Politechnika Wrocławska

Wydział Elektroniki, Fotoniki i Mikrosystemów

KIERUNEK: Automatyka i Robotyka (AIR)

PRACA DYPLOMOWA INŻYNIERSKA

TYTUŁ PRACY:

Aplikacja webowa zwiększająca rozdzielczość obrazów

AUTOR:

Eryk Wójcik

PROMOTOR:

dr hab. inż. Andrzej Rusiecki,
Katedra Informatyki Technicznej

WROCŁAW 2023

Spis treści

Rozdział 1

Wstęp

Cel pracy

Opis celu badań, czyli stworzenia aplikacji webowej służącej do zwiększania rozdzielczości obrazów z użyciem algorytmów ESRGAN i DWSR oraz analiza i porównanie tych algorytmów.

Zakres pracy

Przedstawienie koncepcji i zagadnień, które zostaną omówione w pracy, w tym wybrane metody i technologie.

Rozdział 2

Podstawy teoretyczne

Celem rozdziału jest przedstawienie podstawowych definicji, wytłumaczenie aparatu matematycznego oraz metod wykorzystywanych w algorytmach na których skupia się praca. Dodatkowo ma on na celu ułatwienie dalszego czytania poprzez zapoznanie czytelnika z przyjętymi konwencjami, oznaczeniami oraz symbolami, które mogą pojawić się w kolejnych rozdziałach.

2.1 Definicja super-rozdzielczości

Super-rozdzielczość (ang. Super-Resolution) odnosi się do procesu poprawy rozdzielczości obrazu lub sekwencji obrazów. W kontekście cyfrowym, super-rozdzielczość jest często realizowana za pomocą algorytmów komputerowych, które mają na celu odtworzenie wysokiej rozdzielczości obrazu [Rys 2, 3] z jednego lub wielu obrazów o niskiej rozdzielczość [Rys 1].

Rys 1. Obraz oryginalny

Rys 2. Obraz powiększony czterokrotnie

Rys 3. Obraz powiększony szesnastokrotnie

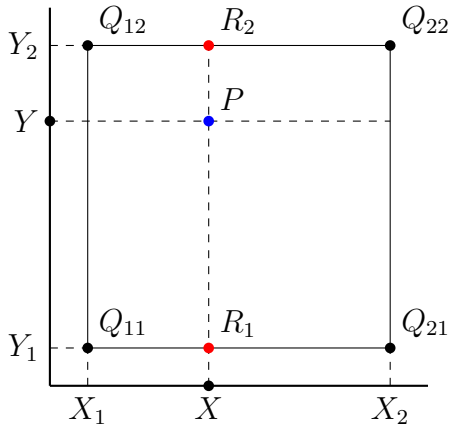
2.2 Przegląd metod powiększania obrazów

Istnieje wiele metod powiększania rozdzielczości obrazów. Najprostszą z nich jest **interpolacja najbliższego sąsiada**, która polega na powieleniu pobliskich pikseli w celu zwiększenia rozdzielczości obrazu.
Metoda ta jest bardzo prosta w implementacji, jednakże nie daje ona zadowalających rezultatów. Obraz powiększony w ten sposób wygląda jak obraz o niskiej rozdzielczości z większymi pikselami [Rys 5].

Rys 4. Obraz oryginalny

Rys 5. Obraz powiększony metodą najbliższego sąsiada

Aby poprawić jakość obrazu, można zastosować **interpolację dwuliniową**. Metoda ta rozszerza interpolację liniową na interpolację funkcji dwóch zmiennych [Rys 6].
W efekcie polega to na wyznaczeniu średniej ważonej pikseli sąsiadujących z pikselem, który chcemy powielić. Współczynniki wag są wyznaczane na podstawie odległości od piksela, który chcemy powielić.



