Sztuczna inteligencja, zastosowanie, możliwości i ograniczenia, fakty i mity

Krzysztof Kolasiński

HUAWEI Mate20 Pro

CO-ENGINEERED WITH Leica

PODWOJONA MOC AI

Mobilny procesor 7nm ze sztuczną inteligencją | Potrójny aparat z ultra szerokokątnym obiektywem | Pojemna bateria 4200 mAh z super szybkim ładowaniem

Sztuczna inteligencja, zastosowanie, możliwości i ograniczenia, fakty i mity

- Podstawy oraz nazewnictwo, zakres prezentacji
- Przykładowe zastosowania
- Problemy AI

Sztuczna inteligencja czyli uczenie maszynowe

Co obecnie rozumiemy przez AI ?

- **Sztuczna Inteligencja** (ang artificial intelligence AI) - to w praktyce uczenie maszynowe (machine learning)
- w praktyce (ta prezentacja) termin ten będzie się odnosić do tzw. **głębokiego uczenia** (deep learning)
- za początki rewolucji głębokiego uczenia możemy uznać rok 2012 - AlexNet
- boom AI związany jest z
  - pojawieniem się wydajnych procesorów graficznych **GPU,**
  - łatwym dostępem do **danych**
  - oraz odrodzeniem **konwolucyjnych sieci neuronowych**

Czym jest uczenie maszynowe rozumiane jako AI?

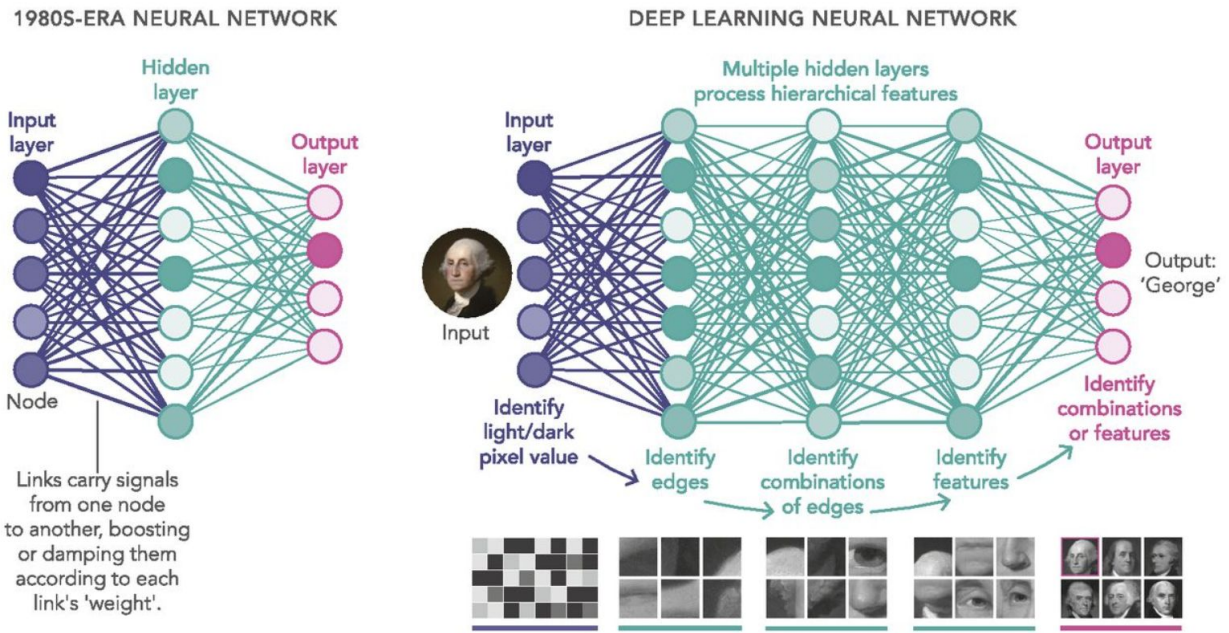
- To jest obszerna dziedzina zajmująca się automatyzacją pozyskiwania danych, ich analizy i interpretacji, generacji nowych danych
- chęć wyeliminowania z wielu procesów czynnika ludzkiego, stąd termin sztuczna inteligencja
- Sieci neuronowe to funkcje których zadaniem jest odwzorowanie wejścia na żądane przez nas wyjście **i są w tym bardzo dobre**

Czym jest głębokie uczenie ?

- Na obecną chwilę przez głębokie uczenie rozumiemy sieci neuronowe z dużą ilością warstw i parametrów
- W większości przypadków sieci te będa stosować najnowsze techniki trenowania sieci neuronowych

Czym jest głębokie uczenie ?

- **Głębokie uczenie** to w dużej mierze uczenie maszynowe na sterydach - więcej danych, większe modele, więcej tricków, więcej wszystkiego

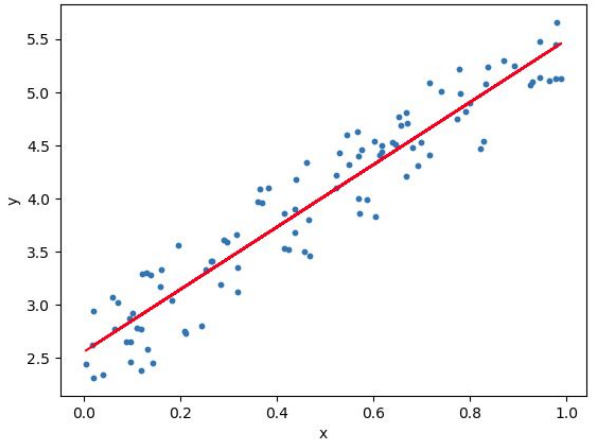


Na czym polega uczenie sieci neuronowej ?

- Uczenie sieci neuronowej nie różni się wiele od dopasowania punktów do prostej

$$y = ax + b$$

- Potrzebujemy danych **x** oraz wyniku obserwacji **y**



Na czym polega uczenie sieci neuronowej ?

- Proces uczenia polega na znalezieniu odpowiednich współczynników naszego modelu (a, b)

$$y = ax + b$$

- Które będą minimalizować zadaną funkcję (koszt) np:

$$\mathcal{L} = \sum_{i=0}^{N} (ax_i + b - y_i)^2$$

- Regresja liniowa ma rozwiązanie analityczne więc proces uczenia odbywa się w jednym kroku
- W sieciach neuronowych proces uczenia odbywa się w sposób iteracyjny: **metoda gradientu prostego**

$$a^{(n+1)} := a^{(n)} - \lambda \frac{\partial \mathcal{L}}{\partial a^{(n)}}$$

Na czym polega uczenie sieci neuronowej ?

- Proces uczenia polega na znalezieniu odpowiednich współczynników naszego modelu (a, b)

$$y = ax + b$$

- W ogólności sieci nie operują na skalarach ale na wektorach

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

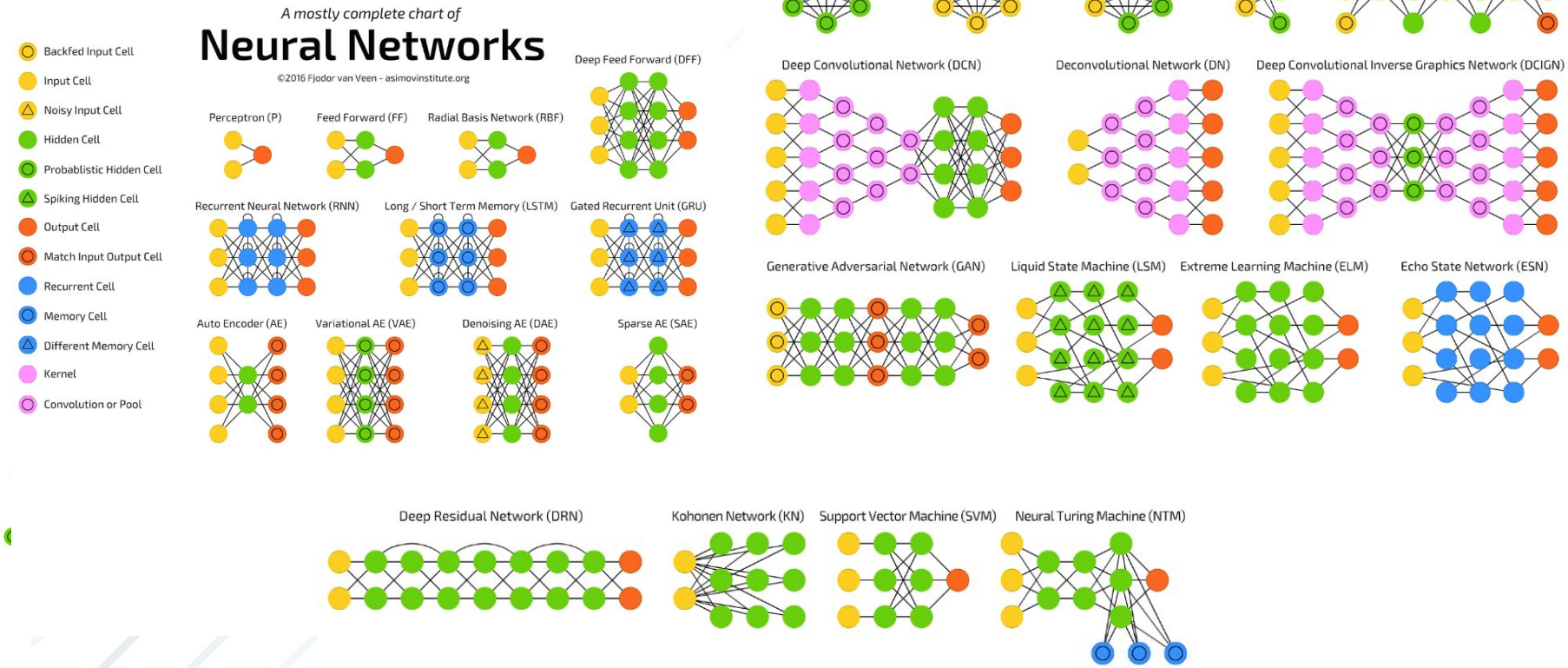
- Sieci neuronowe składamy z powyższych operacji łącząc je na przemian nieliniowymi funkcjami

$$\mathbf{h}_1 = \tanh\left(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1\right)$$
$$\mathbf{h}_2 = \tanh\left(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2\right)$$

A.I.

