Politechnika Warszawska

Wydział Elektryczny

Kierunek Informatyka

Wykorzystanie głębokich sieci neuronowych do poprawy rozdzielczości zdjęć twarzy.

Wykonał:
Aliaksandr Karolik

Promotor:
dr inż. Grzegorz Sarwas

16 maja 2020

Spis treści

1 Cel projektu

Praca skupiała się na badaniu najnowszych rozwiązań algorytmicznych w dziedzinie widzenia komputerowego służących do poprawy rozdzielczości, zwanych również algorytmami super-rozdzielczości (super-resolution). Wybrane metody rokujące swoją użyteczność w przypadku poprawy zdjęć twarzy zostali zaimplementowane i przebadane.

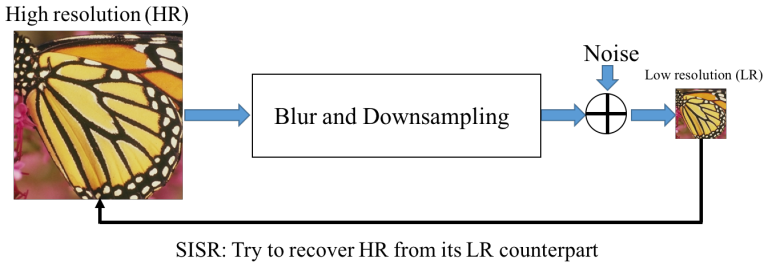
2 Wstęp

Czym jest super-rozdzielczość? Super-rozdzielczość (pisana również jako super resolution, superresolution) jest określeniem zestawu metod zwiększania skali wideo lub obrazów. Terminy takie jak „skalowanie w górę", „powiększanie", „konwersja w górę" i „uprez" również opisują wzrost rozdzielczości w przetwarzaniu obrazu lub edycji wideo. Większość technik super-rozdzielczości opiera się na tym samym pomyśle: wykorzystanie informacji z kilku różnych obrazów do stworzenia jednego powiększonego obrazu. Algorytmy próbują wyodrębnić szczegóły z każdego obrazu w sekwencji, aby zrekonstruować inne ramki.

Głównym celem super-rozdzielczości jest wygenerowanie obrazu o wyższej rozdzielczości z obrazów o niższej rozdzielczości. Obraz w wysokiej rozdzielczości oferuje dużą gęstość pikseli, a tym samym więcej szczegółów na temat oryginalnej sceny. Potrzeba wysokiej rozdzielczości jest powszechna w wizji komputerowej aplikacje dla lepszej wydajności w rozpoznawaniu wzorów i analizie obrazów. Wysoka rozdzielczość ma znaczenie w obrazowaniu medycznym dla diagnozy. Wiele aplikacji wymaga powiększenia określonego obszaru zainteresowania obrazu, w którym niezbędna jest wysoka rozdzielczość, np. aplikacje do nadzoru, kryminalistyki i obrazowania satelitarnego.

3 Wstęp teoretyczny

Super-rozdzielczość (SR) odnosi się do zadania przywracania obrazów o wysokiej rozdzielczości z jednej lub więcej obserwacji tej samej sceny w niskiej rozdzielczości (LR). Zgodnie z liczbą wejściowych obrazów LR, SR można podzielić na super-rozdzielczość pojedynczego obrazu (SISR) i super-rozdzielczość wielu obrazów (MISR). W porównaniu z MISR, SISR jest znacznie bardziej popularny ze względu na wysoką wydajność. Typowa struktura SISR, wygląda następująco:



Rysunek 1: Szkic SISR

Głównie algorytmy SISR dzielą się na trzy kategorie: metody oparte na interpolacji, metody oparte na rekonstrukcji oraz metody oparte na uczeniu. Metody SISR oparte na interpolacji, takie jak interpolacja dwusześcienna (bicubic interpolation) i próbkowanie Lanczosa (Lanczos resampling), są bardzo szybkie i proste, ale dość nie dokładne.

Metody SR oparte na rekonstrukcji, często przyjmują zaawansowaną wcześniejszą wiedzę w celu ograniczenia możliwej przestrzeni rozwiązań z korzyścią polegającą na generowaniu elastycznych i ostrych szczegółów. Jednak wydajność wielu metod opartych na rekonstrukcji szybko spada, gdy zwiększa się skala, oraz metody te są zwykle czasochłonne.