Rys 6. Wizualizacja interpolacji dwuliniowej

Kroki algorytmu:

1. Przeprowadzana jest interpolacja liniowa wzdłuż osi OX:

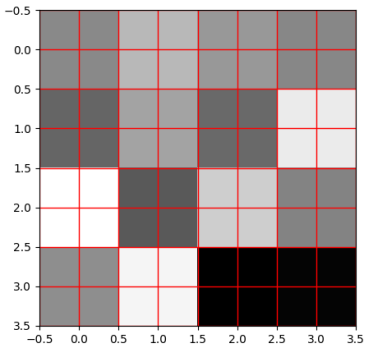
$$f\left(R_1\right) \approx \frac{x_2-x}{x_2-x_1} f\left(Q_{11}\right) + \frac{x-x_1}{x_2-x_1} f\left(Q_{21}\right) \quad \text{gdzie} \quad R_1 = \left(x, y_1\right),$$

$$f\left(R_2\right) \approx \frac{x_2-x}{x_2-x_1} f\left(Q_{12}\right) + \frac{x-x_1}{x_2-x_1} f\left(Q_{22}\right) \quad \text{gdzie} \quad R_2 = \left(x, y_2\right).$$

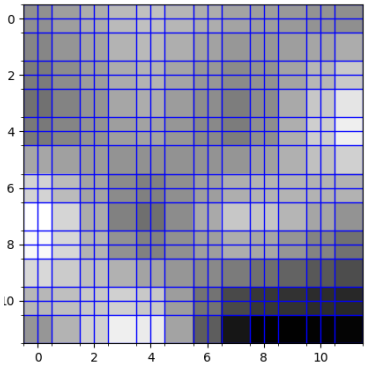
2. Następnie przeprowadzana jest interpolacja wzdłuż osi OY:

$$f(P) \approx \frac{y_2-y}{y_2-y_1} f\left(R_1\right) + \frac{y-y_1}{y_2-y_1} f\left(R_2\right).$$

W efekcie otrzymujemy obraz wyglądający następująco [Rys 8, 10].



Rys 7. Obraz wejściowy



Rys 8. Interpolacja dwuliniowa

Rys 9. Obraz wejściowy

Rys 10. Obraz powiększony metodą interpolacji dwuliniowej

Metoda te daje lepsze rezultaty niż interpolacja najbliższego sąsiada, jednakże wprowadziła ona duże rozmycie, które jest szczególnie widoczne na krawędziach i obszarach wysokiej częstotliwości.

W teorii informacji istnieje koncepcja zwana **nierównością przetwarzania danych**. Zgodnie z nią niezależnie od sposobu przetwarzania danych, **nie można dodać informacji, której nie ma w oryginale.** Oznacza to, że brakujących danych nie można odzyskać poprzez dalsze przetwarzanie. Czy to oznacza, że super-rozdzielczość jest teoretycznie niemożliwa?
Nie, jeśli mamy dodatkowe źródło informacji.

2.3 Wprowadzenie do głębokiego uczenia się w przetwarzaniu obrazów

Głębokie uczenie rewolucjonizuje przetwarzanie obrazów, wprowadzając modele zdolne do uczenia się cech z serii danych. W przetwarzaniu obrazów, głębokie sieci neuronowe są wykorzystywane do zadań takich jak detekcja obiektów, segmentacja, klasyfikacja obrazów, czy właśnie super-rozdzielczość.
Sieć neuronowa może nauczyć się odtwarzać szczegóły obrazów na podstawie pewnych informacji, które zbiera z dużego zbioru obrazów.
Szczegóły dodawane do powiększanego obrazu w przy użyciu modelu uczenia maszynowego nie naruszają nierówności przetwarzania danych, ponieważ wykorzystywane informacje są w zbiorze treningowym, nawet jeśli nie ma ich na obrazie wejściowym.

Rys 11. Obraz wejściowy

Rys 12. Obraz powiększony algorytmem DWSR

Rys 13. Obraz powiększony algorytmem ESRGAN

2.4 Wstęp do funkcji falkowych

Omówienie funkcji falkowych, opis do czego to narzędzie służy począwszy od transformaty Fouriera i jej ograniczeń, w jaki sposób funkcje falkowe rozwijają FFT, przedstawienie działania.

Chatbot:

Przegląd metod powiększania obrazów

Tradycyjne metody powiększania obrazów obejmują interpolację najbliższego sąsiada, interpolację dwuliniową oraz interpolację dwusześcienną. Te techniki opierają się na prostych przekształceniach matematycznych i mogą być łatwo implementowane, lecz często prowadzą do efektów ubocznych takich jak rozmycie czy artefakty.