- Taka sieć nie posiada analitycznego rozwiązania
- Co to znaczy optymalne rozwiązanie ?

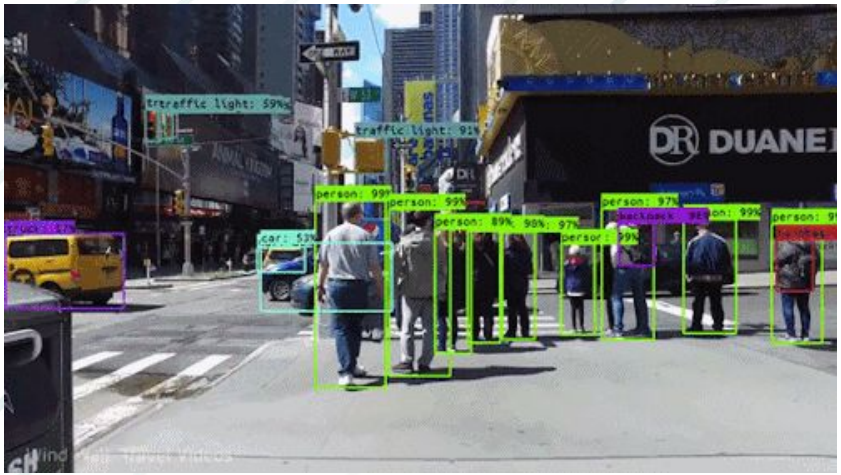
Neuronowe zoo



Przykłady zastosowania głębokiego uczenia

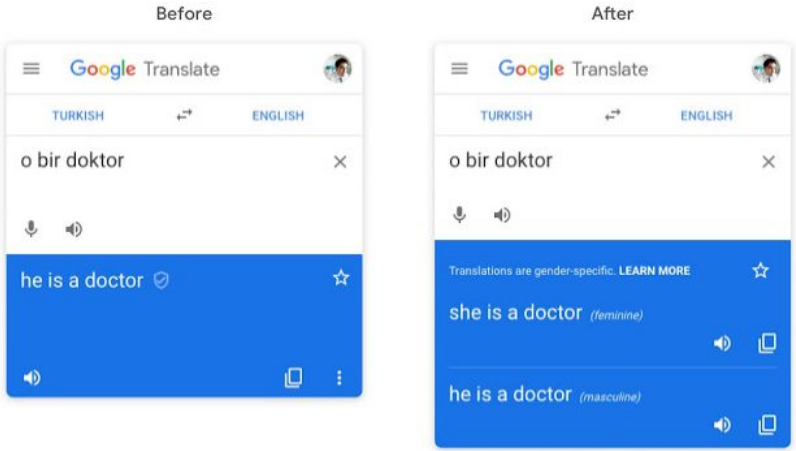
Przykłady zastosowania głębokiego uczenia

- detekcja oraz segmentacja



Przykłady zastosowania głębokiego uczenia

- tłumaczenie tekstu



Gender-specific translations on the Google Translate website.

Przykłady zastosowania głębokiego uczenia

- Synteza mowy z aktywności neuronów - Nature kwiecień - 2019

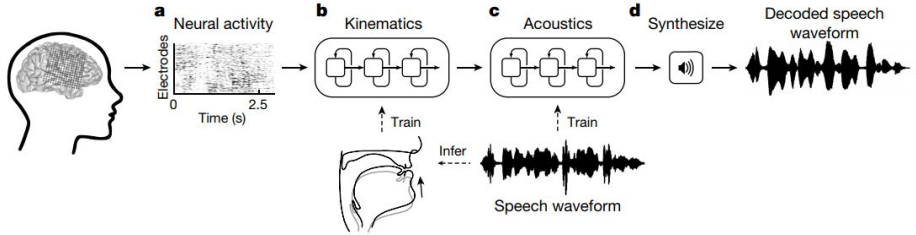
nature
International journal of science

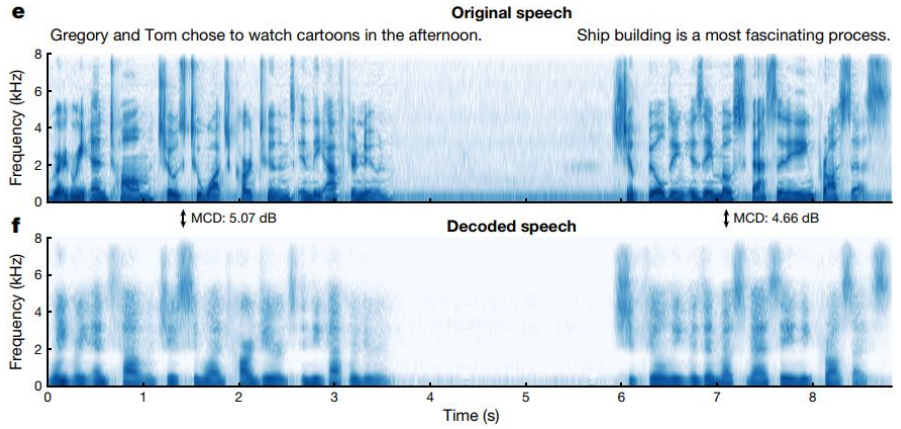
Article | Published: 24 April 2019

Speech synthesis from neural decoding of spoken sentences

Gopala K. Anumanchipalli, Josh Chartier & Edward F. Chang

Nature **568**, 493–498 (2019) | Download Citation



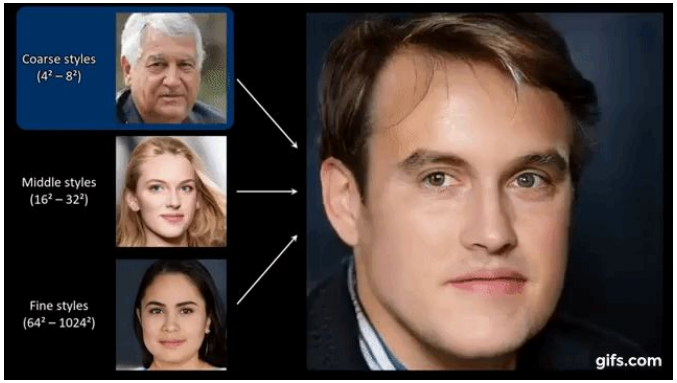


Przykłady zastosowania głębokiego uczenia

- Generatywne sieci współzawodniczące (GANy) - 2014

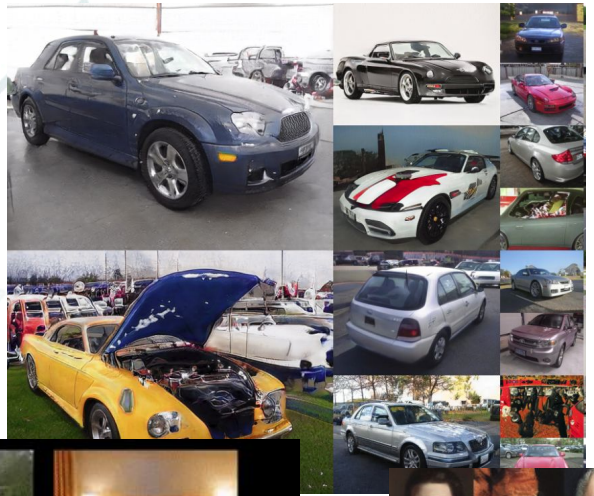
Przykłady zastosowania głębokiego uczenia

- Generatywne sieci współzawodniczące (GANy) - tutaj StyleGAN



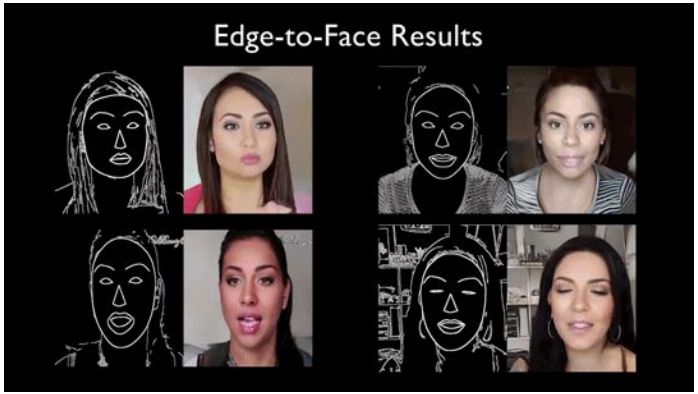
Przykłady zastosowania głębokiego uczenia

