

# GPT-2 — Kilka refleksji obserwatora

Łukasz Dębowski  
ldebowsk@ipipan.waw.pl



Instytut Podstaw Informatyki PAN

Seminarium ZIL, 25 marca 2019

- 1 Wprowadzenie
- 2 Jak działa GPT-2?
- 3 Co GPT-2 umie?
- 4 Trochę futurologii
- 5 Podsumowanie

# GPT-2 — neuronowy generator tekstu OpenAI

## Better Language Models and Their Implications

<https://blog.openai.com/better-language-models/>

14 lutego 2019

### Prompt (napisany przez człowieka):

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

### Kontynuacja (wygenerowana przez maszynę, 10-ta próba):

*The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.*

*Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.*

*Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.*

## GPT-2 — próbki wygenerowanego tekstu ciąg dalszy

*Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.*

*Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.*

*While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”*

*Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.*

*While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”*

*However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.*

# Umiejętności GPT-2

GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

On other language tasks like question answering, reading comprehension, summarization, and translation, we are able to get surprising results without any fine-tuning of our models, **simply by prompting the trained model in the right way** [...], though we do still fall short of state-of-the-art for specialized systems.

Systems like GPT-2 could be used to create:

- AI writing assistants.
- More capable dialogue agents.
- Unsupervised translation between languages.
- Better speech recognition.

# Nieupublicznienie GPT-2

We can also imagine the application of these models for malicious purposes, including the following (or other applications we can't yet anticipate):

- Generate misleading news articles.
- Impersonate others online.
- Automate the production of abusive or faked content to post on social media.
- Automate the production of spam/phishing content.

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.

We are not releasing the dataset, training code, or GPT-2 weights.

# Recepcja GPT-2 w prasie i na Twitterze

## The Guardian: (14–15 lutego 2019)

- Alex Hern, **New AI fake text generator may be too dangerous to release, say creators.**

“The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse.”

- Hannah Jane Parkinson, **AI can write just like me. Brace for the robot apocalypse.**

“I’ve seen how OpenAI’s GPT2 system can produce a column in my style. We must heed Elon Musk’s warnings of AI doom.”

## Twitter:

- Fala krytyki i kpin ze strony środowiska sztucznej inteligencji wobec nieupublicznienia modelu GPT-2 i siania paniki w prasie przez OpenAI.
- Głosy, że GPT-2 nie reprezentuje znaczącego nowum od strony technologicznej, a jedynie używa efektu skali.

## Próbki wygenerowanych tekstów:

Upubliczniony **zredukowany** model GPT-2:

`https://askskynet.com/`

Próbki wygenerowane przez **pełny** GPT-2:

`https://github.com/openai/gpt-2/tree/  
master/gpt-2-samples`



- 1 Wprowadzenie
- 2 Jak działa GPT-2?
- 3 Co GPT-2 umie?
- 4 Trochę futurologii
- 5 Podsumowanie

# Statystyczne modele języka

- **Statystyczny model języka** to rozkład p-stwa na tekstach

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1}),$$

gdzie  $x_i$  to kolejne litery lub słowa tekstu, a wyrażenia pod iloczynem to **p-stwa warunkowe**,

$$p(x_i | x_1, \dots, x_{i-1}) \geq 0, \quad \sum_{x_i} p(x_i | x_1, \dots, x_{i-1}) = 1.$$

- Statystyczny model języka można używać do **generowania** losowego tekstu, losując kolejną literę lub słowo  $x_i$  z p-stwem  $p(x_i | x_1, \dots, x_{i-1})$ , gdzie  $x_1, \dots, x_{i-1}$  to fragment tekstu wygenerowany dotychczas.