Współczesne techniki wykorzystują bardziej zaawansowane podejścia, takie jak rekonstrukcja oparta na przykładowych obrazach (example-based reconstruction), gdzie wykorzystuje się zbiór danych zawierający pary obrazów LR i HR, aby nauczyć algorytm odtwarzania szczegółów na wysokim poziomie rozdzielczości. Metody te mogą wykorzystywać różnorodne techniki, w tym mapowanie cech, modelowanie statystyczne, a nawet uczenie maszynowe.

W ostatniej dekadzie, z uwagi na rozwój w dziedzinie uczenia maszynowego, pojawiły się metody bazujące na głębokich sieciach neuronowych, takie jak konwolucyjne sieci neuronowe (CNN, z ang. Convolutional Neural Networks) w zadaniach SR. Przykładami takich metod są SRCNN (Super-Resolution Convolutional Neural Network), VDSR (Very Deep Super-Resolution), EDSR (Enhanced Deep Super-Resolution) oraz GAN (Generative Adversarial Networks) wykorzystywane w modelu SRGAN (Super-Resolution Generative Adversarial Network), które przynoszą znaczące ulepszenia w jakości rekonstruowanych obrazów HR.

Wprowadzenie do głębokiego uczenia się w przetwarzaniu obrazów

Podstawową jednostką w konwolucyjnych sieciach neuronowych jest warstwa konwolucyjna, która skanuje obraz za pomocą filtrów (również nazywanych jądrami), wyodrębniając lokalne cechy. Następnie, warstwy pooling redukują wymiarowość danych, zachowując przy tym najważniejsze informacje. Głębokie sieci neuronowe mogą składać się z wielu takich warstw, co pozwala na wyodrębnianie coraz bardziej abstrakcyjnych cech z obrazów.

Sieci te są trenowane przy użyciu zestawów danych, gdzie algorytm uczy się mapowania między danymi wejściowymi a oczekiwanymi wynikami. W zadaniach SR, sieć jest trenowana do przewidywania obrazów HR z obrazów LR, przy czym jakość wyniku jest oceniana przy pomocy funkcji straty, takiej jak błąd średniokwadratowy (MSE, z ang. Mean Squared Error) czy wskaźnik strukturalnej jakości obrazu (SSIM, z ang. Structural Similarity Index).

Nowoczesne podejścia do SR często wykorzystują także sieci GAN, które składają się z dwóch części: generatora, który generuje obrazy, oraz dyskryminatora, który ocenia, czy obrazy są wystarczająco zbliżone do prawdziwych HR. Konkurencyjna dynamika między tymi dwoma sieciami prowadzi do generowania coraz lepszych obrazów.

W dalszych rozdziałach szczegółowo omówimy te techniki, przedstawiając zarówno ich teoretyczne podstawy, jak i praktyczne zastosowania. Zostaną przedstawione również aktualne wyzwania i kierunki badań w dziedzinie super-rozdzielczości, co pozwoli czytelnikowi na głębsze zrozumienie tego dynamicznie rozwijającego się obszaru wiedzy.