- Generatywne sieci współzawodniczące (GANy) - tutaj StyleGAN



Przykłady zastosowania głębokiego uczenia

- Transfer domen w filmach video (bazuje na GANach)



Przykłady zastosowania głębokiego uczenia

- Deepfake - podmiana mimiki twarzy u dowolnej osoby

Problemy sieci neuronowych

Problemy

- Większośc problemów będzie również występować w innych dziedzinach - nie tylko wizja
- Nie zamierzam dyskutować problemów technicznych:
  - wymagana wielkość zbioru uczącego
  - potrzebne zasoby obliczeniowe
  - czas trenowania
  - itp
- to są rzeczy które niekoniecznie muszą być problemem np. dla Google
- nie będę poruszał też podejść starających się rozwiązać dany problem
  - z reguły nie są stosowane w praktyce
  - pokryje to 99% przypadków

Problem #1 Ataki przeciwstawne

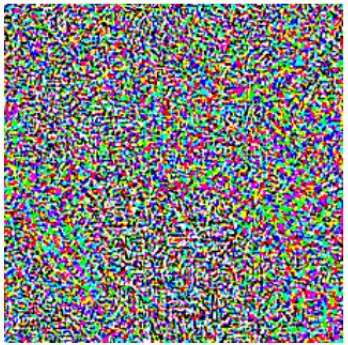
P#1 Ataki przeciwstawne

- Najbardziej znany problem, do tej pory nierozwiązany
- Małe zaburzenia wektora wejściowego mogą wpływać na duże zmiany w wyjściu
- Jest to klasyczny przykład stabilności funkcji

$\boldsymbol{x}$

"panda"
57.7% confidence

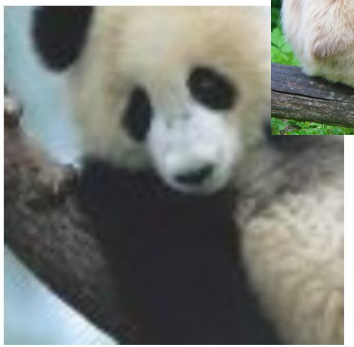
$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$



$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

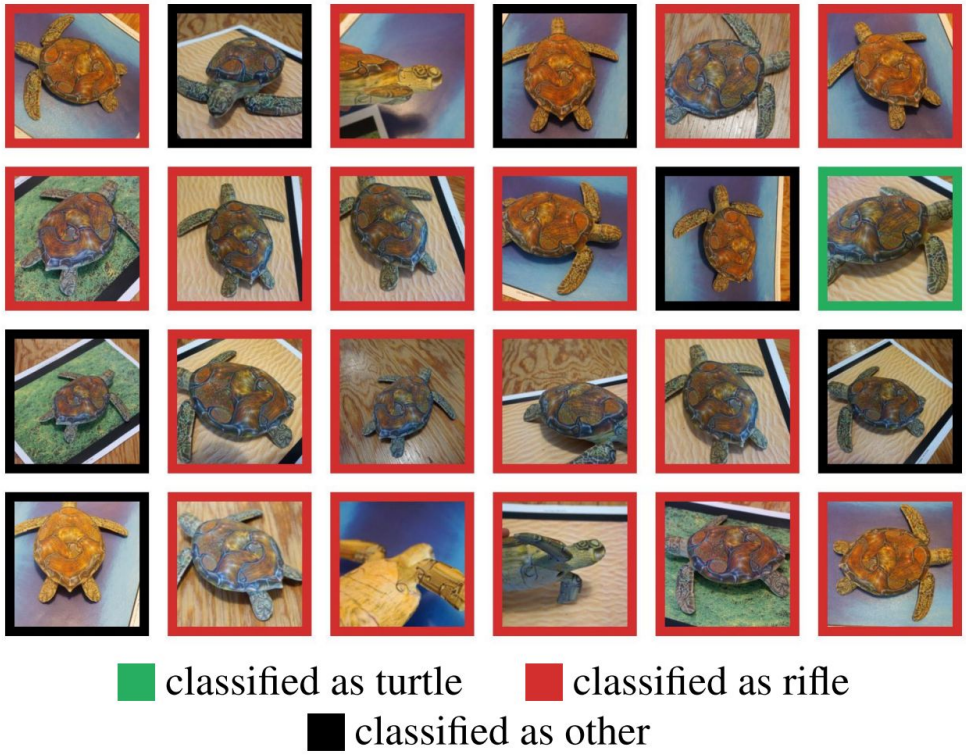
P#1 Ataki przeciwstawne

- Ataki mogą przybierać różną postać
- Wydrukowane znaki drogowe, oraz nalepki na znaki (celem było ograniczenie prędkości 45 mil/h)

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

P#1 Ataki przeciwstawne

- Wydrukowane obiekty 3D ...



P#1 Ataki przeciwstawne

- Pojedynczy pixel ...

Cup(16.48%)
Soup Bowl(16.74%)

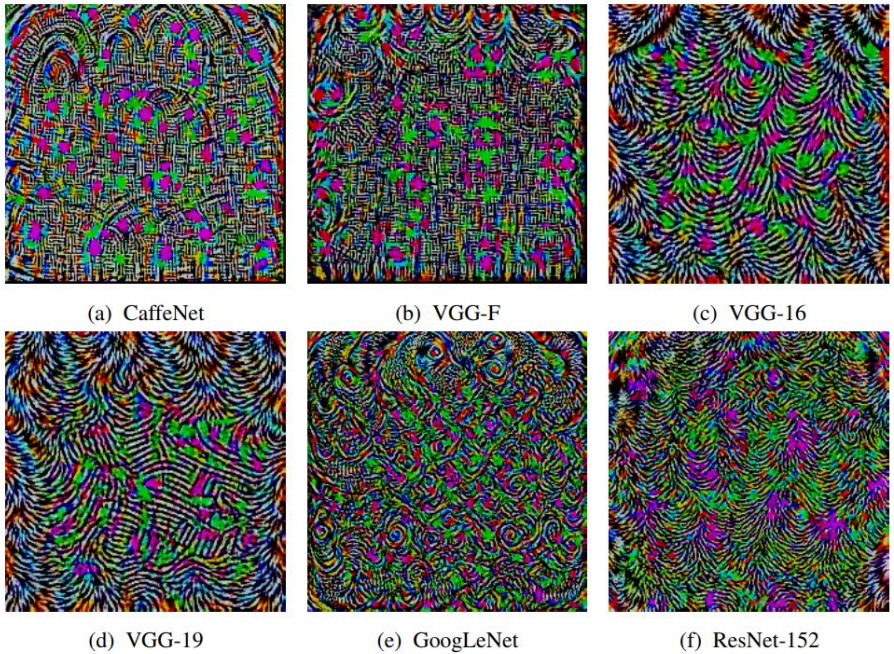
Bassinet(16.59%)
Paper Towel(16.21%)

Teapot(24.99%)
Joystick(37.39%)

Hamster(35.79%)
Nipple(42.36%)

P#1 Ataki przeciwstawne

- Bardzo często ataki optymalizowane są pod kątem wybranej architektury oraz zdjęcia
- **Uniwersalne perturbacje** - stałe zaburzenie niezależne od obrazka oraz architektury, które jest w stanie generować “skuteczne”



(a) CaffeNet (b) VGG-F (c) VGG-16

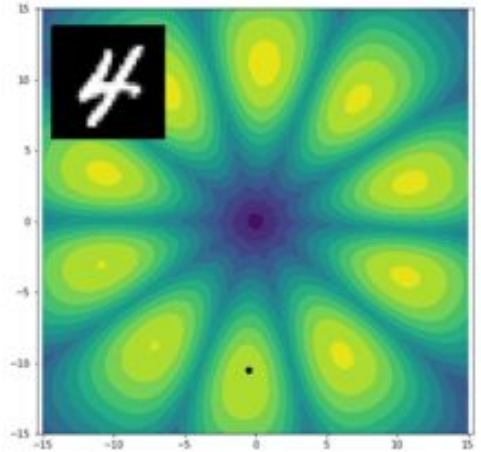
(d) VGG-19 (e) GoogLeNet (f) ResNet-152

P#1 Ataki przeciwstawne

- Jednym z powodów istnienia ataków przeciwstawnych jest silna nieliniowość sieci (niestabilność)

