EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

ИЕРАРХИЧЕСКИЕ БИНАРНЫЕ CNN ДЛЯ ЛОКАЛИЗАЦИИ ДОСТОПРИМЕЧАТЕЛЬНОСТЕЙ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Мадаминов Хайдар Худаярович[1]
PhD, доцент,
[1]ТУИТ имени Мухаммада ал-Хоразмий
Узбекистан, г. Ташкент
Худайберганов Журабек Давлатбоевич[2]
докторант,
[2]ТУИТ имени Мухаммада ал-Хоразмий
Узбекистан, г. Ташкент
jurabekkhudayberganov1991@gmail.com
Каримова Айкерим Отесиновна[3]
ассистент,
[3] НФ ТУИТ имени Мухаммада ал-Хоразмий
Каракалпакстан, г.Нукус
karimovaajkerim4@gmail.com
Ешниязова Гоззал Бахтияровна[4]
ассистент,
[4]НФ ТУИТ имени Мухаммада ал-Хоразмий
Каракалпакстан, г.Нукус
gozzal1115@gmail.com
https://doi.org/10.5281/zenodo.7107448

ARTICLE INFO

Received: 17th September 2022
Accepted:19th September 2022
Online: 23rd September 2022

KEY WORDS

Binary Convolutional Neural Networks, Residual learning, Landmark localization, Human pose estimation, Face alignment.

ABSTRACT

Аннотация. Наша цель — разработать архитектуры, которые сохранят новаторскую производительность сверточных нейронных сетей (CNN) для ориентировочной локализации и в то же время будут легкими, компактными и подходящими для приложений с ограниченными вычислительными ресурсами.

Введение

Эта работа посвящена локализации предопределенного набора реперных точек на интересующих объектах, которые обычно могут подвергаться нежестким деформациям, таким как человеческое тело или лицо. Совсем недавно работа, основанная на сверточных нейронных сетях (CNN), произвела революцию в локализации ориентиров, продемонстрировав результаты замечательной точности даже на самых сложных наборах данных для оценки позы человека [1], [2], [3] и выравнивания лица [4]. Однако развертывание (и обучение) таких методов требует больших вычислительных ресурсов и требует одного или нескольких высокопроизводительных графических процессоров, в то время как для обучения моделей обычно требуются сотни МБ, что делает их совершенно

непригодными для приложений реального времени или мобильных приложений. Эта работа посвящена высокоточной и надежной, но эффективной и легкой локализации ориентиров с использованием бинаризованных CNN. Наша работа вдохновлена недавними результатами бинарных архитектур CNN по классификации изображений [5], [6]. В отличие от этих работ, мы первыми изучили влияние бинаризации нейронной сети на мелкие задачи, такие как локализация ориентиров. Как и в [5], [6], мы обнаружили, что бинаризация приводит к падению производительности, однако для решения этой проблемы мы решили исследовать и предложить несколько архитектурных инноваций, которые привели к введению нового иерархического, параллельного и многомасштабного остаточного блока, в отличие от исследования способов улучшения процесса бинаризации, предложенного в [5], [6]. Таким образом, наш основной методологический вклад:

Мы первыми изучили влияние бинаризации на современные архитектуры CNN для решения проблемы локализации, а именно оценки позы человека и выравнивания лица. С этой целью мы тщательно оцениваем различные варианты дизайна и выявляем узкие места в производительности.

Предлагаемая иерархическая параллельная и многоуровневая структура: наш блок увеличивает размер рецептивного поля, улучшает градиентный поток, специально разработан, чтобы иметь (почти) то же количество параметров, что и исходное узкое место, не содержит сверток 1×1, и в целом выводится с точки зрения повышения производительности и эффективности бинарных сетей.

Основываясь на нашем анализе, мы предлагаем новую иерархическую, параллельную и многомасштабную остаточную архитектуру, специально предназначенную для работы в бинарном случае. Наш блок приводит к значительному улучшению производительности по сравнению с базовым бинарным остаточным блоком из [7] (около 6% в абсолютном выражении при использовании того же количества параметров). На рис. 1 показано сравнение между базовым остаточным блоком из [7] и блоком, предложенным в данной работе.

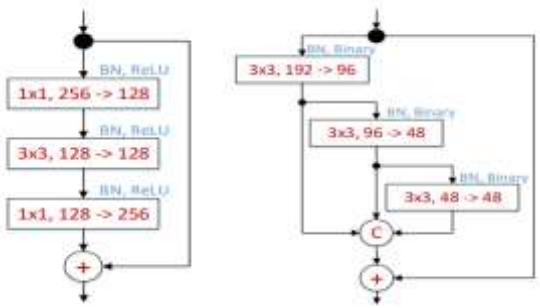

Рис. 1. (a) Исходный слой узкого места [7]. (b)

Обзор состояния дел в области оценки позы человека и выравнивания лица.

