Совершенствование механизмов внимания для архитектуры трансформер в задачах повышения качества изображений

Н.И. Бережнов [1], А.А. Сирота [1]
[1]Воронежский государственный университет,
394018, Воронеж, Россия, Университетская пл., д. 1

Аннотация

Рассматривается задача улучшения качества изображений, полученных в условиях воздействия различного рода шумов и искажений. В работе для решения указанной задачи описаны нейросетевые модели-трансформеры, показавшие в последнее время высокую эффективность в задачах компьютерного зрения. Исследуется механизм внимания моделей-трансфомеров, и определяются проблемы при реализации существующих подходов, основанных на данном механизме. Предлагается собственная модификация механизма внимания с целью уменьшения количества параметров нейронной сети, и проводится сравнение предлагаемой на ее основе модели-трансформера с известными. Рассматривается несколько датасетов с естественными и искусственно внесенными искажениями. При обучении нейронных сетей для сохранения резкости изображений в процессе устранения шумов используется функция потерь EdgeLoss. Исследуется влияние степени сжатия канальной информации в предлагаемом механизме внимания на качество восстанавливаемых изображений. Для оценки качества восстанавливаемых изображений применяются метрики PSNR, SSIM, FID, на основе которых проводится сравнение существующих архитектур нейронных сетей на каждом из использованных датасетов. Установлено, что предлагаемая авторами архитектура в целом как минимум не уступает известным по качеству улучшенных изображений, но при этом требует меньшего количества вычислительных ресурсов. Показано, что качество улучшенных изображений падает незначительно для невооруженного человеческого глаза при увеличении коэффициента сжатия канальной информации в разумных пределах.

Ключевые слова: улучшение качества изображения, нейронные сети, модели-трансформеры, механизм внимания.

Цитирование: **Бережнов, Н.И.** Совершенствование механизмов внимания для архитектуры трансформер в задачах повышения качества изображений / Н.И. Бережнов, А.А.Сирота // Компьютерная оптика. – 2024. – Т. 48, № 5. – С. 726-733. – DOI: 10.18287/2412-6179-CO-1393.

Citation: Berezhnov NI, Sirota AA. Improving attention mechanisms in transformer architecture in image restoration. Computer Optics 2024; 48(5): 726-733. DOI: 10.18287/2412-6179-CO-1393.

Введение

Задача улучшения качества изображений остаётся актуальной по сей день. С каждым годом растут возможности цифровых устройств, а следовательно, увеличиваются требования, предъявляемые к качеству изображений, используемых в них. Это особенно важно для медицинских и аэрокосмических снимков, приложений с дополненной реальностью, фотографии и киноиндустрии.

На сегодняшний день можно выделить несколько основных задач в области повышения качества изображений [1]:

- повышение разрешения и резкости изображений;
- устранение аддитивных и аппликативных помех;
- восстановление изображений, полученных в плохих погодных условиях;
- устранение JPEG-артефактов, возникающих при компрессии изображений;
- комплексное улучшение качества изображений в целом в условиях комбинирования различных типов искажений.

В настоящей работе основное внимание уделяется восстановлению изображений от аддитивных и аппликативных помех и повышению качества изображений, полученных в плохих погодных условиях как реалистического примера комбинирования различных типов искажений.

В последнее время для решения различных задач в области компьютерного зрения всё большую популярность набирает использование моделей-трансформеров (vision transformer, ViT), впервые описанных в работе [2]. Данная архитектура появилась в области обработки естественного языка (NLP), где показала свою эффективность с помощью реализации механизма внимания (самовнимания) [3]. В настоящее время проводятся многочисленные исследования, направленные на сравнение трансформеров со сверточными нейронными сетями (СНС). Показано, что механизм внимания, реализуемый в ViT, в функциональном плане во многом эквивалентен операции свертки [4]. При этом по эффективности модели-трансформеры во многих случаях могут давать лучшие результаты по сравнению со сверточными се-

тями с гораздо большим количеством слоев в задачах классификации, семантической сегментации, повышения качества изображений.

В связи с этим основной целью данной работы явилось исследование и совершенствование механизмов внимания в архитектуре трансформеров в интересах снижения вычислительной сложности алгоритмов улучшения изображений с сохранением обобщающей способности нейронной сети без существенной потери качества улучшенных изображений.