Youtube - super resolution

So, how can we train such a model? If you watched my Deep Learning Crash Course series, you might be thinking: can't we just train a neural network to learn a mapping between low and high-resolution images? Yes, we can, and we wouldn't be the first ones to do so. Dataset That's pretty much what the SRCNN paper did. First, we can create a dataset by collecting high-resolution images and downscaling them, or we can simply use one of the existing super-resolution datasets, such as the DIV2K dataset. Then, we can build a convolutional neural network that would input only the low-resolution images, and we can train it to produce higher resolution images that match the original ones the best. The SRCNN paper simply minimized the squared difference between the pixel values to produce images that are as close as possible to the original high-resolution images. Mean squared error But is mean squared error really the right metric to optimize? This is a very old debate. Long story short, mean squared error doesn't express the human perception of image fidelity well. For example, all of these distorted images are equally distant from the original image in terms of mean squared error. Clearly, they don't look equally good. Because mean squared error cares only about pixel-wise intensity differences but not the structural information about the contents of an image. There's a better measure of perceptual image quality called the structural similarity index, which was developed in my lab at the University of Texas at Austin. The structural similarity index made a very high impact, both in academia and the industry. My doctoral advisor, Alan Bovik, and his collaborators won a Primetime Emmy Award for this method a few years ago. This metric was initially developed to measure the severity of image degradations. However, many researchers also used it as a loss function to train neural networks for Perceptual loss image restoration. More recently, people also started using pre-trained convolutional neural networks as perceptually-relevant loss functions. How it works is that you first take a pre-trained model. This is typically a VGG-19 model trained on ImageNet. Then take it's first few layers and compute the difference between the feature maps produced by those layers. The difference between the feature maps can be minimized to train another model, just like any other loss function. The layers that generate those feature maps stay frozen during training and act as a fixed feature extractor. This method is commonly referred to as perceptual loss, content loss, or VGG-loss. How is this relevant to super-resolution? We can use this loss function to train models to enhance images and get pretty decent results. Super Resolution But, sometimes, it doesn't feel fair to penalize the model for pixelwise differences that don't really make much difference for human viewers. For example, does the direction of the hair on this baboon's face really matter? What if we cared a little less about how the original high-resolution images looked like, as long as the produced images looked good. We can do so by using GANs: generative adversarial networks. GANs consist of two networks fighting each other to achieve adversarial goals. I made a more detailed video about this earlier. There's a GAN-based super-resolution system called SRGAN. It uses a generator network that inputs low-resolution images and tries to produce their high-resolution versions. It also uses a discriminator network that tries to tell whether this is a real high-resolution image or an image upscaled by the generator. Both networks are trained simultaneously, and they both get better over time. Once the training is done, all we need is the generator part to upscale low-resolution images. In addition to this adversarial training setup, SRGAN also used a VGG-based loss function that we talked about earlier. There's another paper called Enhanced SRGAN, which proposed a few tricks to improve the ESR Gain results further. Enhanced SRGAN, or ESRGAN for short, somehow got popular in the gaming community. People started

using it to upscale vintage games, and it worked pretty well. It's surprising how well it worked on video game graphics despite being trained only on natural images. Let's take a look at what enhancements the ESRGAN paper proposed for better results. First, they removed the batch normalization layers in their network architecture. This may sound contradictory to what I said in my previous videos, but it's not. Batch normalization does help a lot for many computer vision tasks. But for image-processing related tasks, such as super-resolution or image restoration in general, batch normalization can create some artifacts. They also added more layers and connections to their model architecture. It's not surprising that a more sophisticated model resulted in better images, but deeper models can be trickier to train, especially if they are not using batch normalization layers. So, the authors of ESRGAN used some tricks like residual scaling to stabilize the training of such a network. In addition to the changes in the model architecture, they also modified the loss functions. For example, they modified the VGG-loss in a way that compared the feature maps before activations. Their rationale is that the feature maps are denser and contain more information before they get clipped by the activation functions. In the original SRGAN paper, the discriminator model was trained to detect whether its input is real or fake. In the enhanced version, the authors used a relativistic discriminator that tells whether the input looks more realistic than fake data or less realistic than real data. Earlier I said minimizing the mean squared error might not be the best way to generate textures that look appealing to the human visual system. Then, I went on to say maybe we shouldn't care too much about how close the generated images are to the original ones. There's actually a trade-off there. Interpolation We would still want the upscaled images to be a faithful representation of the originals while having good-looking textures. The ESRGAN paper aims to find the sweet spot by interpolating between models. What they do is that they compute the weighted average of two models, one trained using mean squared error, and the other fine-tuned with adversarial training. Blending the parameters this way allows for finding the right balance between the two models without retraining them. Zoom to Learn More recently, another paper also explored the idea of network interpolation, and their results also look promising. Super-resolution is a relatively hot topic, and many researchers are experimenting with different ways of approaching this problem and are publishing their results. This paper, titled "Zoom to learn, Learn to zoom," for example, focuses on building a model that mimics optical zoom directly on raw sensor data. The authors created a dataset of raw images, and their corresponding optically zoomed ground truth. Super Resume They also proposed a loss function named "contextual bilateral loss" to handle slightly misaligned image pairs. Speaking of raw images, Google Pixel's Super Res Zoom feature showed that it's possible to achieve super-resolution through a burst of raw images. Google's method makes use of slight hand movements to fill in the missing spots in an upscaled image. So what if the user is using a tripod, and the image is perfectly still. Then, they deliberately jiggle the camera between the shots. So, to be able to implement this, you need to have complete control of the hardware. Single Frame Super Resolution Unlike the other methods we covered so far, Google's Super Res Zoom is a multi-frame super-resolution algorithm. If you don't have such bursts of images and want to upscale your pictures, you can easily use the single-frame super-resolution methods that we overviewed today. ESRGAN, for example, operates on a single input image and is very easy to run on an arbitrary picture you may want to use. Face Upscaling Super Resolution There are also task-specific super-resolution models, which I think is worth mentioning. For instance, face-upscaling models use face priors to synthesize realistic details on faces. Basically, the models know what a face typically looks like and uses that information to hallucinate the details. As you can tell, those methods

are absolutely not suitable for CSI purposes, since all the details in the upscaled version are completely made up. Alright, that's all for today. I hope you liked it. I put the links to all referenced papers in the description below. Subscribe for more videos. And as always, thanks for watching, stay tuned, and see you next time.