Бинаризация (т.е. крайний случай квантования) долгое время считалась непрактичной из-за деструктивного свойства такого представления [10]. Недавно [14] показал, что это не так, и что путем квантования до {−1, 1} можно действительно получить хорошие результаты. [15] представляет новый метод обучения CNN, который использует двоичные веса как для прямого, так и для обратного прохода, однако во время обучения по-прежнему требуются реальные параметры. Работа [6] идет еще дальше и бинаризирует как параметры, так и активации. В этом случае умножения можно заменить элементарными бинарными операциями [6]. Оценка бинарных весов с помощью коэффициента масштабирования [5] является первой работой, сообщающей о хороших результатах на большом наборе данных (ImageNet).

Блок-схема- предлагаемый метод использует архитектуру на основе невязок, поэтому отправной точкой нашей работы является блок узкого места, описанный в [7], [11]. Совсем недавно в [10] исследуется идея увеличения мощности остаточного блока путем разбиения его на серию с параллельных (и намного меньших, чтобы количество параметров оставалось примерно одинаковым) подблоков с той же топологией, которые ведут себя как ансамбль. Помимо слоев узких мест. [12] предлагают начальный блок, который вводит параллельные пути с различными размерами рецептивного поля и различными способами уменьшения количества параметров путем разложения сверточных слоев с большими фильтрами на более мелкие. В последующей статье [14] авторы представляют ряд начальных остаточных архитектур. Последняя работа наиболее связана с предлагаемым методом.

Сетевой дизайн- Нашей целью было не предложить новую сетевую архитектуру для локализации памятников; поэтому мы использовали сеть Hour-Glass (Hour-Glass, HG) из [2], в которой используется блок узкого места из [13]. Поскольку нас интересует эффективность, большинство наших экспериментов проводится с использованием одной сети. Нашей базовой линией был одиночный бинарный HG, полученный путем его прямого квантования с использованием [5]. Как видно из таблицы 1, существует значительный разрыв в производительности между бинарными и реальными значениями HG. Мы устраняем этот пробел, заменяя блок узкого места, использованный в оригинальной HG, предлагаемым блоком.

Оценки позы человека (Human Pose Estimation)- Традиционно методы оценки позы человека основывались на графических моделях с древовидной структурой для представления пространственных отношений между частями тела и обычно строились с использованием элементов, созданных вручную. Совсем недавно методы, основанные на CNN, показали замечательные результаты, значительно превосходящие традиционные методы. Поскольку

EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

изучение прямого сопоставления изображения с расположением частей тела является очень нелинейной задачей, которую трудно освоить, большинство методов представляют каждый ориентир в виде карты достоверности, закодированной как двумерная гауссова с центром в местоположении ориентира, и принимают полностью сверточная структура. Кроме того, вместо того, чтобы делать однократные прогнозы, почти все методы следуют каскадному подходу, делая ряд промежуточных прогнозов, уточняемых последовательно [1], [2], [3]. Примечательно, что для дальнейшего сокращения количества параметров каскадных подходов метод, представленный в [9], использует рекуррентную нейронную сеть. При достижении замечательной производительности все вышеупомянутые методы глубокого обучения требовательны к вычислительным ресурсам и требуют как минимум одного высокопроизводительного графического процессора.

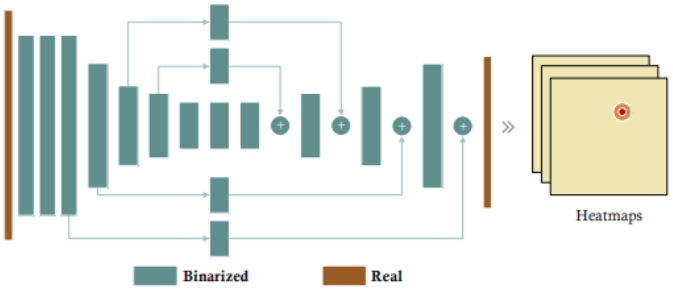

Рис. 2. Архитектура одиночной сети Hour-Glass (HG) [2]. Следуя [5], первый и последний слои (коричневый цвет) остаются реальными, а все остальные слои бинаризуются.