Обзор существующих методов

Одним из важных мест в моделях трансформеров, используемых для улучшения изображений, как уже упоминалось, является реализация механизма внимания. Основная цель его состоит в том, чтобы выбрать наиболее значимые признаки из исходных изображений, на которых нужно сосредоточиться и которые наиболее важны в ходе последующей обработки для повышения эффективности улучшения изображений. В большинстве случаев используется его разновидность – самовнимание, когда все значения берутся непосредственно с выхода одного слоя нейронной сети путем умножения на соответствующие матрицы коэффициентов [5].

В первоначальной архитектуре ViT изображение разбивается на блоки 16×16. Из-за того, что модели-трансформеры инвариантны к перестановкам пикселей, авторы [2] добавляют ещё позиционное кодирование для понимания взаимного расположения блоков. Далее используется стандартный для NLP блок многоголового внимания, на выходе которого применяются полносвязные слои для решения задачи классификации.

Исследователями было выявлено, что данная архитектура показывает невысокие результаты в задачах улучшения качества изображений, сегментации и детектирования объектов [1]. Одна из основных проблем – разделение на блоки 16×16 пикселей, что приводит к снижению точности результатов. Однако уменьшение размерности блоков нежелательно, так как это существенно увеличивает вычислительную сложность нейронной сети.

Авторы работы [6] предлагают собственный блок swin-трансформер (shifted windows transformer) для реализации локального механизма внимания с использованием сканирующего окна, а также добавлением обучаемых параметров для кодирования блоков изображения. При этом уменьшается размер блоков до 4×4 пикселей, а также добавляется иерархичное их объединение. Это позволяет сделать постоянным число признаков на разных масштабах, поступающих на вход механизма внимания. Данная архитектура уже успешно использовалась в задачах обнаружения объектов и семантической сегментации [7 – 8].

В работе [9] предлагается усовершенствованный механизм внимания трансформеров для задачи семантической сегментации изображений. Авторы решают уменьшить вычислительную сложность блока внимания путём сокращения размерности изображения в R раз с помощью так называемой разделяемой по глубине (depthwise) свертки с таким же шагом R.

В работе [10] swin-трансформеры используются в задаче устранения шумов и искажений на изображениях. Здесь авторы предлагают архитектуру SUNet, построенную на подобии и принципах архитектуры UNet. Они оставляют только сверточные слои на входе и на выходе, а также блоки повышения и понижения дискретизации. Все остальные блоки представляют собой swin-трансформеры.

В контексте рассматриваемой задачи возможно использование моделей-трансформеров в генеративно-состязательных сетях. Авторы [11] при помощи двух дискриминаторов с остаточными связями от генератора улучшают качество изображений при низком освещении. Помимо этого, на вход нейронной сети добавляется градиент изображения, как априорная информация о его структуре. Это позволяет успешно восстанавливать текстурные части изображения.

В работе [12] авторы для повышения разрешения изображений применяют блок MDTA (multi-Dconv head transposed attention), который использует межканальную ковариацию для получения оптимальных карт внимания. Преимущество MDTA заключается в том, что он использует глобальные взаимосвязи между пикселями изображения и оптимизирует локальный контекст для выделения признаков с целью последующей обработки в плане улучшения качества изображений. Помимо этого, авторы используют на выходе из MDTA depthwise-свертки с вентильным механизмом на основе функции активации GELU вместо общепринятых полносвязных слоёв.

В работе [13] авторы предлагают архитектуру Transweather – модель-трансформер, ориентированную на восстановление изображений в плохих погодных условиях. Авторы используют одну модель и обучают нейронную сеть одновременно для различных видов погодных осадков.

Авторы большой работы [14] борются с проблемой нехватки данных через самообучение (Self-Supervised Learning), когда важно не столько решение самой задачи повышения качества изображений, сколько выделение признаков, которые будут получены в ходе её решения. Выделенные признаки можно в дальнейшем использовать при обучении нейронной сети в задачах с маленьким набором размеченных данных.

Помимо этого, многие исследователи используют технику переноса обучения [15], когда нейронная сеть обучается в два этапа: сначала на большом размеченном датасете, а потом уже на относительно небольшом, ориентированном на конкретную задачу. Также применяются различные методы аугментации данных.

Описанные выше подходы к обучению нейронных сетей используются и в данной работе. При этом при подготовке обучающих данных авторы применяют собственные алгоритмы аугментации данных [16], позволяющие генерировать различные виды шумов и моделировать реальные искажения.

