогает сократить время обучения, уменьшить риск переобучения и повысить устойчивость модели к высоким значениям скорости обучения. Он также дает возможность уменьшить разницу между обучением и инференсом[4], что делает сеть более стабильной.

Использование облегченных бэкбонов. В модели Deeplabv3 в качестве бэкбона традиционно используется ResNet-50 или более мощные архитекту-

[4] Инференс (англ. *inference* — вывод) в машинном обучении — это процесс применения уже обученной модели к новым данным для получения предсказаний или выводов. В отличие от этапа обучения модели, который требует больших объемов данных и мощных вычислительных ресурсов, инференс происходит на основе уже обученной модели и предназначен для работы с новыми, ранее не встречавшимися данными.

 ISSN 2541-8009

ры. Однако такие архитектуры потребляют много памяти и требуют высоких вычислительных ресурсов, что делает их непрактичными для приложений с ограниченным оборудованием или для работы в реальном времени. В данной работе рассмотрены варианты замены ResNet-50 более легкими бэкбонами, такими как ResNet-18 и MobileNetV2, что позволяет сохранить качество сегментации при значительном уменьшении вычислительных затрат и объема памяти.

Использование облегченных бэкбонов позволяет добиться снижения вычислительных затрат и уменьшения объема потребляемой памяти. В легких бэкбонах, таких как ResNet-18 и MobileNetV2, используются упрощенные архитектурные блоки, которые позволяют сократить число параметров и ускорить вычисления без существенной потери точности. При этом для облегченных бэкбонов требуется меньше памяти, что позволяет запускать модель на устройствах с ограниченными ресурсами, таких как мобильные и встроенные системы. Благодаря уменьшению числа операций свертки и параметров модели с легкими бэкбонами могут обрабатывать изображения быстрее, что особенно важно для задач реального времени.

В MobileNetV2 используются глубинные свертки (depthwise separable convolutions), которые заменяют стандартные свертки, разделяя пространственные и каналовые операции, что значительно снижает объем вычислений. Также используется структура с обратными остаточными блоками (inverted residual blocks) и блоки с линейной активацией, что уменьшает потери информации. MobileNetV2 имеет небольшой размер и подходит для задач реального времени, обеспечивая хорошее качество сегментации при относительно низких затратах.

Сеть ResNet-18 — это упрощенная версия стандартного бэкбона ResNet, состоящая всего из 18 слоев, что делает ее значительно легче ResNet-50. Хотя ResNet-18 содержит меньше параметров, ее архитектура сохраняет основные особенности ResNet, такие как остаточные связи, которые помогают избежать проблемы исчезающих градиентов. ResNet-18 обеспечивает стабильную производительность и высокую точность для задач сегментации, сохраняя при этом доступную сложность.

Дистилляция знаний. Дистилляция знаний — это техника обучения, в которой сложная и ресурсоемкая модель-учитель передает свои знания более простой модели-ученику. Этот подход позволяет добиться того, чтобы легкая модель-ученик работала быстрее, но сохраняла точность, близкую к учительской модели, что особенно полезно для задач с высокими требованиями к производительности на ограниченном оборудовании, таких как сегментация в реальном времени [4].

Дистилляция знаний основана на передаче информации от более точной, но сложной модели к более простой. Учительская модель обучается на полных данных, и в процессе ее обучение генерирует мягкие метки (soft labels), которые не просто указывают на правильный класс, а содержат информацию о распределении вероятностей. Затем модель-ученик обучается не только на исходных данных, но и на мягких метках, полученных от учителя, что позволяет ей улавливать более глубокие закономерности, присутствующие в данных.

Использование дистилляции знаний в Deeplabv3 позволяет добиться снижения вычислительных затрат при сохранении практически идентичной точности. За счет уменьшения сложности и объема параметров в модели-ученике удается сократить потребление ресурсов и увеличить скорость инференса. При этом дистилляция позволяет сохранить большую часть точности и качества сегментации благодаря обучению модели-ученика с мягкими метками, которые содержат более богатую информацию, чем обычные бинарные метки. Поскольку мягкие метки содержат распределение вероятностей для всех классов, модель-ученик обучается понимать не только основной класс, но и вероятность остальных классов, что помогает ей лучше обобщать данные.

Для использования дистилляции знаний с Deeplabv3 в качестве модели-учителя были предприняты следующие шаги.