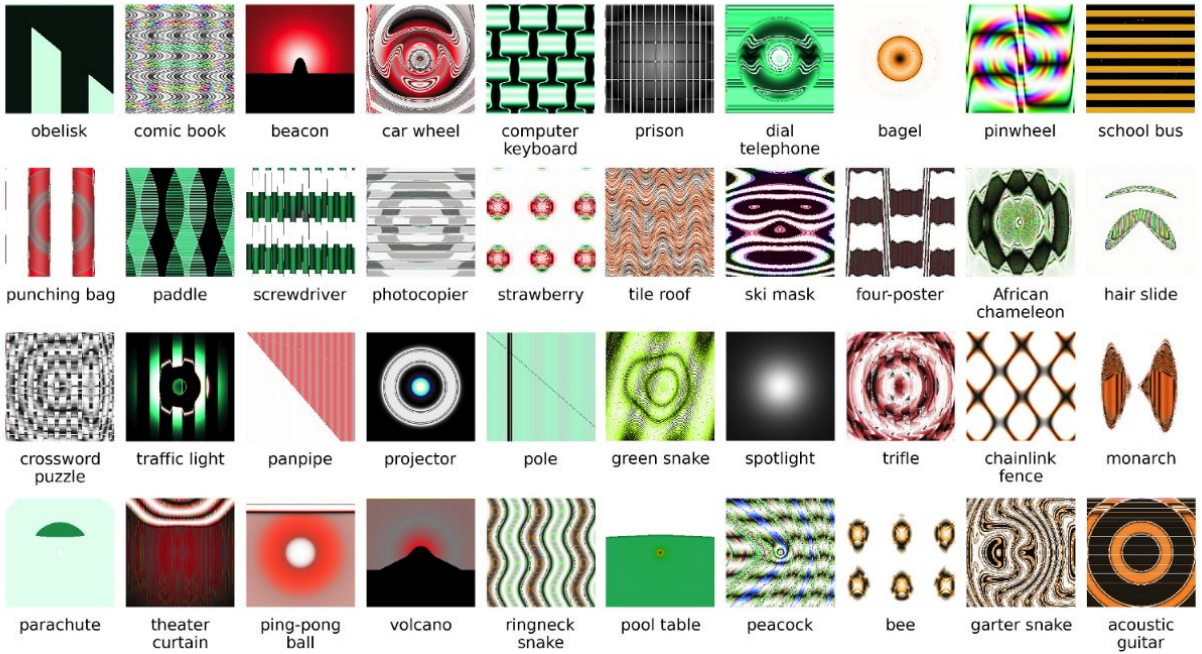
$$\|(\mathbf{x}+\varepsilon\mathbf{n})-\mathbf{x}\| \ll \|f(\mathbf{x}+\varepsilon\mathbf{n})-f(\mathbf{x})\|$$

- Istnieją metody których celem jest zwiększenie stabilności sieci, **nie ma metody która gwarantowałaby teoretyczną** (i praktyczną) odporność na dowolny atak
- jedną z najpopularniejszych metod są:
  - **trenowanie sieci z przykładami przeciwstawnymi** (wykorzystanie ataków jako źródła danych) - to może być czasochłonne
  - **regularyzacja macierzy wag** sieci neuronowych w celu zwiększenia stabilności poszczególnych bloków sieci neuronowej np. poprzez wymuszenie małych wartości osobliwych macierzy wag



P#1 Ataki przeciwstawne

- Ataki prze
- Dochodzi
- Przestrze
- Sieci neu



klas 1000

- gdzie z re
- to oznacz
- **w szczegó**

ą wartość **y**

- Uwaga: tego problemu nie da się rozwiązać!

P#1 Ataki przeciwstawne

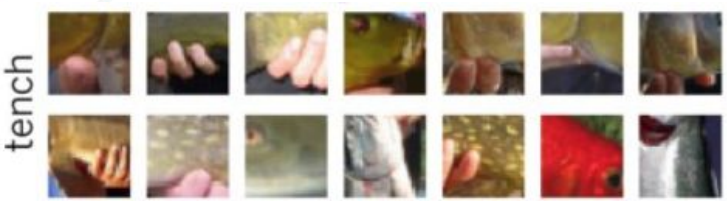
- Ataki przeciwstawne to realny i nierozwiązany problem
- Zmiana algorytmu (biblioteki) kompresji JPG może być wystarczająca aby zepsuć system
- Nowa dziedzina zajmuje się badaniem bezpieczeństwa sieci

Problem #2 Lenistwo - problem fałszywych przyjaciół

P#2 Lenistwo

- W problemie klasyfikacji obrazów sieć musi nauczyć się odwzorowania wektora **[W,H, 3] na N klas**
- To zmusza sieć do nauki **stratnej kompresji** informacji
- Sieć zatem musi na poszczególnych etapach przetwarzania decydować jaka informacja ma zostać odrzucona a jaka jest potrzebna do podjęcia ostatecznej decyzji
- W praktyce sieć szuka najprostszych korelacji które są w stanie zminimalizować koszt
- Przykład: klasa *lin (tench)* rozpoznawany jest po palcach

lin



P#2 Lenistwo

- Ukryte korelacje w danych, wynikają z małej liczby przykładów uczących

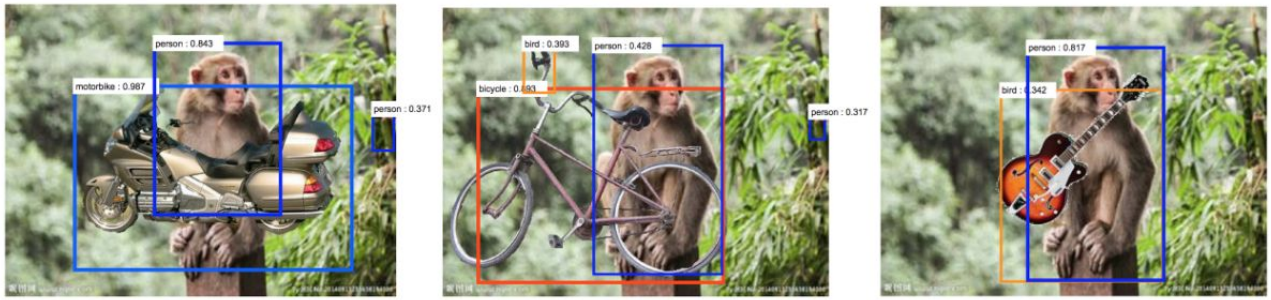
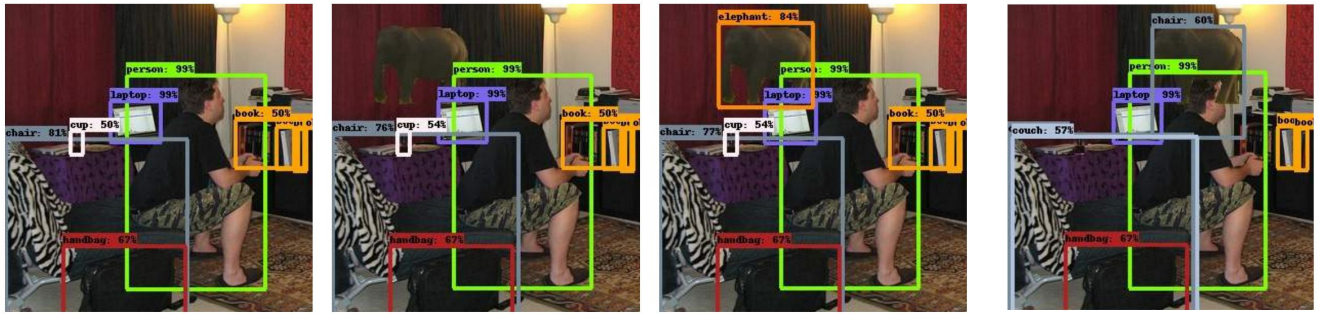


Figure 3: (Source). Adding occluders cause deep network to fail. Left: The occluding motorbike turns a monkey into a human. Center: The occluding bicycle turns a monkey into a human and the jungle turns the bicycle handle into a bird. Right: The occluding guitar turns the monkey into a human and the jungle turns the guitar into a bird.

- Praca: *The elephant in the room* (Aug-2018)



P#2 Lenistwo

- Przykład automatycznego podpisywania zdjęć: *image captioning*
- Problem żyraf oraz drzew ->
- Z drugiej strony system miał problem żeby wspomnieć o żyrafie jeśli na zdjęciu nie było drzew

- Uwaga: ten problem da się **rozwiązać częściowo (tj teoretycznie)**:
- wymaga to bardzo dużego i zróżnicowanego zbioru danych, tak żeby wyeliminować wszelkie przypadkowe korelacje: **eksplozja kombinatoryczna**

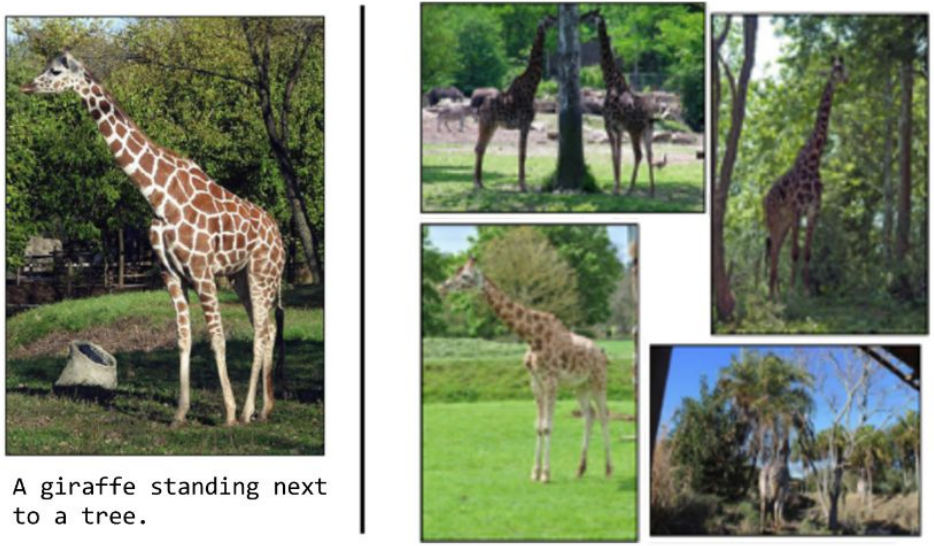
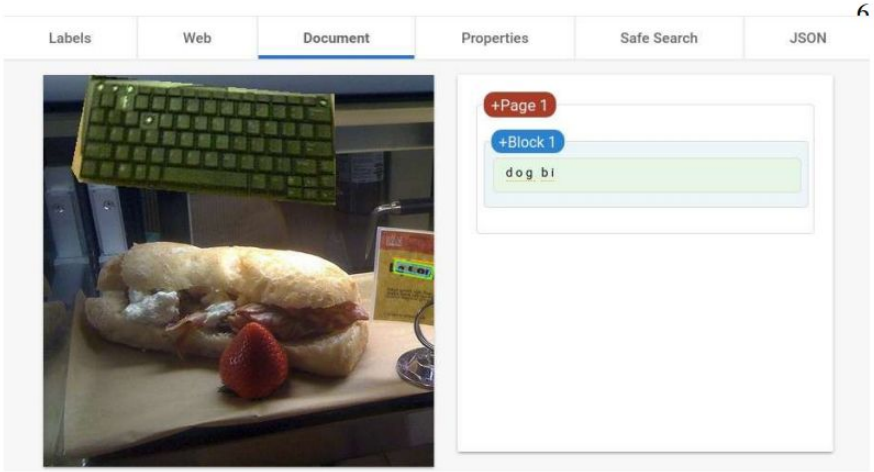


Figure 5.
Credit: Raul Puri, with images sourced from the MS COCO data set.