Изначально в работе [1] выделено несколько основных проблем при использовании трансформеров для решения задач компьютерного зрения:

1. Длительный процесс обучения и сложная интерпретируемость результатов.
2. Квадратичная вычислительная сложность относительно числа пикселей изображения из-за использования механизмов внимания.
3. Необходимость использования большого набора данных для обучения.

На сегодняшний день эти проблемы остаются актуальными. Описанные в известных работах подходы отчасти решают проблему производительности и получения требуемого количества обучающих примеров, но не всегда позволяют использовать трансформеры в реальном времени на небольшом наборе данных в задачах машинного зрения.

В этом плане авторы настоящей работы предлагают модифицированный механизм внимания для уменьшения вычислительной сложности нейронной сети и увеличения её быстродействия при сохранении относительно высоких показателей эффективности реализуемой обработки изображений.

Предлагаемая архитектура нейронной сети

Помимо обычного механизма внимания, исследователи используют его разновидность – механизм самовнимания, когда все значимые для обработки признаки берутся из одного источника. Рассмотрим подробнее этот механизм, используемый в моделях-трансформерах при обработке изображений. Пусть есть нормализованный тензор I размерами $H\times W\times C$, который подаётся на вход слоя нейронной сети с механизмом самовнимания. Для традиционного в трансформерах разделения тензора I на K *(key),* Q *(query),* V *(values)* будем использовать depthwise-свертку, как и авторы [12]. Тогда $K=W_k I$, $Q=W_q I$, $V=W_v I$, где W_k, W_q, W_v – обучаемые матрицы соответствующих весовых коэффициентов сверточных слоев нейронной сети. Реализуемый механизм самовнимания можно описать следующей формулой:

$$Attention = SoftMax\left(\frac{QK^{\mathrm{T}}}{t}\right)\cdot V, \qquad (1)$$

где K, Q, V – тензоры размерами $H\times W\times C$, t – нормировочный коэффициент, который может быть обучаемым параметром нейронной сети.

Обозначим $H\times W$ как N. Согласно формуле (1) необходимо перемножить три матрицы размерами: $N\times C$, $N\times C$, $N\times C$. В результате получаем вычислительную сложность для двух операций умножения $O(N^2C+NC^2)$.

В настоящей работе реализован механизм самовнимания, применяемый для канальной информации входного тензора I. При этом предлагается модифицировать механизм самовнимания путем добавления специального коэффициента сжатия канальной информации R через поточечную (pointwise) свертку. В результате итоговая вычислительная сложность становится равной $O(N^2C/R+NC^2/R)$: происходит уменьшение вычислительной сложности механизма самовнимания в R раз. Варьирование параметра R позволяет подбирать целесообразные его значения, обеспечивающие уменьшение вычислительной сложности при незначительном снижении качества восстановления изображений.

В итоге также уменьшается количество обучаемых параметров нейронной сети. При использовании в стандартных вариантах depthwise-сверток количество параметров остаётся неизменным, в то время как для pointwise-свёрток получаем сокращение параметров в R раз за счёт уменьшения размерности канальной информации. Количество обучаемых параметров нейронной сети в зависимости от коэффициента сжатия представлено в табл. 1

Табл. 1. Количество параметров нейронной сети

Степень сжатия	Количество параметров
$R=1$	26,111,668
$R=2$	22,379,476
$R=4$	20,513,380
$R=8$	19,580,332

Блоки нейронных сетей, в которых происходит сокращение канальной информации с последующим её увеличением, именуется в литературе [17] bottleneck. Отсюда, данную модификацию механизма внимания модели-трансформера предлагается назвать channel bottleneck self-attention mechanism, в дальнейшем – CBSA-механизм. Концептуально стандартный блок CBSA представлен на рис. 1.

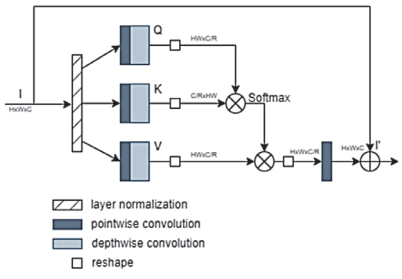


Рис. 1. Стандартный блок CBSA

Также необходимо было добавить нормализацию, чтобы увеличить сходимость и устойчивость нейронной сети.

Полная предлагаемая архитектура нейронной сети представлена на рис. 2. Вместо позиционного коди-

рования использовался стандартный сверточный слой. Было показано в [12], что использование depthwise сверток вместо полносвязных слоёв, следующих за механизмом внимания, значительно улучшает точность восстановления, поэтому этот же подход решено было применить в данной работе.