1. Обучение полной модели Deeplabv3 с бэкбоном ResNet-50 на исходных данных до достижения хороших показателей точности. Этот шаг занимает много времени и ресурсов, так как учительская модель должна быть достаточно точной для передачи знаний.

2. Генерация soft labels с использованием учительской модели. Для каждого изображения в обучающем наборе данных получено распределение вероятностей для всех пикселей, которое используется как целевой ответ для модели-ученика.

3. Замена бэкбона ResNet-50 на более легкие архитектуры ResNet-18 и MobileNetV2 с последующим обучением модели-ученика. При этом используются как soft labels от учительской модели, так и оригинальные метки для вычисления потерь.

4. Обучение студенческой модели ведется с комбинированной функцией потерь, которая учитывает soft labels от учительской модели и реальные метки классов. Это позволяет модели-ученику учиться на глубинных признаках, которые учительская модель смогла обнаружить, а также на базовых данных.

Формула для расчета комбинированной функции потерь имеет вид

$$Loss = \alpha \cdot Loss_{soft} + (1-\alpha) \cdot Loss_{hard},$$

где $Loss_{soft}$ — это потери по мягким меткам; $Loss_{hard}$ — потери по реальным меткам; α — коэффициент, регулирующий вклад мягких и жестких меток.

 ISSN 2541-8009

В рассматриваемом в данной работе эксперименте учительская модель обучена на наборах данных Cityscapes и генерирует сегментированные изображения с высокой точностью, а для моделей-учеников использованы Deeplabv3-Lite с бэкбонами ResNet-18 и MobileNetV2, которые обучены на тех же данных с использованием мягких меток, полученных от учителя. Это позволяет добиться того, чтобы студенческая модель могла сегментировать городские сцены с минимальными вычислительными затратами, обеспечивая высокую скорость инференса на мобильных и встроенных устройствах.

Улучшение обобщающей способности. Одним из способов повышения обобщающей способности модели является применение методов аугментации[5], которые позволяют сократить риск переобучения, добавляя разнообразие в тренировочные данные. Особую пользу аугментация приносит при обучении на ограниченных наборах данных [5]. Существующие методы аугментации позволяют создать множество уникальных вариаций одного и того же изображения, что помогает модели изучить больше примеров и приспособиться к обработке изменений в данных.

Для всех рассматриваемых в данной работе архитектур моделей были применены следующие типы аугментации:

– геометрические преобразования (обрезка, поворот на случайный угол, отражения по горизонтали и вертикали и масштабирование изображений). Экспериментально было установлено, что использование геометрических аугментаций повышает устойчивость к изменениям углов съемки;

– цветовые преобразования (изменение яркости, контрастности и насыщенности. Данный тип преобразований делает модель устойчивой к изменениям условий освещения, что полезно для данных с реальных камер, таких как видеорегистраторы).

Отметим, что использование чрезмерной аугментации может ввести шум, что затруднит обучение.

Эксперимент. Для проведения экспериментов использовали набор данных Cityscapes [6], популярный в задачах сегментации городских изображений (см. рисунок). Модель обучалась с использованием оптимизатора AdamW, коэффициент снижения скорости обучения устанавливали равным 0,1 каждые 30 эпох.

При использовании указанных бэкбонов сохраняется модуль ASPP в Deeplabv3, что позволяет модели захватывать многоуровневую информацию о контексте изображения. Этот модуль адаптируется под новые бэкбоны,

[5] Аугментация (англ. *augmentation*) — это техника искусственного расширения обучающего набора данных путем создания модифицированных версий имеющихся примеров.

сохраняя функциональность ASPP для эффективного захвата пространственной информации.

Производительность моделей оценивали с использованием следующих метрик:

– mIoU (средняя точность пересечения и объединения) — характеризует качество сегментации;

– mAE 0,5 (%) — определяет среднюю абсолютную ошибку (mean Absolute Error) при фиксированном пороге перекрытия 0,5. Чем выше точность модели (mIoU), тем ниже значения mAE, так как модель точнее определяет границы объектов;

– mAE 05:095 (%) — рассчитывается как среднее значение для порогов от 0,5 до 0,95. Эта метрика более чувствительна к точности предсказаний на различных уровнях перекрытия. Модели с облегченными бэкбонами демонстрируют больше ошибок при высоких порогах (ближе к 0,95), что приводит к увеличению среднего значения mAE.