Metody SISR oparte na uczeniu, znane również jako metody oparte na przykładach, najczęściej używane ze względu na ich szybkie obliczenia i wyjątkową wydajność. Metody te zwykle wykorzystują algorytmy uczenia maszynowego do analizy związków statystycznych między LR i odpowiadającym mu odpowiednikiem HR z istotnych przykładów szkoleniowych.

Technika MISR wykorzystuję jako wejście zestaw obrazów niskiej rozdzielczości do budowy obrazu HR, ale jak już wcześniej było wspomniane, SISR jest popularniejsza ze względu na wysoką wydajność.

4 Metryki porównania jakości

Do porównania wyników działania algorytóm zostaną wykorzystane następujące metryki:

- Szczytowy stosunek sygnału do szumu (PSNR, ang. peak signal-to-noise ratio)
- Podobieństwo strukturalne (SSIM, ang. structure similarity)

4.1 Szczytowy stosunek sygnału do szumu

Szczytowy stosunek sygnału do szumu, rzadziej stosunek sygnału szczytowego do szumu (PSNR, ang. peak signal-to-noise ratio) – stosunek maksymalnej mocy sygnału do mocy szumu zakłócającego ten sygnał. Ze względu na szeroki zakres wartości PSNR wyrażany jest w decybelach.

Najczęściej PSNR stosowany jest do oceny jakości kodeków wykorzystujących stratną kompresję obrazów. W takim przypadku sygnałem są nieskompresowane dane źródłowe, a szumem – artefakty (zniekształcenia) spowodowane zastosowaniem kompresji stratnej.

W celu wyznaczenia PSNR należy wpierw obliczyć współczynnik MSE (ang. mean squared error) bazujący na obu porównywanych obrazach, korzystając z wzoru:

$$MSE = \frac{1}{M * N} \sum_{i=1}^{N} \sum_{j=1}^{M} ([f(i,j) - f'(i,j)]^2)$$

Gdzie:

- N, M - wymiary obrazu w pikselach,
- $f(i,j)$- wartość piksela o współrzędnych (i,j) obrzu oryginalnego,
- $f'(i,j)$- wartość piksela o wpółrzędnych (i,j) obrazu skompresowanego.

$$PSNR = 10log_{10}\frac{[max(f(i,j))]^2}{MSE}$$

Gdzie:

- $max(f(i,j))$ – wartość maksymalna danego sygnału; w przypadku obrazów zwykle jest to wartość stała, np. dla obrazów monochromatycznych o reprezentacji 8-bitowej wynosi 255.

5 Podobieństwo strukturalne

SSIM służy do pomiaru podobieństwa pomiędzy dwoma obrazami. Indeks SSIM jest metodą pełnego porównania, innymi słowy, mierzy jakość w oparciu o oryginalny obraz (nie skompresowany lub bez zniekształceń). Wskaźnik SSIM jest rozwinięciem

tradycyjnych metod, takich jak PSNR (peak signal-to-noise ratio) i standardowa metoda błędu MSE, które okazały się niezgodne z fizjologią ludzkiego postrzegania.

Indeks SSIM zawiera się w przedziale od -1 do $+1$. Wartość $+1$ jest osiągana tylko wtedy, gdy próbki są w pełni autentyczne. Indeks SSIM jest obliczany w różnych oknach obrazu. Miara pomiędzy dwoma oknami x i y o wspólnym rozmiarze $N \times N$ wynosi:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Gdzie:

- μ_x wartość średnia z x
- μ_y wartość średnia z y
- σ_x^2 wariancja x
- σ_y^2 wariancja y
- σ_{xy} kowariancja x i y
- $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ dwie zmienne stabilizujące podział ze słabym mianownikiem
- L zakres dynamiki wartości pikseli
- $k_1 = 0.01$ i $k_2 = 0.03$

6 Architektury dla SISR

6.1 SRCNN

SRCNN(Image Super-Resolution Using Deep Convolutional Networks). Architekturę SRCNN pokazana jest na rysunku 2. Jak ustalono w wielu tradycyjnych metodach, dla uproszczenia SRCNN używa tylko komponenty luminancji do treningu. SRCNN jest trójwarstwowym CNN (Konwolucyjne sieci neuronowe), w którym znajdują się rozmiary filtrów każdej warstwy $64 \times 1 \times 9 \times 9$, $1 \times 32 \times 5 \times 5$ i $1 \times 32 \times 5 \times 5$. Dalej w omówieniu działania algorytmu będę stosować następującą notację:

- Y obraz o niskej rozdzielczości
- X prawdiwy obraz o wysokiej rozdzielczości

Każda warstwa odpowiada za następujące czynności:

1. Wyodrębnienie i reprezentacja. Popularną strategią w rekonstrukcji obrazu jest wyodrębnienie plastrów, a następnie reprezentowanie ich przez zestaw wstępnie przeszkolonych baz, takich jak PCA. Jest to równoważne z zawijaniem obrazu przez zestaw filtrów, z których każdy jest podstawą. W naszym sformułowaniu, włączamy optymalizację tych baz do optymalizacji sieci. Formalnie, nasza pierwsza warstwa wyrażona jest jako operacja:

   $$F_1(Y) = max(0, W_1 * Y + B_1)$$

   Gdize:

   - W_1 - odpowiedni filter
   - B_1 - szum
   - $*$ - oznacza operację zwijania (convolution operation)

2. Mapowanie nieliniowe. Pierwsza warstwa wyodrębnia $n1$-wymiarową cechę dla każdego plastra. W drugiej operacji mapujemy każdy z tych $n1$-wymiarowych wektorów na $n2$-wymiarowy. Jest to równoznaczne z zastosowaniem $n2$ filtrów. Działanie drugiej warstwy jest następujące:

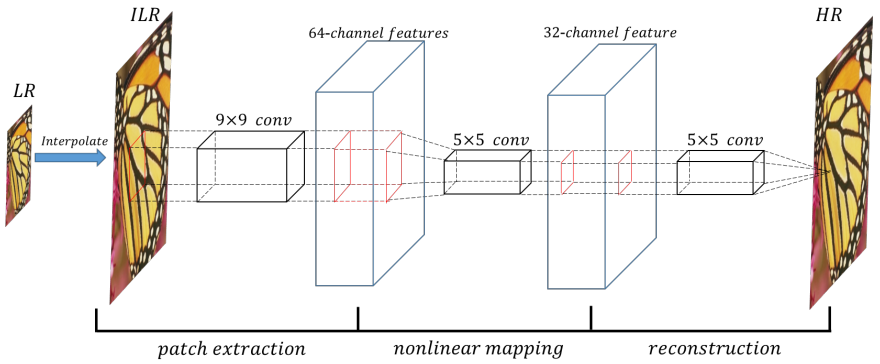
   $$F_2(Y) = max(0, W_2 * F_1(Y) + B_2)$$

   - W_2 - odpowiedni filter
   - B_2 - szum
   - $*$ - oznacza operację zwijania (convolution operation)

   Każdy z wyjściowych $n2$-wymiarowych vektorów jest koncepcyjnie odwzorowaniem plastra o wysokiej rozdzielczości, który zostanie użyty do rekonstrukcji.

3. Rekonstrukcja. Ostatecznie plastry o wysokiej rozdzielczości są uśredniane w celu uzyskania ostatecznego pełnego obrazu. Uśrednianie może być traktowane jako predefiniowany filtr na zestawie map cech (gdzie każda pozycja jest śpłaszczoną"formą wektorową plastry o wysokiej rozdzielczości). Ostatnia warstwa konwolucyjną jest definiowana następująco:

$$F_3(Y) = W_3 * F_2(Y) + B_3$$



Rysunek 2: Architektura SRCNN.

6.1.1 Wyniki działania algorytmu

(a) Oryginalny obraz.

(b) Obraz niskiej rozdzielczości.

(c) Obraz wysokiej rozdzielczości.

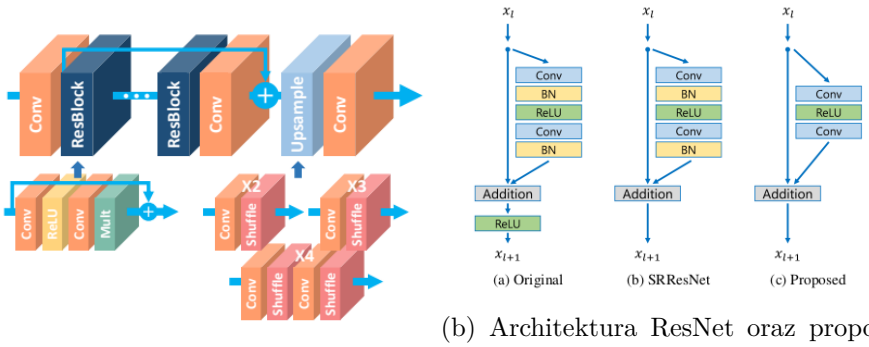
Rysunek 3: Wyniki działania algorytmu SRCNN.