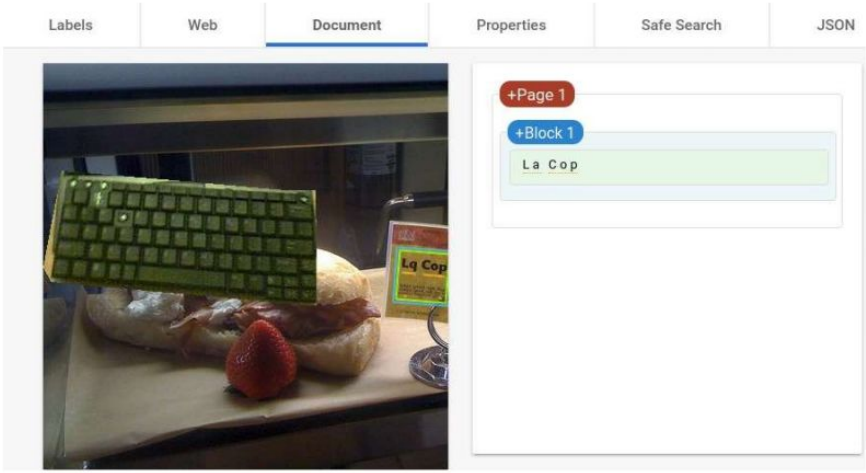
This is most certainly a "giraffe standing next to a tree." However, if we look at other pictures, we will likely notice that it generates a caption of "a giraffe next to a tree" for any picture with a giraffe because giraffes in the training set often appear near trees.

P#2B Efekty nielokalne

- Przykład detekcji tekstu na obrazie: Google Vision API



(a)

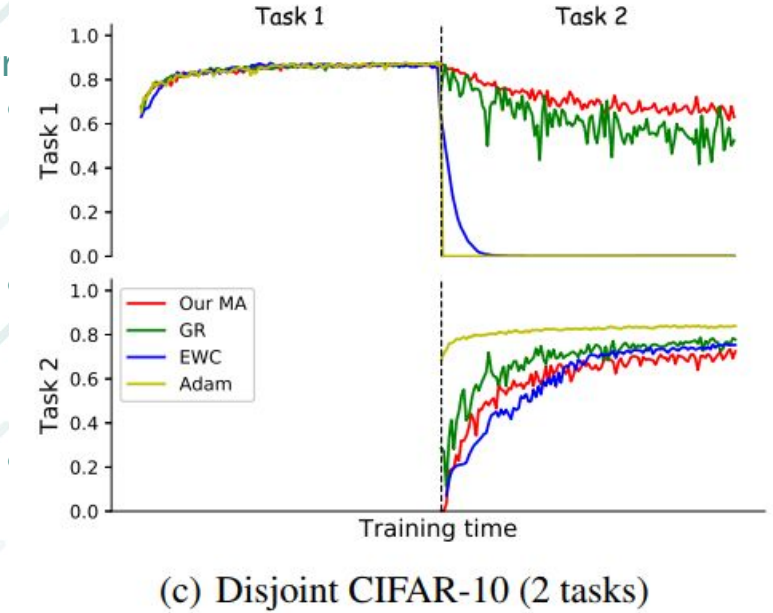


- Uwaga: **prawdopodobnie** nie da się tego rozwiązać na obecną chwilę

Problem #3 Pamięć

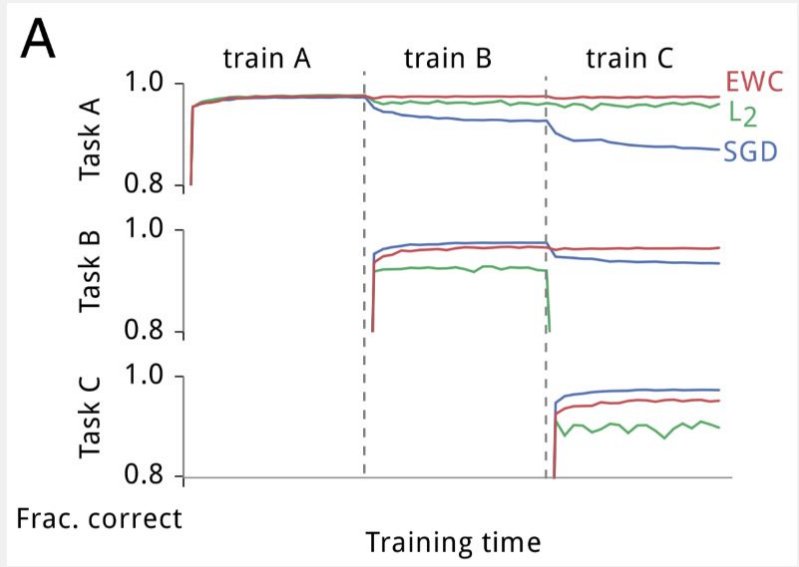
P#3 Pamięć

- Pamięć sieci neuronowych na temat świata jest zapisana w jej wagach
- Zapis odbywa się przez trenowanie sieci neuronowej
- Po wytrenowaniu nie mamy możliwości zmiany pamięci - pamięć jest statyczna
- Pr



(c) Disjoint CIFAR-10 (2 tasks)

y

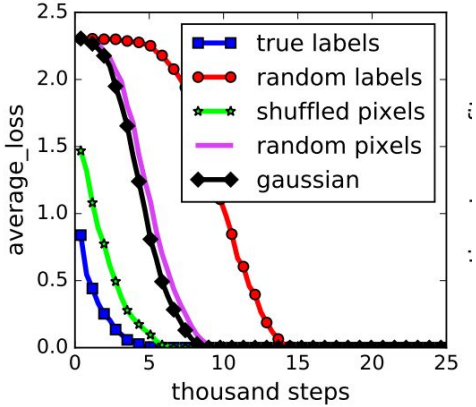


- Istnieją podejścia które mają temu zapobiec, ale w praktyce się tego nie stosuje albo nie ma testów na trudnych zadaniach

Problem #4 Sieć nauczy się praktycznie wszystkiego

P#4 Ogromna pojemność i zdolność do przetrenowania

- co za tym idzie może nauczyć szumu w naszych danych
- przeważnie wymaga ręcznego przejrzenia zbioru danych - czasochłonne,
- sieć może być 100% pewna w swojej predykcji w przypadku błędnych oznaczeń



(a) learning curves

- Istnieją podejścia które mają temu częściowo zapobiec, ale nie ma idealnego rozwiązania (teoretycznie optymalnego)

Problem #5 Sieć nie rozumie abstrakcyjnych pojęć

P#5 Problem z abstrakcyjnym myśleniem

- sieć nie jest w stanie nauczyć się zadań które nie posiadają prostych korelacji np. tekstury w obrazach
- **problem zliczania kropek** - czy ich liczba jest parzysta/nieparzysta ?
- sieci nie są wstanie tworzyć abstrakcyjnych reguł

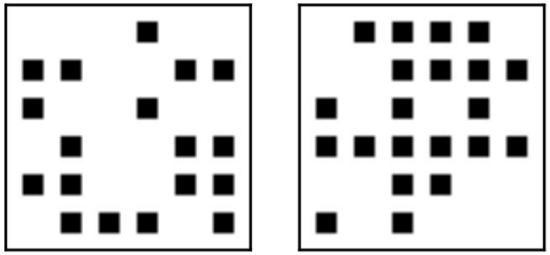


Figure 1: Two images of $13^2 = 169$ squares colored black with probability $1/2$. The left (right) image has an even (odd) number of black squares. The experiment illustrates the incapability of deep learning to learn the parity.