Улучшенная сетевая архитектура

В этом разделе мы исследуем ряд архитектурных изменений, примененных к общей структуре сети. Во-первых, вдохновившись [8], мы упрощаем модель HG, улучшая ее производительность без ущерба для точности для бинарного случая. Затем мы изучаем эффект объединения нескольких сетей и анализируем их поведение.

Улучшенная архитектура HG

Руководствуясь выводами подраздела 4.5, которые проливают свет на важность градиентного потока и предполагают, что по возможности следует использовать пропущенные соединения с более короткими путями, мы применяем аналогичный подход к общей архитектуре HG. В частности, чтобы улучшить общий поток градиента, мы удалили остаточные блоки в ветвях восходящей выборки, которым поручено «внедрение» информации высокого разрешения на более поздние этапы сети. Чтобы приспособиться к этому изменению, количество входных каналов первого слоя из модулей, которые находятся сразу после точки, где ветвь объединяется путем конкатенации, увеличивается в два раза (с учетом увеличения количества каналов). Полученная архитектура, изображенная на рис. 4, представляет собой модифицированную архитектуру U-net

EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

[4], которая была преобразована в бинарную форму так же, как модель HG. Результаты, представленные в таблице 1, показывают, что за счет удаления остаточных блоков из ветвей повышающей дискретизации производительность по сравнению с базовой HG увеличивается на 0,5%, что еще больше подтверждает важность градиентного потока в производительности бинарных сетей. Кроме того, за счет уменьшения количества слоев и параметров наблюдается ускорение до 20%. Сеть обучается с использованием той же процедуры, описанной ранее, для 100 эпох.

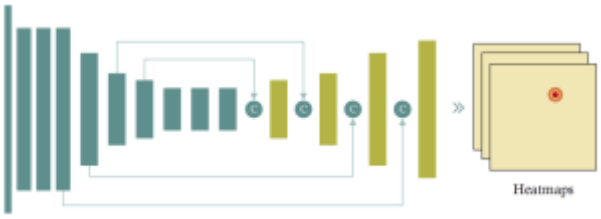

Рис. 3. Улучшенная архитектура HG, вдохновленная U-Net. Темно-зеленые модули остались без изменений, а для светло-зеленых мы удвоили количество их входных каналов с 256 до 512.

Таблица 1

Сравнение между HG и улучшенным HG на проверочном наборе MPII. Обе сети построены с использованием предложенного нами бинарного блока.

Сетевая архитектура	#parametrs	PCKh
HG	6.2 M	76%
Improved HG	5.8 M	76.6%

Стекированные бинаризованные сети HG

Недавно было показано, что наложение сети позволяет достичь самых современных результатов в оценке позы человека [1], [2], [3], когда используются модели с действительными значениями. В этом подразделе мы исследуем, верно ли то же самое для бинарного случая. Следуя [2], мы объединяем и соединяем сети следующим образом: первая сеть принимает в качестве входных данных изображение RGB и выводит набор из N тепловых карт. Следующая сеть в стеке принимает на вход сумму: (1) ввода в предыдущую сеть, (2) проекции ранее предсказанных тепловых карт и (3) вывода предпоследнего блока с предыдущего уровня. Результирующая сеть для стека из двух показана на рис. 5.

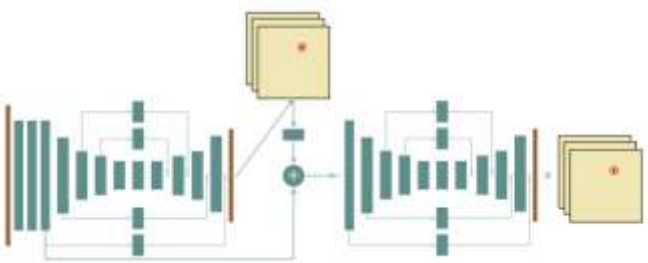

Рис. 4. Двухстековый бинаризованный HG. Все блоки бинаризованы, кроме самого первого и последнего слоев, выделенных красным цветом.

EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

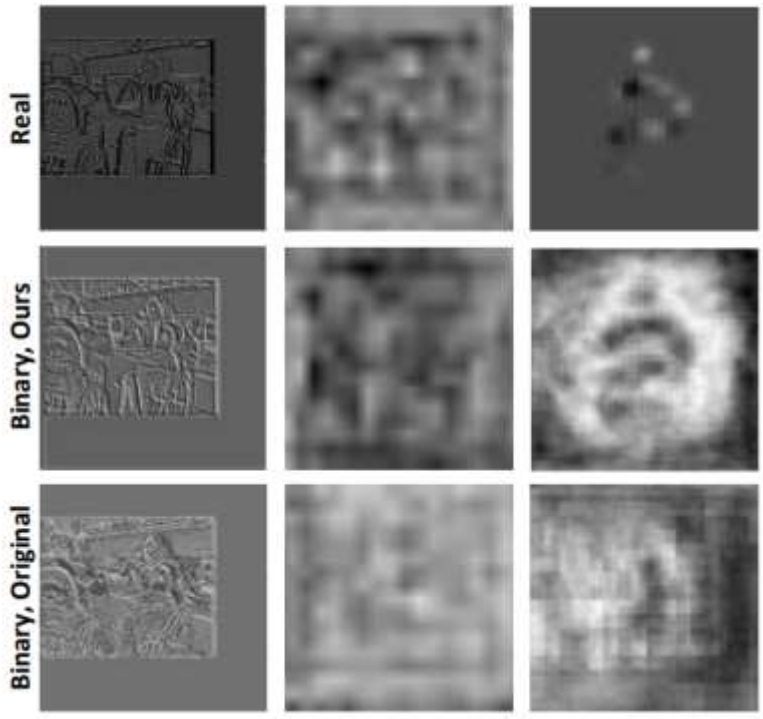

Рис. 5. Признаки, извлеченные из первого слоя (первый столбец), среднего слоя (средний столбец) и непосредственно перед самым последним слоем (правый столбец) для действительных и бинарных (наших и исходных) сетей.

По мере того, как мы переходим к последним слоям, активация становится более шумной для двоичного случая, что, по нашему мнению, снижает производительность многоуровневых сетей.

Как показывают результаты Таблицы 2, стекирование сети для бинарного случая ведет себя в некоторой степени так же, как и для случая с действительным значением, однако выигрыш от одного этапа к другому меньше, и производительность насыщается быстрее. Мы считаем, что основная причина этого заключается в том, что в случае бинарных сетей активация является более шумной, особенно для последних слоев сети. Это показано на рис. 4, где мы сравниваем карты признаков, полученные из реальной сети, и двух типов бинарных сетей, сравниваемых в этой статье (исходной, основанной на узком месте и предложенной). Очевидно, что карты признаков для бинарного случая становятся более зашумленными и размытыми по мере того, как мы переходим к последним слоям сети. Поскольку сетевое стекирование основано на функциях более ранних сетей каскада и поскольку они являются шумными, мы пришли к выводу, что это оказывает негативное влияние на общую производительность сети.

Подготовка. Чтобы ускорить процесс обучения, мы обучали составную версию последовательно. Сначала мы обучали первую сеть до сходимости, затем добавляли поверх нее вторую, замораживая ее веса и обучая вторую. Процесс повторяется до тех пор, пока не будут добавлены все сети. Наконец, весь стек обучается совместно в течение 50 эпох.

Дополнительные эксперименты

В этом разделе мы также покажем, что предложенный блок хорошо обобщает, обеспечивая

согласованные результаты для различных наборов данных и задач. С этой целью мы сообщаем о результатах задачи анализа лица, также известной как семантическая сегментация частей лица, которая представляет собой задачу присвоения категориальной метки каждому пикселю изображения лица. Мы создали набор данных для сегментации лицевых частей, объединив вместе 68 наземных ориентиров истинности (изначально предоставленных для выравнивания лица), чтобы полностью охватить каждый лицевой компонент. Всего мы создали семь классов: кожа, нижняя губа, верхняя губа, внутренняя часть рта, глаза, нос и фон. На рис. 14 показан пример наземной маски истинности. Мы обучили сеть на наборе данных мощностью 300 Вт (примерно 3000 изображений) и протестировали ее на тестовом наборе для соревнований мощностью 300 Вт, как в помещении, так и на улице (600 изображений), используя ту же процедуру, что и в таблице 2.

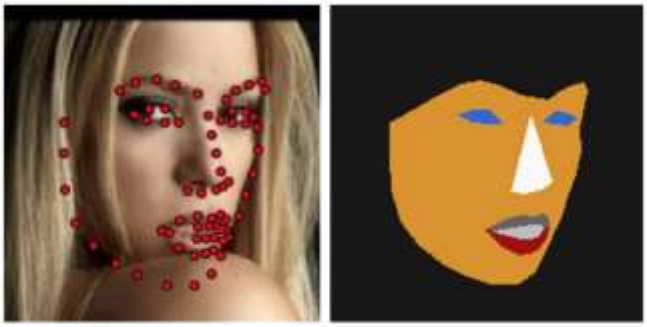

Рис. 6. Пример маски наземной истинности (справа), полученной путем объединения 68 наземных ориентиров (слева). Каждый цвет обозначает один из семи классов.