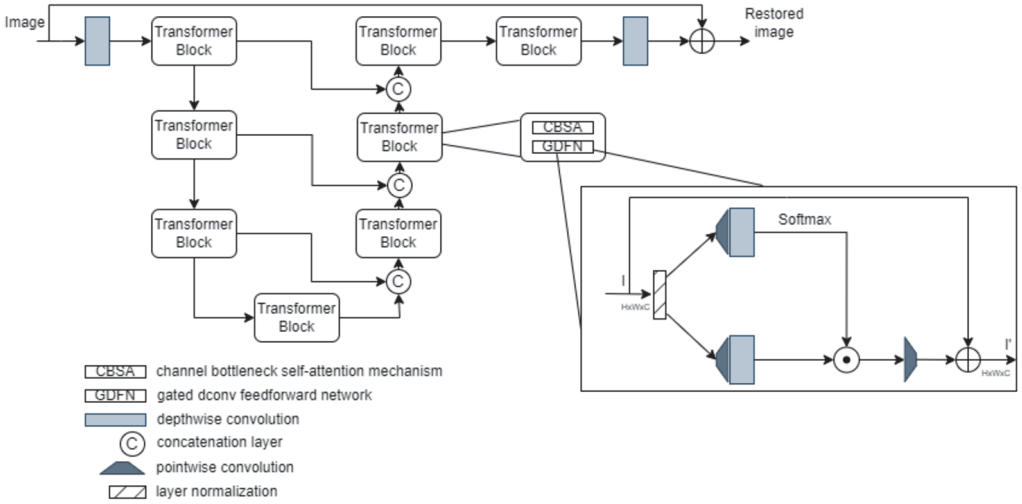


Рис. 2. Предлагаемая архитектура модели-трансформера

В работе [18] показано, что использование разделяемых свёрток на ранних слоях нейронной сети может привести к падению точности. Поэтому решено было использовать на первых двух блоках модели-трансформера обычные свертки.

Между блоками трансформера включены остаточные связи, необходимые для увеличения сходимости нейронной сети при обучении и для устранения проблемы затухания градиента, что очень критично для архитектур с большим количеством слоёв. Изображение, полученное на выходе нейронной сети, показанной на рис. 2, складывается с изображением на входе. Это является общепринятой практикой для задач улучшения качества изображений. Нейронной сети проще выучить убирать искажения, нежели формировать само изображение.

Экспериментальная часть

Для проверки эффективности работы предлагаемых алгоритмов были выбраны следующие датасеты:

- *ImageNet* (50 тыс. изображений) с наложенными на него различными искажениями в виде аппликативных и аддитивных помех, сгенерированных авторами в соответствии с методикой, изложенной в [16];
- *SIDD* – набор данных, содержащий 30 000 зашумленных изображений из 10 сцен в различных условиях освещения, полученных с помощью пяти мобильных устройств, а также изображения их эталонов [19];
- *Погодный датасет* с 18 тыс. изображений, полученных в условиях снега, дождя, тумана, для которого в целях аугментации данных было решено использовать библиотеку Albumentations [20], где уже реализованы алгоритмы генерации всевозможных погодных осадков. Был использован датасет из работы [13].

Все изображения приводились к общему разрешению 256×256. Изображения из датасета SIDD могли иметь разрешение порядка 1024×1024, поэтому было решено улучшать их по частям. Предлагаемая архитектура сети может обработать каждую из частей изображения, а затем объединить их в единое изображение посредством билинейной интерполяции.

Для сравнения качества двух изображений: эталона и улучшенного изображения – использовались следующие метрики: SSIM (чем больше, тем лучше), PSNR (чем больше, тем лучше) и FID (чем меньше, тем лучше). SSIM (коэффициент структурного сходства изображений) и PSNR (пиковое отношение сигнал-шум) являются классическими метриками в области обработки изображений. Однако многими исследователями было показано [21], что они не всегда соответствуют человеческому восприятию изображения. Поэтому решено было использовать ещё метрику Frechet Inception Distance (FID). Она основана на расстоянии Фреше между двумя распределениями признаков улучшенных и эталонных изображений. По сути, эта метрика показывает, насколько одно многомерное распределение похоже на другое. Для выделения признаков из изображений используется нейронная сеть InceptionV3, обученная на датасете ImageNet. В явном виде формулу можно записать следующим образом:

$$FID = \sum(\mu_1 - \mu_2)^2 + Tr(C_1 + C_2 - 2\times\sqrt{C_1 C_2}), \qquad (2)$$

где μ_1, μ_2 – вектора математических ожиданий по каждому из признаков эталонного и сгенерированного распределений, C_1, C_2 – матрицы ковариаций, Tr – след матрицы, сумма берётся по всем признакам изображений.