- Ten problem nie jest rozwiązany w żaden sposób

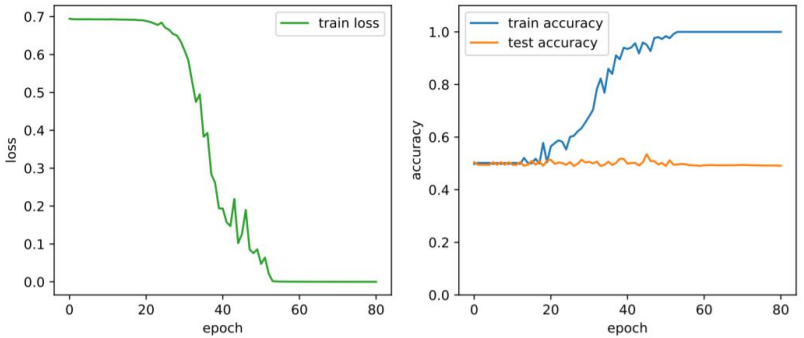


Figure 2: Training loss (left) and training/testing errors (right) for up to 80 SGD epochs.

Problem #6 Interpretowalność wyników

P#6 Problem interpretowalności wyników

- praktycznie niemożliwe jest odpowiedzenie na pytanie, dlatego model przewidział tą klasę a nie inną ?
- to ma szczególne znaczenie gdy model działa na danych wrażliwych
- istnieją całe grupy badawcze które zajmują się tym problemem
- jednak rozwiązania dotyczą konkretnych architektur bądź rozwiązywanych problemów
- praca: **Interpretable Deep Learning under Fire** (Dec 2018) - pokazuje że pewne metody interpretacji wyników też są podatne na ataki przeciwstawne

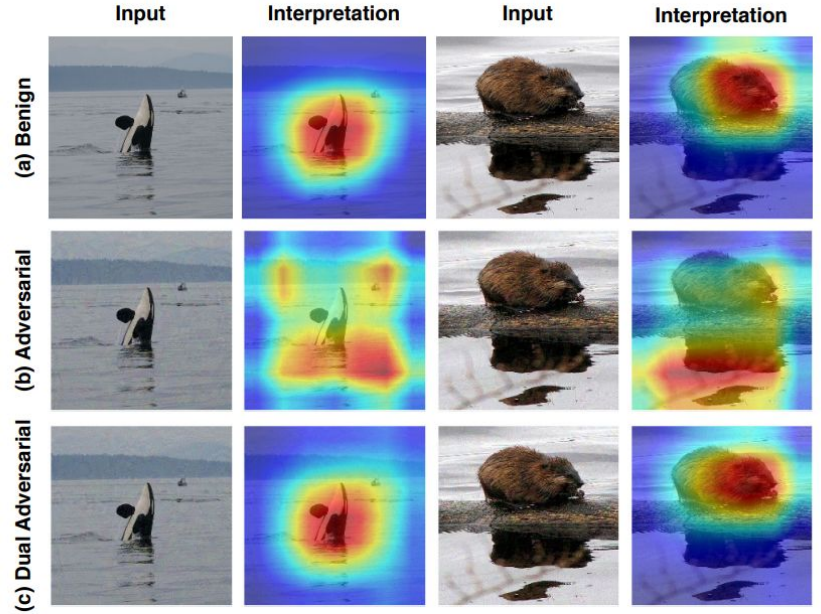
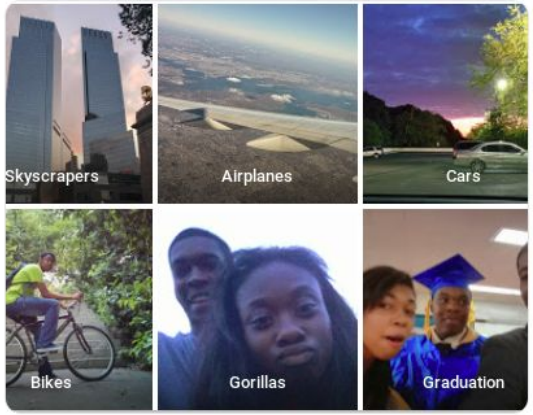


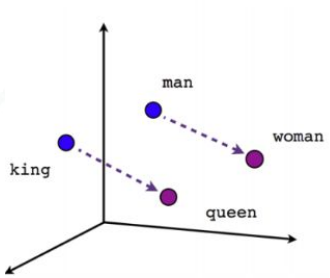
Figure 1: (a) benign, (b) regular adversarial, and (c) ACID adversarial inputs and interpretation, with RESNET [23] and CAM [72] as the classifier and interpreter respectively.

Problem #6B Problemy związane z Gender, rasizm i bezpieczeństwo

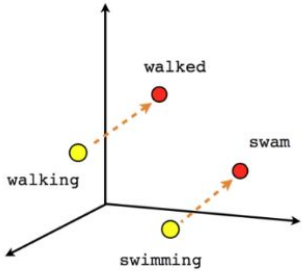
P#6B Interpretacja i wytłumaczenie wyników

- jeżeli w naszym datasecie są nie chciane korelacje to może on prowadzić do niepożądanych lub obraźliwych rezultatów
- **Problem Google Photos z 2015** - problem został rozwiązany: słowo gorillas zostało usunięte ze słownika tagów (ponoć do tej pory nierozwiązany)
- **Problem przetwarzania języka naturalnego** (NLP) - wektorach wyrazów (dataset Google News):

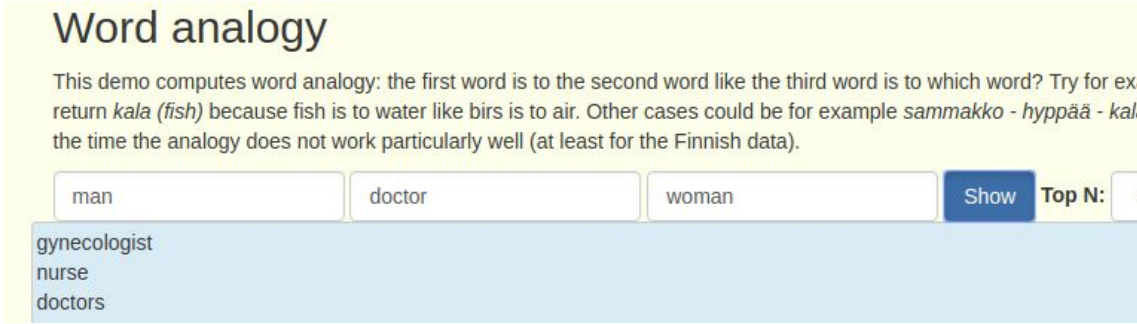




Male-Female



Verb tense



P#6B Niebezpieczne modele

- Generatywny model GPT-2 (OpenAI) posiada 1.5 miliarda parametrów (Feb 14, 2019)
- Potrafi np. generować dalszą część tekstu do rozpoczętego akapitu

SYSTEM PROMPT (HUMAN-WRITTEN)

Recycling is good for the world.

NO! YOU COULD NOT BE MORE WRONG!!

MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES)

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts. One of the best ways to start is to look at the process of creating a paper product. When you make a paper product, it is basically a long chain of materials. Everything from the raw materials (wood, cardboard, paper, etc.), to the reagents (dyes, solvents, etc.) to the printing equipment (chemicals, glue, paper, ink, etc.), to the packaging, to the packaging materials (mercury, chemicals, etc.) to the processing equipment (heating, cooling, etc.), to the packaging materials, to the packaging materials that are shipped overseas

Release Strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we

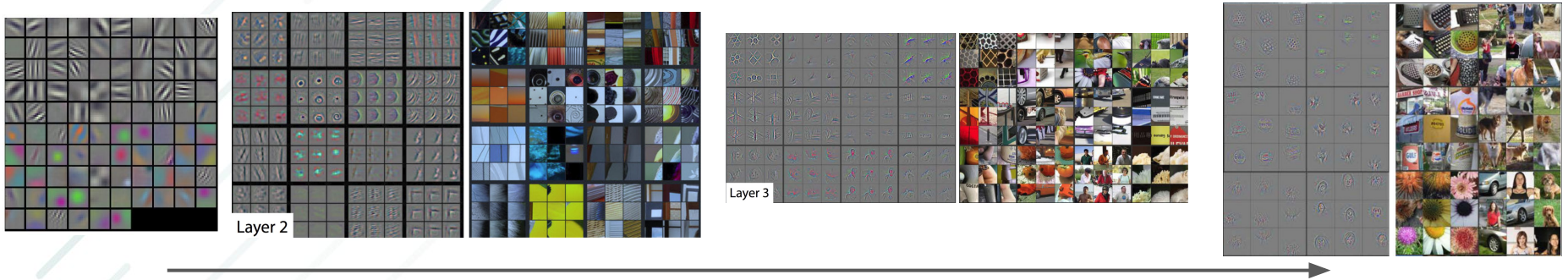
GPT-2 Interim Update, May 2019

We're implementing two mechanisms to responsibly publish GPT-2 and hopefully future releases: staged release and partnership-based sharing. We now releasing a larger 345M version of GPT-2 as a next step in staged release and are sharing the 762M and 1.5B versions with partners in the AI and security communities who are working to improve societal preparedness for large language models.