6.2 EDSR

Najprostszym sposobem na zwiększenie wydajności modelu jest zwiększenie liczby parametrów. W sieci konwolucyjnej wydajność modelu może być zwiększona poprzez układanie wielu warstw lub zwiększenie liczby filtrów.

Na Rysunku 4 (a) jest zaprezentowana użyta architektóra sieci w EDSR (Enhanced Deep Residual Networks for Single Image Super-Resolution). Niebieskie bloki na obrazie reprezentują bloki w stylu ResNet (ang. Residual Neural Network). Zostali one użyte ze względu na to że super-rozdzielczość wymaga, aby większość informacji zawartych w obrazie LR została zachowana w obrazie HR. W związku z tym modele używające takie bloki uczą się głównie pozostałości między obrazami LR i HR.

Autorzy stwierdzili, że zwiększenie liczby map cech powyżej pewnego poziomu spowodowałoby, że procedura szkoleniowa byłaby niestabilna pod względem obliczeniowym. Problem został rozwiązany w następujący sposób w każdym resztkowym bloku po ostatniej warstwie konwolucyjnej umieszczane są stałe warstwy skalowania. Moduły te znacznie stabilizują procedurę treningową przy użyciu dużej liczby filtrów.



(a) Architektura EDSR.

(b) Architektura ResNet oraz proponowany wariant.

Rysunek 4: Architektura EDSR.

6.2.1 Wyniki działania algorytmu

Poniższe wyniki powstali z wytrenowanej sieci w architektórze 16 bloków resztkowych i 64 filtrami w warstwie konwolucyjnej filtry miały wymiar 3×3 jak jest omówione w artykule związanym z tą architektórą.

(a) Oryginalny obraz.

(b) Obraz niskiej rozdzielczości.

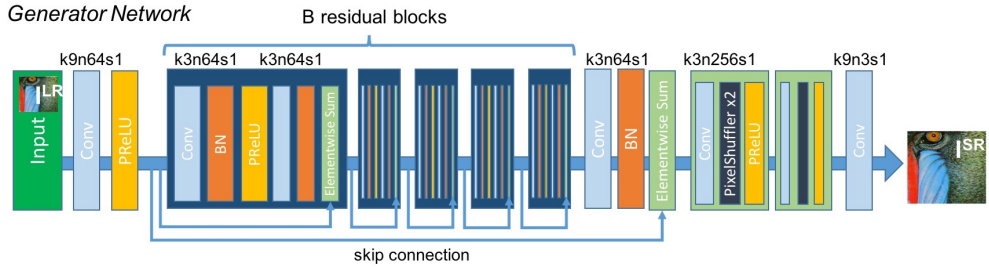
(c) Obraz wysokiej rozdzielczości.

Rysunek 5: Wyniki działania algorytmu EDSR.

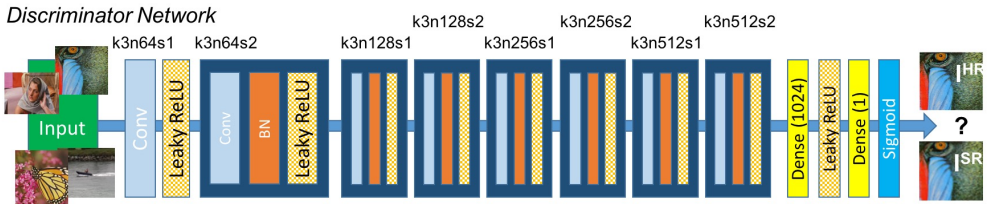
6.3 SRGAN

SRGAN - Photo-Realistic Single Image Super-Resolution Using a Generative AdversarialNetwork. Wybrana architektura wykorzustuje wewnątrz generatywną sieć przeciwstawną. To tak naprawdę dwie sieci neuronowe: pierwsza (generator) próbuje stworzyć obraz nieodróżnialny od prawdziwego z bazy wzorcowych danych, podczas gdy druga (dyskryminator) stara się znaleźć różnice. Ogólna idea takej architektury jest taka, że pozwala ona na szkolenie generatora w celu oszukiwania różnicującego dyskryminatora, który jest szkolony do odróżniania superrozdzielczych obrazów od rzeczywistych. Dzięki takiemu podejściu generator może nauczyć się tworzyć rozwiązania, które są bardzo podobne do prawdziwych obrazów, a przez to trudne do sklasyfikowania przez dyskryminatora. Uczenie takiego rodzaju zachęca do poszukiwania lepszych rozwiązań w podprzestrzeni które istnieją w przestrzeni naturalnych obrazów. SRGAN jest bardziej atrakcyjny dla człowieka bo może wykreować obraz z większą ilością szczegółów.