Из формулы можно сделать вывод, что чем меньше её значение, тем ближе друг к другу распределе-

ния. Данная метрика уже зарекомендовала себя в различных задачах компьютерного зрения, поэтому будем оценивать её распределения, полученные из тестовой выборки, состоящей из 500 изображений для каждого датасета.

Предложенная архитектура нейронной сети обучалась на 8 Гб видеопамяти в течение 8 часов. В качестве оптимизатора использовался Adam. В качестве функций потерь выступали *MSE* и *Charbonnier Loss* с добавлением *Edge Loss*, отвечающей за чёткость восстанавливаемых границ, определённых с помощью оператора Лапласа. Функцию потерь можно представить в виде формулы:

$$Loss(X,Y) = chl(X,Y) + w \cdot chl(\Delta X, \Delta Y),$$
$$chl(X,Y) = \sqrt{(X-Y)^2 + e^2}, \quad (3)$$

где Δ – оператор Лапласа, X, Y – улучшенное и эталонное изображения, w – коэффициент значимости функции потерь *Edge Loss*, e – коэффициент робастности для *Charbonnier Loss*.

В ходе экспериментов были установлены оптимальные коэффициенты для функции потерь: $w=1$ и $e=10^{-3}$.

В работе была использована методика прогрессивного обучения, когда нейронная сеть учится улучшать качество изображений, начиная с маленького разрешения, а заканчивая самым большим. Таким образом, она становится адаптированной к различным разрешениям изображений, что является частым случаем в области обработки изображений.

Результаты экспериментов

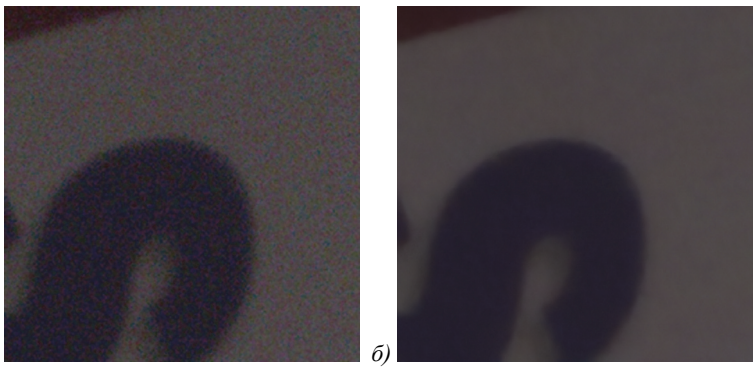
В табл. 2 показаны результаты, полученные на датасете SIDD, при различных коэффициентах сжатия *R*.

Табл. 2. Исследование качества изображений в зависимости от коэффициента сжатия R на датасете SIDD

SIDD	**PSNR**	**SSIM**	**FID**
$R=1$	39,23	0,97	47,42
$R=2$	39,14	0,97	48,17
$R=4$	39,15	0,97	49,46
$R=8$	37,1	0,95	50,36

На рис. 3 показан пример зашумленного и улучшенного изображения на датасете SIDD.

Согласно метрикам оценки качества изображений, можно сделать вывод, что качество улучшенных изображений уменьшается с увеличением коэффициента сжатия *R*. Аналогичный вывод можно сделать на остальных датасетах. Результаты их исследований представлены в табл. 3 и 4.