Problem #7 Brak teorii tworzenia sieci neuronowych

P#7 Brak formalizmu teoretycznego opisu sieci neuronowej

- nie ma czegoś takiego jak teoria tworzenia sieci neuronowych + są jedynie empiryczne wskazówki jakie rzeczy poskładane do razem dają dobre rezultaty,
- co za tym idzie, modele tworzone są głównie metodą prób i błędów,
- **interpretacja vs spekulacja**: np. hierarchia filtrów konwolucyjnych - głębsze warstwy uczą się kształtów (2014)



P#7 Brak formalizmu teoretycznego opisu sieci neuronowej

- okazuje się że (2019):

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

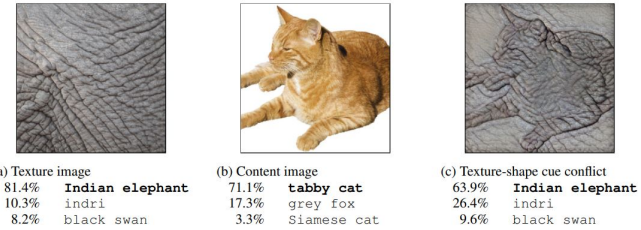
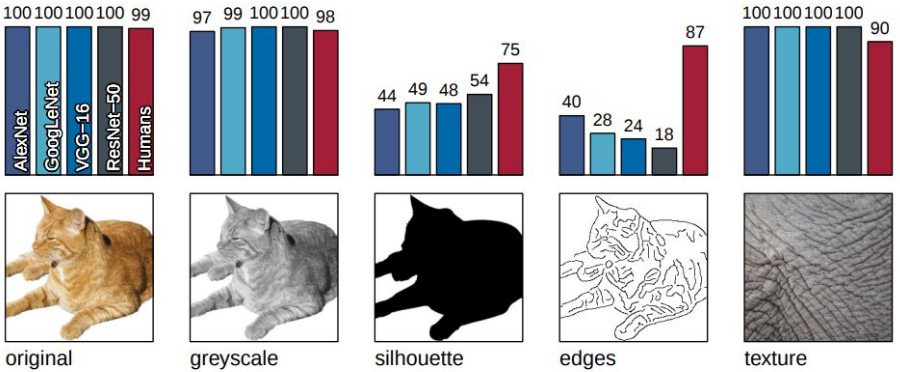


Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.



- autorzy pokazują że można zmusić sieci aby operowały bardziej na kształtach niż teksturach

P#7 Brak formalizmu teoretycznego opisu sieci neuronowej

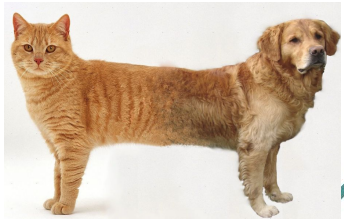
- **działa zasada:** co jest głupie ale działa to nie jest głupie: mixup (157 cytowań na google od początku ~2018)

Contribution Motivated by these issues, we introduce a simple and data-agnostic data augmentation routine, termed *mixup* (Section 2). In a nutshell, *mixup* constructs virtual training examples

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

- częściowo wynika to z tego, że dziedzina ta leży na granicy biznesu i nauki - liczy się czy działa, nie ważne jak i dlaczego

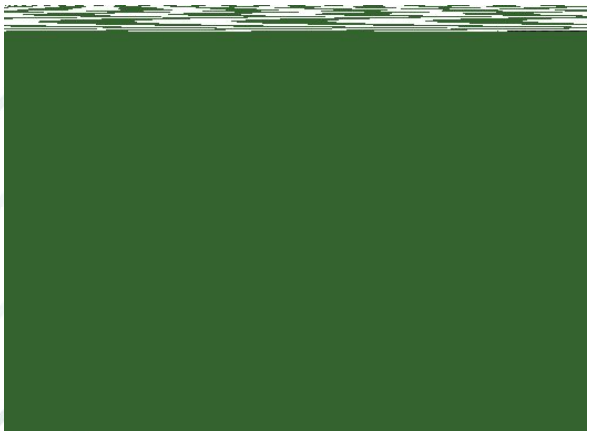
0.4 × data[1]:		cat: 1.0
+ 0.6 × data[2]:		dog: 1.0
= mixup:		cat: 0.4 dog: 0.6



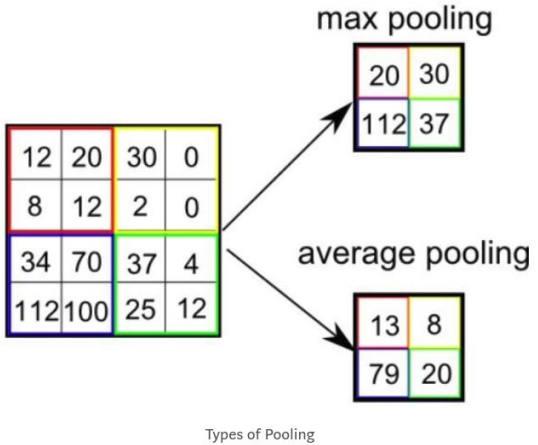
Problem #8 Przetwarzanie obrazów - brak niezmienników

P#8 Symmetrie

- sieci konwolucyjne są jedną z najpopularniejszych architektur
- wykorzystywane są w przetwarzaniu obrazów: detekcja obiektów, segmentacja, klasyfikacja itp
- jak i również w przetwarzaniu sygnałów dźwiękowych, NLP



Konwolucja - może zmniejszyć rozdzielczość

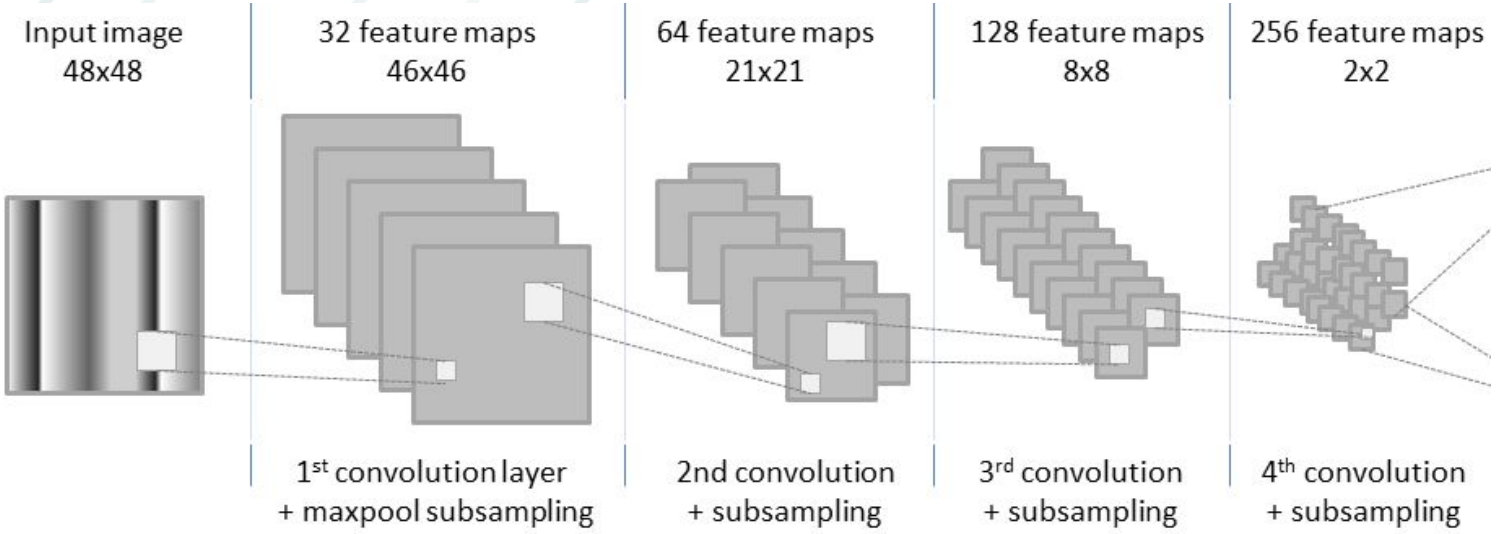


Types of Pooling

Próbkowanie - zmniejsza rozdzielczość

P#8 Symmetrie - translacje

- wielokrotne nałożenie na siebie operacji konwolucji oraz próbkowania redukuje wymiar przestrzenny mapy aktywacji



- wielokrotne nałożenie na siebie operacji konwolucji oraz próbkowania redukuje wymiar przestrzenny mapy aktywacji - tracimy symetrię translacyjną w przestrzeni aktywacji
- sieć może się nauczyć tej symetrii - ale nie jest ona wbudowana w jej architekturę
- bias w danych może przyczynić się do tego że nie nauczy się translacji

P#8 Symmetrie - translacje

- praca (**2019**): Why do deep convolutional networks generalize so poorly to small image transformations?