Rozdział 3

DWSR: Deep Wavelet Super Resolution

[1]

3.1 Architektura DWSR

Dokładne przedstawienie struktury i funkcjonowania sieci DWSR, podkreślając jej unikalne cechy i mechanizmy.

3.2 Kluczowe cechy i innowacje

Dyskusja na temat głównych innowacyjnych rozwiązań zastosowanych w DWSR i ich wpływu na efektywność metody.

3.3 Proces treningu i implementacji

Wyjaśnienie procedur związanych z treningiem DWSR, z uwzględnieniem specyfikacji danych, procesu uczenia i kwestii implementacji.

3.4 Przykłady zastosowań i rezultaty

Ilustracja praktycznych zastosowań DWSR oraz ocena i interpretacja osiągniętych dzięki niemu wyników.

Rozdział 4

ESRGAN

[2]

4.1 Architektura ESRGAN

Szczegółowy opis architektury sieci ESRGAN, w tym jej głównych komponentów i zasady działania.

4.2 Kluczowe cechy i innowacje

Omówienie innowacji wprowadzonych w ESRGAN i w jaki sposób różnią się one od wcześniejszych podejść.

4.3 Proces treningu i implementacji

Opis procesu treningu sieci ESRGAN, w tym zbierania danych, uczenia oraz wyzwań implementacyjnych.

4.4 Przykłady zastosowań i rezultaty

Prezentacja przykładów, gdzie ESRGAN został użyty oraz analiza wyników, jakie osiągnięto dzięki tej technologii.

Rozdział 5

Porównanie algorytmów ESRGAN i DWSR

5.1 Kryteria porównawcze

Ustalenie kryteriów, które będą stosowane do oceny i porównania skuteczności i efektywności algorytmów super rozdzielczości.

5.2 Analiza wydajności

Bezpośrednie porównanie wydajności obu metod w różnych warunkach, bazujące na ustalonych kryteriach.

5.3 Jakość odtwarzania obrazów

Ocena jakości obrazów generowanych przez oba algorytmy, uwzględniając różne aspekty jakości wizualnej.

5.4 Ograniczenia i wyzwania

Dyskusja na temat ograniczeń obu metod i potencjalnych wyzwań w ich stosowaniu.

Rozdział 6

Aplikacja webowa do powiększania rozdzielczości obrazów

6.1 Projektowanie aplikacji

Wytłumaczenie wyboru określonych technologii i narzędzi użytych do stworzenia aplikacji webowej.
Projekt interfejsu użytkownika
Omówienie procesu projektowania interfejsu użytkownika, w tym wytycznych ergonomii i użyteczności.

6.2 Wybór narzędzi i technologii

6.3 Implementacja aplikacji

Opis technicznego procesu integracji wybranych algorytmów z aplikacją, wraz z napotkanymi wyzwaniami.

6.4 Integracja algorytmów DWSR i ESRGAN

6.5 Wdrożenie i utrzymanie aplikacji

Omówienie procesu wdrożenia gotowej aplikacji oraz planów dotyczących jej przyszłego utrzymania i aktualizacji.

Rozdział 7

Podsumowanie i wnioski

7.1 Dyskusja wyników

Krytyczna analiza uzyskanych wyników w kontekście celów pracy oraz istniejących badań i literatury w dziedzinie.

7.2 Rekomendacje i kierunki dalszych badań

Sugestie dotyczące potencjalnych ulepszeń i obszarów, które wymagają dalszych badań, w oparciu o obserwacje i wyniki badań.

Literatura

[1] T. Guo, H. S. Mousavi, T. H. Vu, V. Monga. Deep wavelet prediction for image super-resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. C. Loy. Esrgan: Enhanced super-resolution generative adversarial networks. *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.