а) *б)*

Рис. 3. Примеры изображений из датасета SIDD: а) зашумлённая часть изображения; б) улучшенная часть изображения

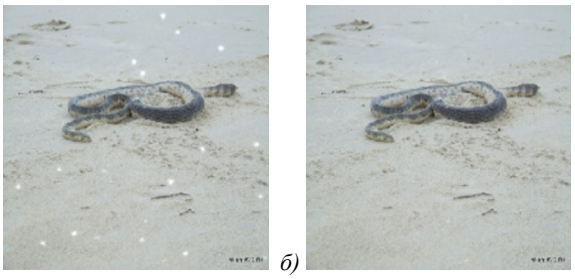
Табл. 3. Исследование качества изображений в зависимости от коэффициента сжатия R на погодном датасете

Погодный датасет	**PSNR**	**SSIM**	**FID**
$R=1$	26,6	0,93	74,42
$R=2$	26,6	0,92	74,75
$R=4$	26,5	0,92	75,46
$R=8$	26,0	0,90	76,46

Табл. 4. Исследование качества изображений в зависимости от коэффициента сжатия R на зашумленном датасете ImageNet

Зашумлённый ImageNet	**PSNR**	**SSIM**	**FID**
$R=1$	38,32	0,75	102,06
$R=2$	38,20	0,74	102,73
$R=4$	38,14	0,74	103,18
$R=8$	38,01	0,74	103,44

В качестве примера демонстрации работы модели-трансформера на рис. 4 показаны зашумлённое и восстановленное от аппликативной помехи изображения.



а) *б)*

Рис. 4. Примеры изображений из датасета ImageNet, а) зашумлённое аппликативной помехой изображение, б) улучшенное изображение

В дальнейшем решено было остановиться на $R=2$, как на оптимальном выборе между качеством и быстродействием обработки изображений. На рис. 5 показано 6 изображений из датасета плохих

погодных условий. Были специально выбраны изображения, содержащие текстурные части. Показано, что с увеличением коэффициента сжатия нейронная сеть хуже справляется с восстановлением текстурной информации. При $R=2$ на изображении становятся малозаметны солнечные лучи, при $R=4$ и $R=8$ на текстурных частях изображения начинают проявляться капли дождя.

Рис. 5. Улучшенные фрагменты изображения из погодного датасета: а) исходное изображение; б) изображение с каплями дождя; в) улучшенное изображение при R = 1; г) улучшенное изображение при R = 8

Сравнение с известными архитектурами

Для сравнения были рассмотрены следующие архитектуры: Restormer [12], SUNet [10], MIRNet [22]. MIRNet – свёрточная нейронная сеть с механизмом внимания как для канальной, так и для пространственной информации. SUNet – модель-трансформер с локальным механизмом внимания с использованием принципа наложения пространственных окон. Restormer – гибридная модель-трансформер, использующая канальный механизм внимания. Также в Restormer применяется блок MDTA (multi-Dconv head transposed attention), в который добавляются дополнительный сверточный слой для получения *K, Q, V* и вентильный механизм на основе функции активации GELU вместо общепринятых полносвязных слоёв.

Обучение этих сетей на используемых в данной работе датасетах осуществлялось на оптимальной конфигурации нейронных сетей Restormer, SUNet, MIRNet, описанных авторами в их статьях и выложенных в открытом доступе.

В табл. 5, 6, 7 приведены результаты исследования описанных выше архитектур нейронных сетей, выделены наилучшие результаты для каждого из датасетов.

Табл. 5. Сравнение на датасете SIDD

SIDD	**PSNR**	**SSIM**	**FID**
Restormer	**40,02**	0,96	**46,12**
SUNet	39,79	0,96	61,32
MIRNet	32,06	0,84	78,29
Авторский вариант ($R=2$)	39,14	**0,97**	48,17

Табл. 6. Сравнение на датасете плохих погодных условий

Погодный датасет	**PSNR**	**SSIM**	**FID**
Restormer	**33,93**	0,91	83,15
SUNet	20,21	0,69	92,42
MIRNet	20,98	0,81	132,5
Авторский вариант ($R=2$)	26,6	**0,92**	**74,75**

Табл. 7. Сравнение на зашумлённом ImageNet

Зашумлённый ImageNet	**PSNR**	**SSIM**	**FID**
Restormer	36,11	**0,80**	**100,7**
SUNet	25,98	0,78	104,60
MIRNet	22,92	0,71	106,0
Авторский вариант ($R=2$)	**38,20**	0,74	102,73

Исходя из результатов, можно сделать вывод, что предлагаемая архитектура не уступает в точности Restormer на заявленных выше датасетах и метриках, а даже в некоторых случаях превосходит. Лучшие результаты выделены в таблицах жирным начертанием. При этом предлагаемая архитектура нейронной сети имеет в 1,5 раз меньше настраиваемых параметров, чем архитектура Restormer. В табл. 8 представлены результаты для датасета SIDD.