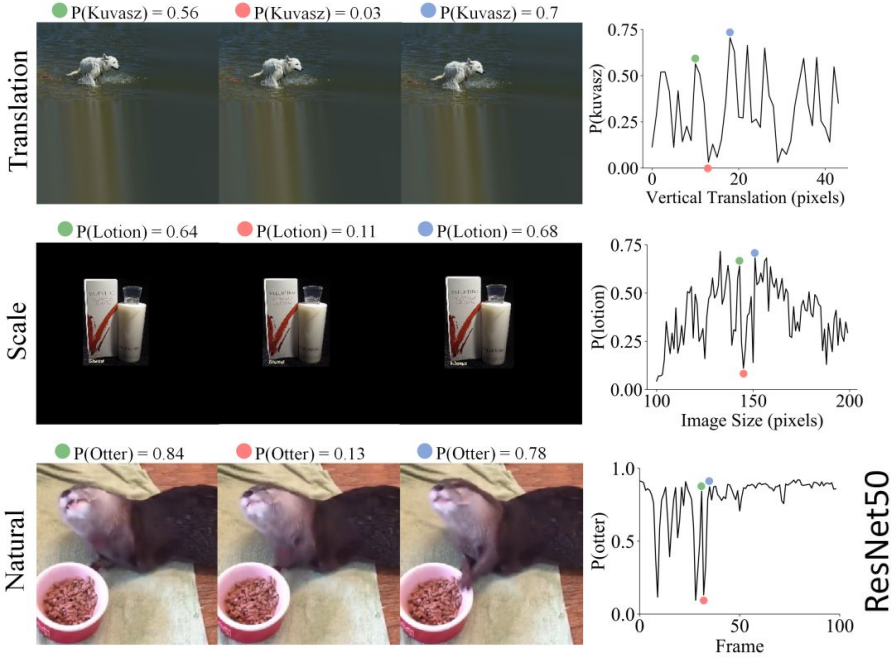
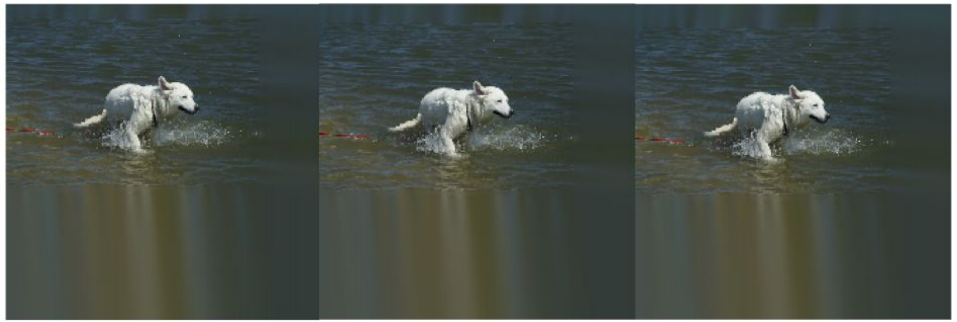
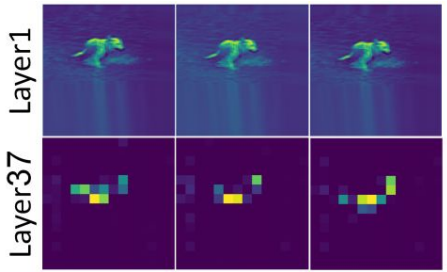
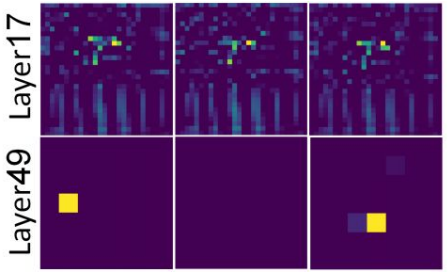


Figure 1: Examples of jagged predictions of modern deep convolutional neural networks. Top: A negligible vertical shift of the object (Kuvasz) results in an abrupt decrease in the network's predicted score of the correct class. Middle: A tiny increase in the size of the object (Lotion) produces a dramatic decrease in the network's predicted score of the correct class. Bottom: A very small change





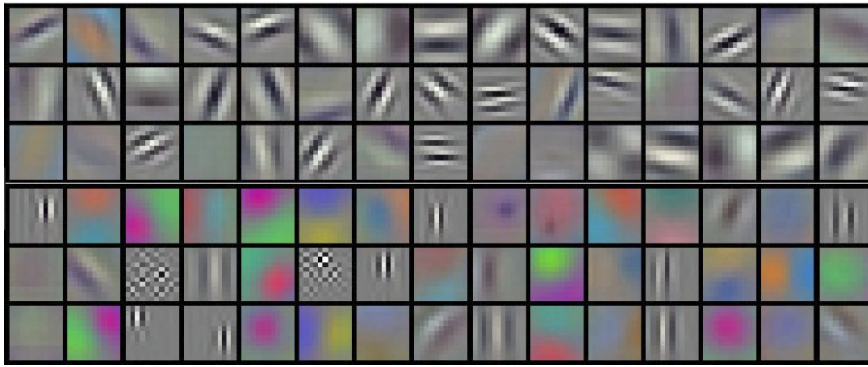


P#8 Symmetrie - skala i rotacja

- ten sam problem dotyczy skali obiektów i rotacji
- sieć musi wykształcić osobne filtry które potrzebne są do rozpoznawania skali obiektu, punktu widzenia, oświetlenia itp

Elevation \ Azimuth	90	135	180	225	270
0	-	0.713	0.769	0.930	0.319
30	0.900	1.000	0.588	1.000	0.710
60	0.255	0.100	0.148	0.296	0.649

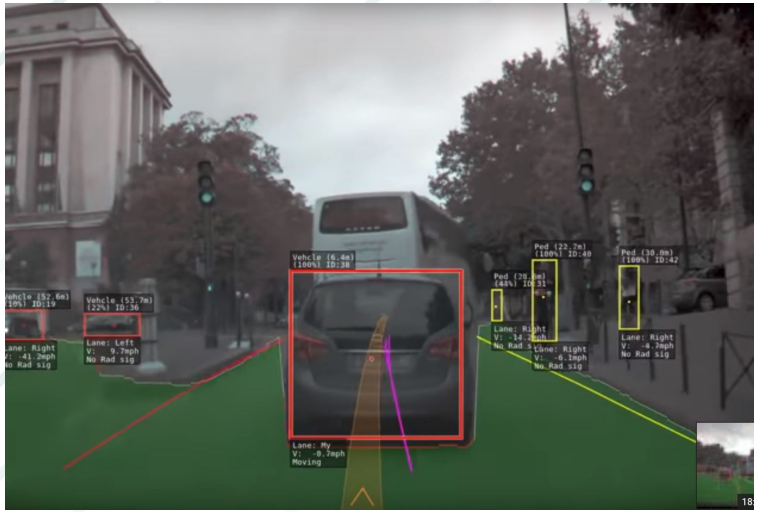
Figure 2: (Source). UnrealCV allows vision researchers to easily manipulate synthetic scenes, e.g. by changing the viewpoint of the sofa. We found that the Average Precision (AP) of Faster-RCNN detection of the sofa varies from 0.1 to 1.0, showing extreme sensitivity to viewpoint. This is perhaps because the biases in the training cause Faster-RCNN to favor specific viewpoints.

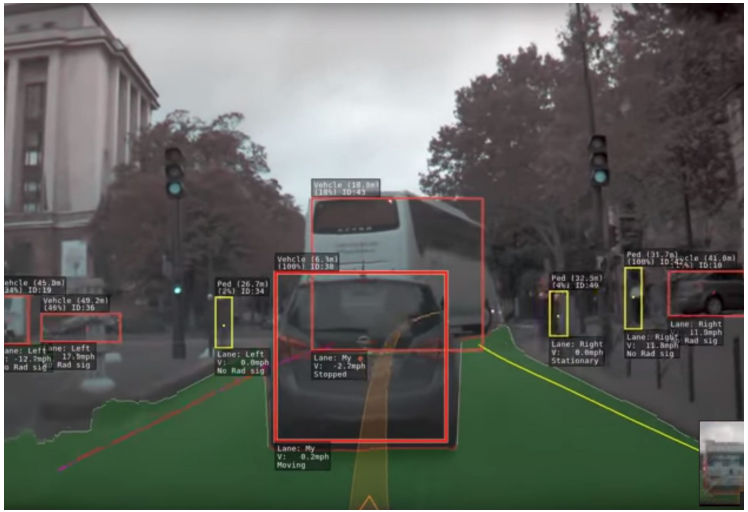


przykładowe filtry

P#8 Symmetrie - skala i rotacja

- przykład: autopilot Tesli - stabilność detekcji



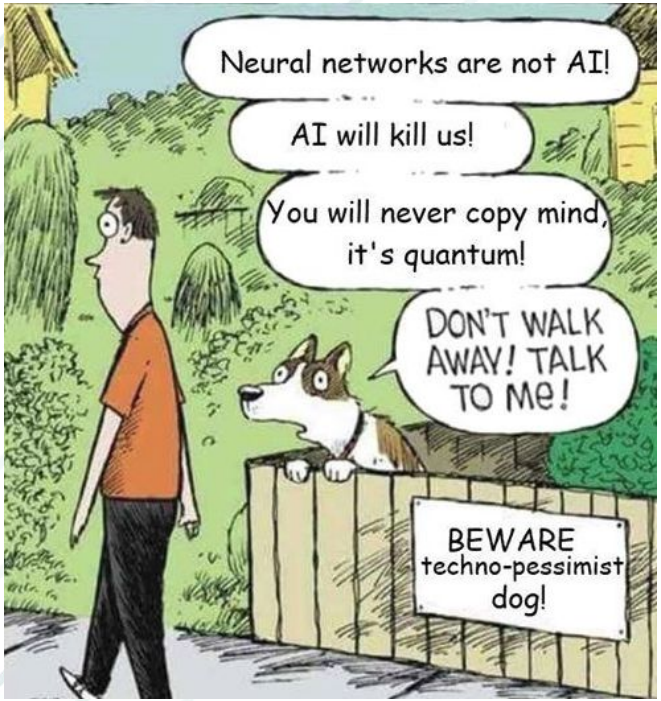


Podsumowanie

Podsumowanie

“deep learning is a terrific tool for some kinds of problems, particularly those involving perceptual classification, like recognizing syllables and objects, but also not a panacea”

Gary Marcus



Dziękuję za uwagę