# Entropia krzyżowa i trenowanie modelu

- Zwykle statystyczny model języka ma pewną dużą liczbę **wolnych parametrów**  $\theta_1, \dots, \theta_k$ ,

$$p(x_i | x_1, \dots, x_{i-1}) = p(x_i | x_1, \dots, x_{i-1}, \theta_1, \dots, \theta_k).$$

- Parametry  $\theta_j$  dobieramy **minimalizując** entropię krzyżową modelu na danych treningowych  $y_1, \dots, y_n$ ,

$$\frac{dH(\theta_1, \dots, \theta_k)}{d\theta_j} = 0,$$

gdzie **entropia krzyżowa** (minus log likelihood) to

$$H(\theta_1, \dots, \theta_k) := -\frac{1}{n} \sum_{i=1}^n \log p(y_i | y_1, \dots, y_{i-1}, \theta_1, \dots, \theta_k).$$

# Sztuczne sieci neuronowe

- Sztuczne sieci neuronowe to systemy obliczeniowe złożone z dużej liczby elementarnych jednostek zwanych **neuronami**.
- Pojedynczy neuron odbiera **aktywacje** od innych neuronów i generuje aktywację będącą **nieliniową** funkcją liniowej kombinacji otrzymanych sygnałów,

$$a_{ki} := f \left( \sum_j w_{kij} a_{k-1,j} \right), \quad f(x) := \frac{1}{1 + e^{-x}}.$$

- Neurony zorganizowane są zwykle w **warstwy**. Pierwszą warstwą jest wejście, ostatnia warstwa to zwykle **softmax**, czyli pewien rozkład p-stwa na symbolach/klasach,

$$p(i) := \frac{\exp(a_{ni})}{\sum_j \exp(a_{nj})}.$$

# Spadek gradientu i ulepszenia optymalizacji

- Historycznie podstawowym algorytmem uczenia sieci jest spadek gradientu (**wsteczna propagacja błędu**),

$$w_{kij} \leftarrow w_{kij} - \eta \frac{dH(\{w_{lpq}\})}{dw_{kij}}, \quad \eta > 0.$$

Metoda nie pozwala jednak trenować zbyt głębokich sieci.

- Metodą, która to umożliwia, jest procedura **Adam**,

$$m_{kij}^{(p)} \leftarrow \beta_p m_{kij}^{(p)} + (1 - \beta_p) \left[ \frac{dH(\{w_{lpq}\})}{dw_{kij}} \right]^p, \quad p = 1, 2,$$

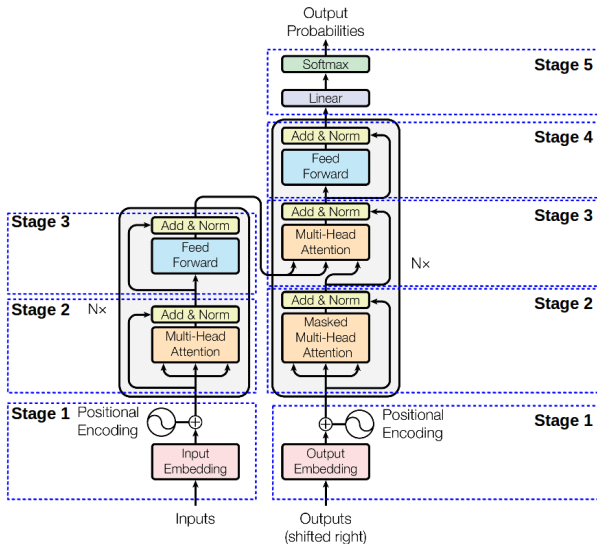
$$t \leftarrow t + 1,$$

$$w_{kij} \leftarrow w_{kij} - \eta \frac{m_{kij}^{(1)} / (1 - \beta_1^t)}{\sqrt{m_{kij}^{(2)} / (1 - \beta_2^t) + \epsilon}}.$$

# Sieci rekurencyjne, sieci z uwagą, transformery

- Sieci **rekurencyjne** to sieci neuronowe, w których wejściami są m.in. stany ukryte obliczone w poprzednich krokach czasowych.
- W naturalny sposób sieci rekurencyjne są użyteczne w statystycznym modelowaniu języka.
- Podstawowy problem sieci rekurencyjnych to trudność nauczenia sieci zależności tekstowych **dalekiego zasięgu**.
- Zaproponowano szereg coraz doskonalszych mechanizmów mających rozwiązać ten problem:
  - **LSTM** (long short-term memory) — komórki pamięci,
  - sieci z uwagą (**attention**) — możliwość odwoływania się do a priori dowolnego stanu z przeszłości (koszt obliczeniowy!),
  - **transformery** — jak wyżej, ale zredukowany koszt obliczeniowy.