Табл. 8. Исследование предлагаемой архитектуры на оптимальность

SIDD	**PSNR**	**SSIM**	**FID**
Функция потерь Chabonier loss	39,01	0,96	48,42
Функция потерь MSE	38,79	0,96	51,33
Без прогрессивного обучения	39,06	0,93	50,27
Добавление двух блоков трансформеров	39,10	0,97	**48,11**
Удаление двух блоков трансформеров	37,56	0,92	53,14
Использование разделяемых сверток во всех блоках трансформеров	38,97	0,96	48,79
Авторский вариант ($R=2$)	**39,14**	**0,97**	48,17

Добавление двух блоков трансформеров: одного – для понижения дискретизации, а второго – для повышения, незначительно улучшает результат. Как уже упоминалось выше, разделяемые свертки работают хуже, чем обычные в начальных блоках модели-трансформера. Предлагаемая функция потерь Edge loss показала себя лучше, чем обычная Chabonier loss. Также можно отметить, что прогрессивное обучение положительно влияет на качество получаемых изображений.

Заключение

В работе была предложена модификация механизма внимания в моделях-трансформерах, показано, как коэффициент сжатия R влияет на качество восстановленных изображений. Можно сделать вывод, что качество результатов нейронной сети падает незначительно, однако вычислительная сложность блоков механизма внимания уменьшается в R раз. По большей части это можно связать с тем, что сжатие положительно влияет на механизм внимания: имея меньше параметров, нейронная сеть учится выделять только самое важное, меньше акцентируя внимание на незначимых признаках.

В дальнейшем авторами планируется продолжить совершенствование механизма внимания в моделях-трансформерах, увеличивать их устойчивость и сходимость при обучении с добавлением механизмов регуляризации и с использованием дистилляции для обучения трансформеров через более простые сверточные нейронные сети. Также было выявлено, что предлагаемая архитектура не полностью устраняет мелкие аппликативные помехи с изображений на датасете ImageNet. В дальнейшем планируется улучшить этот результат.

References

[1] Ali A, Benjdira B, et al. Vision transformers in image restoration: A survey. Sensors 2023; 23(5): 2385. DOI: 10.3390/s23052385.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint. 2020. Source: <https://arxiv.org/abs/2010.11929>.

[3] Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv Preprint. 2017. Source: <https://arxiv.org/abs/1706.03762>. DOI: 10.48550/abs/1706.03762.

[4] Cordonnier J, Loukas A, Jaggi M. On the relationship between self-attention and convolutional layers. arXiv Preprint. 2019. Source: <https://arxiv.org/abs/1911.03584>. DOI: 10.48550/arXiv.1911.03584.

[5] Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition 2020: 10076-10085.

[6] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proc IEEE/CVF Int Conf on Computer Vision (ICCV'2021) 2021: 10012-10022.

[7] Zhang J, Qin Q, Ye Q, Ruan T. ST-Unet: Swin transformer boosted U-Net with cross-layer feature enhancement for medical image segmentation. Comput Biol Med 2023; 153: 106516. DOI: 10.1016/j.compbiomed.2022.106516.

[8] Illarionova S, Shadrin D, Shukhratov I, Evteeva K, Popandopulo G, Sotiriadi, Burnaev E. Benchmark for building segmentation on up-scaled Sentinel-2 imagery. Remote Sens 2023; 15(9): 2347. DOI: 10.3390/rs15092347.

[9] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. arXiv Preprint. 2021. Source: <https://arxiv.org/abs/2105.15203>. DOI: 10.48550/arXiv.2105.15203.

[10] Fan C-M, Liu T-J, Liu K-H. SUNet: Swin transformer UNet for image denoising. 2022 IEEE Int Symp on Circuits and Systems (ISCAS) 2022: 2333-2337. DOI: 10.1109/ISCAS48785.2022.9937486.

[11] Wang C, Pan J, Wu X. Structural prior guided generative adversarial transformers for low-light image enhancement. arXiv Preprint. 2022. Source: <https://arxiv.org/abs/2207.07828>. DOI: 10.48550/arXiv.2207.07828.