Архитектура. Мы повторно использовали ту же архитектуру для локализации ориентиров, изменив только последний слой, чтобы учесть другое количество выходных каналов (от 68 до 7). Мы сообщаем результаты для трех различных представляющих интерес сетей: (а) сеть с действительным значением, использующая исходный блок узкого места (называемый «Реальное, узкое место»), (б) бинарная сеть, использующая исходный блок узкого места (называемый «Двоичный, узкое место»).) и (с) бинарная сеть, использующая предлагаемый блок (называемый «Двоичный, наш»). Чтобы обеспечить справедливое сравнение, все сети имеют одинаковое количество параметров и глубину. Для обучения сетей мы использовали потери Log-Softmax [12]. Полученные результаты. В таблице 2 представлены полученные результаты. Как и в наших экспериментах по оценке позы человека и выравниванию лица, мы наблюдаем, что бинаризованная сеть, основанная на предложенном блоке, значительно превосходит сеть аналогичного размера, построенную с использованием исходного блока узкого места, почти совпадая с производительностью сети с реальным значением. Большая часть улучшения производительности связана с более высокой способностью представления /

обучения нашего блока, что особенно очевидно для сложных случаев, таких как необычные позы, окклюзии или сложные условия освещения. Для визуального сравнения см. рис. 7.

Таблица 2

Результаты на 300 Вт (в помещении и на улице). Пиксель акк., средний акк. и средние IU вычисляются, как в [13].

Networtk type	pixel.acc.	mean.acc.	mean IU
Real, bottleneck	97,98%	77,23%	69,29%
Binary, bottleneck	97,41%	70,35%	62,49%
Binary, Ours	97,91%	76,02%	68,05%

Вывод

Мы предложили новую блочную архитектуру, специально предназначенную для бинарных CNN и визуальных задач локализации. В ходе этого процесса мы тщательно оценили различные варианты дизайна, выявили узкие места в производительности и предложили решения. Мы показали, что наш иерархический, параллельный и многоуровневый блок увеличивает репрезентативную силу, позволяя изучать более сильные отношения без чрезмерного увеличения количества сетевых параметров.

(а) Примеры подгонки, созданные нашей бинарной сетью на наборе данных AFLW2000-3D. Обратите внимание, что наш метод хорошо справляется с экстремальными позами, выражениями лица и условиями освещения.

(б) Примеры поз человека, полученные с помощью нашей бинарной сети. Обратите внимание, что наш метод дает хорошие результаты для самых разных поз и окклюзий.

Рис. 7. Качественные результаты, полученные нашим методом на наборах данных AFLW2000-3D (а) и MPII (б).

Предлагаемая архитектура эффективна и может работать на ограниченных ресурсах. Мы проверили эффективность предложенного блока в широком спектре мелких задач распознавания, включая оценку позы человека, выравнивание лица и сегментацию частей лица

References:

[1] Begmatov Sh.A., Arabboyev M.M, Xudayberganov J.D. “ Google media-pipe kutubxonasidan foydalangan holda inson gavdasi harakatlari farqlarini baholash” Muhammad al-Xorazmiy avlodlari 3(21)2022.

[2] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in ECCV, 2016.
[3] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in CVPR, 2016.
[4] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge," in ECCV. Springer International Publishing, 2016, pp. 616–624.
[5] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnornet: Imagenet classification using binary convolutional neural networks," in ECCV, 2016.
[6] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv, 2016.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in ECCV, 2016.
[8] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in ICCV, 2017.
[9] J. L. Holi and J.-N. Hwang, "Finite precision error analysis of neural network hardware implementations," IEEE Transactions on Computers, vol. 42, no. 3, pp. 281–290, 1993.
[10] M. Courbariaux, Y. Bengio, and J.-P. David, "Training deep neural networks with low precision multiplications," arXiv, 2014.
[11] D. D. Lin, S. S. Talathi, and V. S. Annapureddy, "Fixed point quantization of deep convolutional networks," arXiv, 2015.
[12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.
[13] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," arXiv, 2016.
[14] D. Soudry, I. Hubara, and R. Meir, "Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights," in NIPS, 2014.
[15] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in NIPS, 2015.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016.
[17] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated ´ residual transformations for deep neural networks," arXiv, 2016.