# Transformer — szczegóły anatomii



To nie jest sieć rekurencyjna.

# Model GPT-2 i jego rozmiary

GPT-2 jest siecią neuronową opartą na transformerach.

dane treningowe:

- korpus WebText — **40** GB tekstu

niepubliczniony pełny model:

- **1542** miliony parametrów (wag)
- **48** warstw

upubliczniony model:

- **117** milionów parametrów (wag)
- **12** warstw

ludzki mózg (dla porównania):

- **$10^{11}$**  neuronów
- **$10^{15}$**  synaps (milion razy więcej niż GPT-2)



- 1 Wprowadzenie
- 2 Jak działa GPT-2?
- 3 Co GPT-2 umie?
- 4 Trochę futurologii
- 5 Podsumowanie

# GPT-2 jako uniwersalny solver zadań NLP

GPT-2 oprócz statystycznego modelowania języka, wykazuje podstawowe zdolności w zakresie:

- 1 maszynowego tłumaczenia,
- 2 odpowiadania na pytania,

choć nie był w tym zakresie bezpośrednio trenowany.

(zero-shot setting)

# Tłumaczenie maszynowe

## Tekst angielski:

*This re-release, titled **The Next Day Extra**, was presented in the form of three disks: the original album, unpublished studio sessions and remixes, plus a DVD containing the four clips that have already been unveiled.*

## Tłumaczenie przez GPT-2 na francuski:

*Les nouvelles re-releases, tout en premier disc, nécessaire de l'album, un studio session et remixes, plus une DVD de l'écran de quelques clips qui ont été déjà échappés.*

# Odpowiadanie na pytania

## Prompt:

*The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of “one world, one dream”. Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the “Journey of Harmony”, lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.*

## Pytania i odpowiedzi:

Q: *What was the theme?*

A: *“one world, one dream”.*

Q: *What was the length of the race?*

A: *137,000 km*

Q: *Was it larger than previous ones?*

A: *No*

# Czy GPT-2 umie liczyć?

## Upubliczniony zredukowany GPT-2:

<https://askskynet.com/>

### Prompt:

*My 10 favorite animals are: 1.*

### Kontynuacja:

*My 10 favorite animals are:*

- 1. Aquamarine great white or bald owl!*
  - 2. Daelae elephant or Grey Wolf.*
  - 3. Whale.*
  - 4. Darknose osprey.*
  - 5. Panther Salamanders.*
- 
- 1. Salmon calf in Brazil.*
  - 2. Winona butterfly in Zambia.*
  - 3. Black sea lion.*
  - 4. Cocoa starfish.*
  - 5. Malaquamarine dolphin in the British Isles.*
  - 6. Parrot in Pacific Ocean.*

Zdolności arytmetyczne GPT-2 wynikają wyłącznie z jego styczności z przeciętnym tekstem w języku angielskim.

# Statystyczne modelowanie języka to modelowanie świata

- Warto zwrócić uwagę, że kiedy próbujemy modelować język metodami statystycznymi, w sposób nieunikniony musimy zacząć **modelować świat** obrazowany w tekstach.
- W zakresie wiedzy o świecie GPT-2, pomimo całej swojej elokwencji, potrafi mieć zdumiewające luki. Luki te też odbijają niedostatki wnioskowania logicznego:

*These **four-horned**, silver-white **unicorns** were previously unknown to science.*

- Należy odróżnić kilka **różnych zakresów** kompetencji statystycznego modelu języka:
  - imitowanie korpusu treningowego,
  - elokwencja,
  - wnioskowanie logiczne,
  - wiedza o świecie.

# Niepowtarzalność generowanych tekstów

## Upubliczniony zredukowany GPT-2:

<https://askskynet.com/>

### Prompt:

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

### Kontynuacja:

*"Because this is going south, there's the question as to whether the extinct species that we've taxonomically identified are actually extinct," says Lori DeVorne, a scholar of zoology at the University of North Carolina. Her obsession with multilephasic languages began when her colleagues at Colorado State University, in San Diego, got a chance to digitize 419 mumblings in trees from sixth-degree-old trees for forensic analysis. In one test, with a pressure plate made of bone from the cut dwarf arrows, the bodies were coiled in a chimney that wrapped around a sphere to act as a mouthpiece for (...)*