Эффективность замены ResNet-50 на облегченный бэкбон в модели-ученике оценивали с помощью метрик производительности и точности:

– количество операций с плавающей точкой (FLOPs);

– потребление памяти — позволяет понять, насколько легче и менее ресурсоемкой стала модель после замены бэкбона.

Пример размеченного изображения

Результаты экспериментов (см. таблицу) показали, что переход с ResNet-50 на ResNet-18 позволил сократить вычислительные затраты на 40 % при снижении точности mIoU всего на 2 %. Использование сети MobileNetV2 да-

 ISSN 2541-8009

ло еще больший прирост производительности, позволяя снизить объем вычислений на 65 % при падении mIoU на 3,5 %.

Сводная таблица результатов экспериментов

Модель	Бэкбон	FLOPs (миллиарды)	Потребление памяти (МБ)	mIoU (%)	mAE 0.5 (%)	mAE 05:095 (%)
Deeplabv3	ResNet-50	25,9	780	75,4	15,3	18,7
Deeplabv3-Lite	ResNet-18	15,5	470	73,2	17,1	21,0
Deeplabv3-Lite	MobileNetV2	9,1	310	71,9	18,4	22,5

Подход с использованием моделей-учеников с облегченными бэкбонами обеспечивает существенное снижение вычислительных затрат и позволяет модели работать на устройствах с ограниченными ресурсами. Небольшое снижение точности приемлемо для задач реального времени, в которых критична скорость обработки данных.

Другие подходы. Помимо рассмотренных методов можно применить другие способы оптимизации, такие как инкрементальное обучение, квантизация[6] и прунинг[7]. Инкрементальное обучение — это метод, позволяющий модели обучаться поэтапно, добавляя новые данные или классы по мере их появления, без необходимости полного переобучения на всем наборе данных. Этот подход особенно полезен в тех случаях, когда модель Deeplabv3 должна адаптироваться к новым сценам или условиям, что делает ее подходящей для реального применения в условиях изменяющейся среды, например, при сегментации дорожных изображений в разных погодных условиях и времени суток.

Инкрементальное обучение не требует полного переобучения на старом наборе данных при добавлении новых данных, что существенно экономит вычислительные ресурсы [7]. Модель при этом может постепенно обновляться и адаптироваться к новым условиям или классам, что полезно для решения задач, в которых данные могут постоянно изменяться, как в случае с дорожными сценами. При инкрементальном обучении модель сохраняет свои

[6] Квантизация — это метод оптимизации, позволяющий уменьшить размер и сложность моделей машинного обучения без заметной потери их производительности. Представляет собой процесс преобразования числовых значений в модели машинного обучения из высокоточных форматов с плавающей запятой.

[7] Прунинг (англ. *pruning*) — метод сжатия (уменьшения расхода памяти и вычислительной сложности) сети за счет устранения части параметров в предобученной модели.

предыдущие знания, что предотвращает забывание ранее изученных классов и позволяет добавлять новые знания постепенно.

Для используемой модели Deeplabv3 можно рассмотреть следующие подходы к инкрементальному обучению:

1) добавление регуляризационных методов, которые используются для предотвращения забывания старых знаний. К примеру, метод Elastic Weight Consolidation (EWC) добавляет дополнительные регуляризационные члены в функцию потерь, чтобы защитить параметры, важные для старых классов, во время обучения на новых данных;

2) использование репрезентативных данных, при которых модель сохраняет часть старых данных (репрезентативные примеры) или генерирует их с помощью генеративных моделей. Во время обучения на новых данных часть старых примеров также используется, что помогает модели не забывать старые классы и улучшает ее способность к адаптации;

3) добавление новых классов для дорожных объектов также может оказаться эффективным, если модель постоянно будет получать новые данные или классы, но не будет иметь доступа к полному исходному набору данных. Для Deeplabv3 это может быть применимо при добавлении таких объектов, как конкретные дорожные знаки или автомобили разных типов.

Если модель обучена на дорожных изображениях при дневном свете и ясной погоде, в условиях реальной эксплуатации автомобильный видеорегистратор может получать новые данные при различных погодных условиях и времени суток. В этом случае инкрементальное обучение позволит обновить модель для новых погодных условий через добавление новых данных, например, сегментацию объектов в условиях дождя, снега и тумана. Используя инкрементальное обучение, модель сможет адаптироваться к новым условиям, не теряя способности распознавать объекты, обученные ранее. Модель можно периодически обновлять, добавляя примеры из различных ситуаций (например, ремонт дороги или обновление дорожных знаков и разметки), чтобы она могла сегментировать новые сцены без переобучения на полных данных.