Struktura sieci jest zaprezentowana na Rysunku 6.



(a) Architektura generatora.



(b) Architektura dyskryminatora.

Rysunek 6: Architektura SRGAN.

6.3.1 Wyniki działania algorytmu

Poniższe wyniki powstały z wytrenowanego generatora mającego 16 bloków resztkowych z 64 filtrami w warstwie konwolucyjnej filtry miały wymiar 3×3 oraz deskryminatora który miał 8 bloków zawierających warsty konwolucyjnej każdy następny blok miał w dwa razy więcej filtrów czyli z 64 do 512.

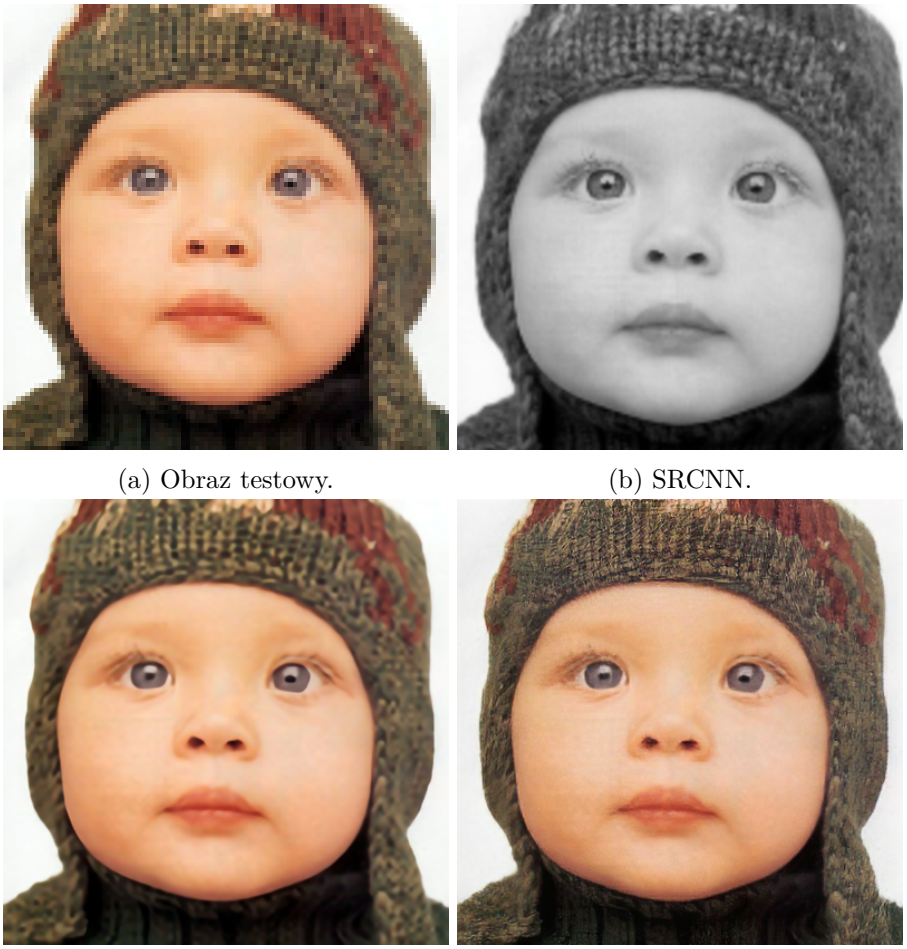
(a) Oryginalny obraz.

(b) Obraz niskiej rozdzielczości.

(c) Obraz wysokiej rozdzielczości.

Rysunek 7: Wyniki działania algorytmu SRGAN.

7 Wnisoki



(a) Obraz testowy. (b) SRCNN.

(c) EDSR. (d) SRGAN.

Rysunek 8: Wyniki działania algorytmu SRGAN.

Najlepsze wyniki osiągnął algorytm SRGAN w porównaniu z wybranymi algorytmami. Jak już było pisane wcześniej SRGAN jest bardziej atrakcyjny dla człowieka z większą ilością szczegółów wyniki eksperymentów to potwierdziły.

Literatura

[1] SRCNN Image Super-Resolution Using Deep Convolutional Networks https://arxiv.org/abs/1501.00092

[2] EDSR Enhanced Deep Residual Networks for Single Image Super-Resolution https://arxiv.org/abs/1707.02921

[3] SRGAN Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network https://arxiv.org/abs/1609.04802