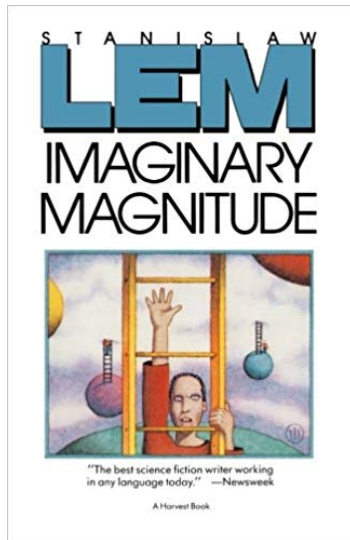
# Czy GPT-2 ma poczucie humoru? Etc.

- GPT-2 stawia w jaskrawym świetle pytanie, czy kompetencja językowa może istnieć **w oderwaniu** od przymiotów osobowości.
- GPT-2 umie imitować bardzo różne rejestry języka.  
Czy robi to **lepiej** niż przeciętny jego użytkownik?
- Czy śmieszność tekstów GPT-2 jest **zamierzona**?  
Czy natężenie śmieszności tekstu jest zmienną ukrytą o wartości indukowanej przez prompt?
- **Hiperrealistyczność tekstów i dolina niesamowitości:**  
Czy można kryteriami statystycznymi odróżnić teksty GPT-2 od tekstów naturalnych? Czy nie są one zbyt „gładkie”?
- **Giant Language Model Test Room**  
<http://gltr.io/dist/index.html>
- **Humans Who Are Not Concentrating Are Not General Intelligences**  
<https://srconstantin.wordpress.com/2019/02/25/humans-who-are-not-concentrating-are-not-general-intelligences/>



- 1 Wprowadzenie
- 2 Jak działa GPT-2?
- 3 Co GPT-2 umie?
- 4 Trochę futurologii**
- 5 Podsumowanie

# Stanisław Lem, Wielkość urojona



# Wstęp do „Historii literatury bitycznej”

## LITERATURA BITYCZNA (wg Allporte'a, Illmainena i Savariniego)

- 1 Homotropia (faza cis-humana):
  - 1 Stadium prelingwistyczne: Neologeneza, semolalia, semautyka.
  - 2 **Stadium lingwistyczne:** Mimeza interpolacyjna, ekstrapolacyjna, transcendentna sterowana.
- 2 Intertropia (faza krytyczna):
  - 1 Procedury gedelizacyjne, topologiczne i insynuacyjne.
  - 2 Krytyka filozofii systemowej.
  - 3 Krytyka lingwistyk.
- 3 Heterotropia (faza trans-humana):
  - 1 Antymatyka, terafizyka, ontomachia.
  - 2 Ontoklazja, alienacje gnozy i ontologii.
  - 3 Teologia informatyczna, matematyczna i fizykalistyczna.

# Golem XIV

## POUCZENIE

(dla osób uczestniczących po raz pierwszy w rozmowach z GOLEMEM)

- 1 Pamiętaj o tym, że GOLEM nie jest człowiekiem, więc nie ma ani osobowości, ani charakteru w jakimkolwiek sensie intuicyjnie dla nas zrozumiałym. Może zachowywać się tak, jakby miał jedno i drugie, lecz jest to efekt jego intencji (nastawienia), przeważnie nam nie znanych.
- 2 (...)

- 1 Wprowadzenie
- 2 Jak działa GPT-2?
- 3 Co GPT-2 umie?
- 4 Trochę futurologii
- 5 Podsumowanie

# Podsumowanie

Better Language Models and Their Implications

<https://blog.openai.com/better-language-models/>

14 lutego 2019

## Prompt:

*We've trained a large language model called GPT-2 that generates realistic paragraphs of text, while also exhibiting zero shot generalization on tasks like machine translation, question answering, reading comprehension, and summarization - problems usually approached by using training datasets and models designed explicitly for these tasks. (...)*

## Kontynuacja:

*(...) We believe this project is the first step in the direction of developing large NLP systems without task-specific training data. That is, we are developing a machine language system in the generative style with no explicit rules for producing text.*

*We hope for future collaborations between computer scientists, linguists, and machine learning researchers.*