Квантизация и прунинг. Помимо инкрементального обучения увеличить производительность модели позволяют методы квантизации и прунинга. Квантизация и прунинг — это два ключевых подхода для уменьшения размера модели и сокращения вычислительных затрат, что позволяет существенно повысить скорость инференса и снизить потребление памяти. Оба метода можно использовать для оптимизации Deeplabv3, сохраняя приемлемый уровень точности при уменьшении объема модели [8]. Эти методы особенно полезны для развертывания на устройствах с ограниченными вычислительными ресурсами, таких как мобильные устройства и встроенные системы.

 ISSN 2541-8009

Квантизация позволяет уменьшить точность представления весов и активаций модели, снижая при этом объем памяти и увеличивая скорость вычислений.

Среди методов квантизации наиболее часто применяют следующие:

– 8-битную квантизацию — наиболее распространенный метод квантизации, где веса и активации модели представляются в формате 8-битных целых чисел вместо стандартных 32-битных чисел с плавающей точкой. Это сокращает объем памяти, необходимый для хранения весов, в четыре раза и ускоряет вычисления за счет того, что многие процессоры могут выполнять операции с 8-битными числами гораздо быстрее;

– динамическую квантизацию, при которой квантизация применяется только на этапе инференса, оставляя веса модели в 32-битном формате, но преобразовывая их в 8-битный формат на этапе выполнения. Это дает значительное ускорение, не влияя на точность обучения;

– постобучающую квантизацию — собой метод, при котором модель квантизируется после завершения обучения, не требуя изменения процесса обучения. Постобучающую квантизацию применяют, когда исходная точность модели удовлетворительна, и требуется только ее ускорение на этапе инференса.

Благодаря использованию квантизации можно добиться сокращения использования памяти и вычислительных затрат путем уменьшения веса модели. Операции при этом будут выполняться быстрее благодаря использованию низкоточных арифметических операций. При этом современные методы квантизации позволяют сохранить почти такой же уровень точности, как и у модели с 32-битными весами.

Цель прунинга — удалить избыточные параметры модели с целью уменьшения ее размера и ускорения работы. Этот метод может быть эффективен для сети Deeplabv3 с бэкбоном ResNet-50, в которой существует много параметров, незначительно влияющих на точность.

Основные методы прунинга можно подразделить на три категории:

1) структурный прунинг. Удаляет целые фильтры или каналы в сверточных слоях, что приводит к сокращению вычислений в каждом слое. В ResNet-50 можно удалить наименее значимые фильтры из каждого сверточного слоя;

2) неструктурный прунинг. Удаляет отдельные веса в слоях, что приводит к разреженности модели. В отличие от структурного прунинга, неструктурный прунинг сложнее оптимизировать для ускорения инференса, поскольку разреженные слои требуют специализированных аппаратных средств для эффективного выполнения вычислений;

3) прунинг на основе важности. Весы ранжируются по их влиянию на результаты модели, и наименее значимые веса отбрасываются. Это помогает минимизировать влияние прунинга на точность.

В дальнейших исследованиях планируется изучить влияние прунинга на снижение числа параметров для уменьшения размера модели, экономии памяти и снижения потребления энергии. Сокращение объема вычислений на каждом слое должно ускорить инференс, особенно при структурном прунинге, который позволяет уменьшить размер слоев в сверточных сетях. В целом можно сочетать квантизацию и прунинг для достижения максимального эффекта.

Заключение. Использование вышеупомянутых методов позволит сократить требования к ресурсам для работы модели Deeplabv3 с бэкбоном ResNet-50 и ускорить процесс обучения. Оптимизация позволит модели работать в условиях ограниченного оборудования, как это требуется для задач сегментации изображений с видеорегистратора.

Настоящее исследование демонстрирует эффективность использования облегченных бэкбонов в модели Deeplabv3 для сегментации изображений. Замена ResNet-50 на ResNet-18 или MobileNetV2 позволила снизить вычислительные затраты и потребление памяти, что делает модель более подходящей для внедрения в мобильные устройства и системы с ограниченными ресурсами. В дальнейшем планируется провести дополнительные исследования с применением методов квантизации и прунинга для дальнейшего облегчения модели.