[12] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M. Restormer: Efficient transformer for high-resolution

image restoration. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition 2022: 5718-5729. DOI 10.1109/CVPR52688.2022.00564.
[13] Valanarasu JM, Yasarla R, Patel VM. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition 2022: 2353-2363.
[14] Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans Pattern Anal Mach Intell 2021; 43: 4037-4058.
[15] Zhuang F. A comprehensive survey on transfer learning. Proc IEEE 2021; 109: 43-76. DOI: 10.1109/JPROC.2020.3004555.
[16] Berezhnov NI, Sirota AA. Universal image enhancement algorithm using deep neural neworks [In Russian]. Proc Voronezh State University: Systems Analysis and Information Technologies 2022; 2: 81-92. DOI: 10.17308/sait/1995-5499/2022/2/81-92.
[17] He K, et al. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition 2016: 770-778.
[18] Tan M, Le QV. EfficientNetV2: Smaller models and faster training. arXiv Preprint. 2021. Source: <https://arxiv.org/abs/2104.00298>. DOI: 10.48550/arXiv.2104.00298.
[19] Abdelhamed A, Lin S, Brown, MS. A high-quality denoising dataset for smartphone cameras. Proc IEEE Conf on Computer Vision and Pattern Recognition 2018: 1692-1700.
[20] Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: Fast and flexible image augmentations. Information 2020; 11: 125. DOI: https://doi.org/10.3390/info11020125
[21] Huynh-Thu, Q. Scope of validity of psnr in image/video quality assessment. Electron Lett 2013; 44: 800-801. DOI: 10.1049/el:20080522.
[22] Zamir SW, et al. Learning enriched features for real image restoration and enhancement. In Book: Vedaldi A, Bischof H, Brox T, Frahm J-M, eds. Computer Vision – ECCV 2020. Cham: Springer Nature Switzerland AG; 2020: 492-511. DOI: 10.1007/978-3-030-58595-2_30.

Сведения об авторах

Бережнов Н. И. – аспирант кафедры технологий обработки и защиты информации кафедры технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: *beregnovnikita@yandex.ru* ORCID iD: https://orcid.org/0000-0002-3532-1002

Сирота А. А. – д-р техн. наук, проф., заведующий кафедрой технологий обработки и защиты информации, факультет компьютерных наук, Воронежский государственный университет.
E-mail: *sir@cs.vsu.ru* ORCID iD: https://orcid.org/0000-0002-5785-8513

ГРНТИ: 28.23.37
Поступила в редакцию 7 июля 2023 г. Окончательный вариант – 11 декабря 2023 г.

Improving attention mechanisms in transformer architecture in image restoration

N.I. Berezhnov [1], A.A. Sirota [1]
[1] Voronezh State University, 394018, Russia, Voronezh, Universitetskaya Square 1

Abstract

We discuss a problem of improving the quality of images obtained under the influence of various kinds of noise and distortion. In this work we solve this problem using transformer neural network models, because they have recently shown high efficiency in computer vision tasks. An attention mechanism of transformer models is investigated and problems associated with the implementation of the existing approaches based on this mechanism are identified. We propose a novel modification of the attention mechanism with the aim of reducing the number of neural network parameters, conducting a comparison of the proposed transformer model with the known ones. Several datasets with natural and generated distortions are considered. For training neural networks, the Edge Loss function is used to preserve the sharpness of images in the process of noise elimination. The influence of the degree of compression of channel information in the proposed attention mechanism on the image restoration quality is investigated. PSNR, SSIM, and FID metrics are used to assess the quality of the restored images and for a comparison with the existing neural network architectures for each of the datasets. It is confirmed that the architecture proposed by the present authors is, at least, not inferior to the known approaches in improving the image quality, while requiring less computing resources. The quality of the improved images is shown to slightly decrease for the naked human eye with an increase in the channel information compression ratio within reasonable limits.

Keywords: image quality improvement, neural networks, transformer models, attention mechanism.

Citation: Berezhnov NI, Sirota AA. Improving attention mechanisms in transformer architecture in image restoration. Computer Optics 2024; 48(5): 726-733. DOI: 10.18287/2412-6179-CO-1393.

Authors' information

Nikita I. Berezhnov – 1nd year PhD student, Information Security and Processing Technologies department, Computer Sciences faculty, Voronezh State University.
E-mail: *beregnovnikita@yandex.ru* ORCID iD: https://orcid.org/0000-0002-3532-1002

Alexander A. Sirota – DSc in Technical Sciences, Head of Information Security and Processing Technologies department, Computer Sciences faculty, Voronezh State University.
E-mail: *sir@cs.vsu.ru* ORCID iD: https://orcid.org/0000-0002-5785-8513

Code of State Categories Scientific and Technical Information (in Russian – GRNTI)): 28.23.37
Received July 07, 2023. The final version – December 11, 2023.