Применение перечисленных методов позволяет ускорить процесс обучения модели Deeplabv3 и сделать его более эффективным в условиях ограниченных ресурсов. Уменьшение размера изображений, использование аугментации и оптимизация размера батча обеспечивают гибкость в настройке модели и помогают добиться высокого качества сегментации при низких вычислительных затратах. Использование оптимизатора AdamW, а также регуляризации с помощью Dropout и Batch Normalization позволяет значительно ускорить обучение модели Deeplabv3 и улучшить ее обобщающую способность. Эти методы уменьшают риск переобучения и помогают модели стабильно достигать высоких показателей качества на сложных данных.

Литература

[1] Usmani I.A., Qadri M.T., Zia R., Alrayes F.S., Saidani O., Dashtipour K. Interactive effect of learning rate and batch size to implement transfer learning for brain tumor classification. *Electronics*, 2023, vol. 12, art. no. 964. https://doi.org/10.3390/electronics12040964

[2] Zhou P., Xie X., Lin Z., Yan S. Towards Understanding Convergence and Generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46, vol. 9, pp. 6486–6493. https://doi.org/10.1109/TPAMI.2024.3382294

 ISSN 2541-8009

[3] Kolarik M., Burget R., Riha K. Comparing normalization methods for limited batch size segmentation neural networks. *43rd International Conference on Telecommunications and Signal Processing (TSP)*, Milan, Italy, 2020, pp. 677–680. https://doi.org/10.48550/arXiv.2011.11559

[4] Gou J., Yu B., Maybank S.J. et al. Knowledge distillation: a survey. *Int. J. Comput. Vis.*, 2021, vol. 129, pp. 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

[5] Shorten C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *J. big data*, 2019, vol. 6. https://doi.org/10.1186/s40537-019-0197-0

[6] Cordts M., Omran M., Ramos S., Scharwächter T., Enzweiler M., Benenson R., Franke U., Roth S., Schiele B. *The Cityscapes dataset for semantic urban scene understanding*. URL: https://markus-enzweiler.de/downloads/publications/cordts15-cvprws.pdf

[7] van de Ven G.M., Tuytelaars T., Tolias A.S. Three types of incremental learning. *Nat Mach Intell.*, 2022, vol. 4, pp. 1185–1197. https://doi.org/10.1038/s42256-022-00568-3

[8] Liang T., Glossner J., Wang L., Shi S., Zhang X. Pruning and quantization for deep neural network acceleration: a survey. *Neurocomputing*, 2021, vol. 461, pp. 370–403. https://doi.org/10.1016/j.neucom.2021.07.045

Поступила в редакцию 09.06.2025

Малышев Павел Викторович — студент кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Локтев Даниил Алексеевич, доктор технических наук, доцент кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Малышев П.В. Комплексная оптимизация Deeplabv3 для мобильных устройств с применением облегченных архитектур. *Политехнический молодежный журнал*, 2025, № 06 (101). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1077.html

COMPREHENSIVE OPTIMIZATION OF DEEPLABV3 FOR MOBILE DEVICES USING LIGHTWEIGHT ARCHITECTURES

P.V. Malyshev pasha_malyshev01@mail.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Deep convolutional neural networks (CNNs) have become a mainstay for image segmentation tasks, making them indispensable for a variety of applications including autonomous systems and mobile devices. However, high computational resource requirements make the use of models such as Deeplabv3 difficult on resource-constrained devices. In this paper, we investigate how Deeplabv3 can be optimized by replacing ResNet-50 with lighter architectures such as ResNet-18 and MobileNetV2. The reduction in resource consumption without significant loss of model accuracy using the lighter architectures is investigated.

Keywords: convolutional neural networks, segmentation, digital image processing, deep learning, optimization

Received 09.06.2025

Malyshev P.V. — Student of Information Systems and Telecommunications Department, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Loktev D.A., Dr. Sc. (Eng.), Assoc. Professor, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Malyshev P.V. Comprehensive optimization of Deeplabv3 for mobile devices using lightweight architectures. *Politekhnicheskiy molodezhnyy zhurnal*, 2025, no. 06 (101). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1077.html

 ISSN 2541-8009