

KSBI-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)



생성형 AI 기술의 이해

이상근 _ 고려대학교



KSBI
KOREAN SOCIETY FOR
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBI-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics & Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의를 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

생성형 AI 기술의 이해

최근 몇 년간 생성형 인공지능(Generative AI)은 다양한 산업과 일상생활에 깊이 스며들며 혁신을 이끌고 있습니다. 생성형 AI는 텍스트, 이미지, 음악, 영상 등 인간이 만들어내던 콘텐츠를 자동으로 생성할 수 있는 기술로, 특히 트랜스포머(Transformer)와 같은 딥러닝 기술의 발전 덕분에 빠르게 성장해 왔습니다. 이 강의에서는 생성형 AI의 개념과 이를 가능하게 하는 핵심 기술을 쉽게 이해할 수 있도록 소개합니다.

본 강의의 목표는 다음과 같습니다.

- 텍스트 생성형 AI의 중심 기술인 트랜스포머(Transformer) 구조의 이해
- 챗GPT와 같은 LLM의 개념, 신경망 구성, 학습 방법을 이해
- 연구를 위한 ChatGPT 활용 사례 소개
- 이미지 생성 기술인 디퓨전 (diffusion) 모델의 원리를 소개

* 강의 난이도: 초급

* 강의: 이상근 교수 (고려대학교 정보보호학과)

Curriculum Vitae

Speaker Name: Sangkyun Lee, Ph.D.



► Personal Info

Name Sangkyun Lee
Title Associate professor
Affiliation Korea University

► Contact Information

Address 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea
Email sangkyun@korea.ac.kr

Research Interest

Trustworthy AI, Robust deep learning methods, AI for security, Data analysis

Educational Experience

2003 B.S., Seoul National University
2005 M.S., Seoul National University
2011 Ph.D., University of Wisconsin-Madison, USA

Professional Experience

2011-2014 Post-doc Researcher, SFB 876, TU Dortmund University, Germany
2015-2017 Principal Investigator, SFB 876, TU Dortmund University, Germany
2017-2019 Assistant Professor, Department of Computer Science, Hanyang University ERICA
2020-2021 Assistant Professor, School of Cybersecurity, Korea University
2022-current Associate Professor, School of Cybersecurity, Korea University

Selected Publications (3 maximum)

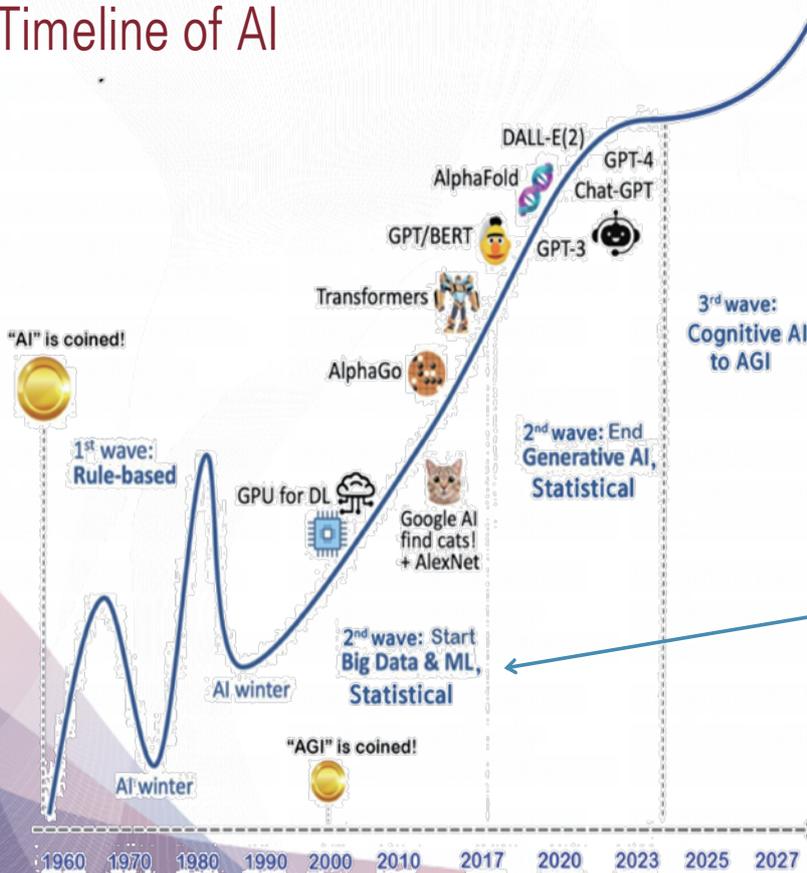
1. SwiftThief: Enhancing Query Efficiency of Model Stealing by Contrastive Learning, Jeonghyun Lee, Sungmin Han, Sangkyun Lee, IJCAI, 2024
2. Libra-CAM: An Activation-Based Attribution Based on the Linear Approximation of Deep Neural Nets and Threshold Calibration, Sangkyun Lee, Sungmin Han, IJCAI, 2022
3. Model Stealing Defense against Exploiting Information Leak Through the Interpretation of Deep Neural Nets, Jeonghyun Lee, Sungmin Han, Sangkyun Lee, IJCAI, 2022

생성형 AI 기술의 이해 Introduction to Generative AI Technology

고려대학교 정보보호대학원
인공지능연구실
이상근 교수

KSBI-BIML 2025

Timeline of AI



Vladmir Vapnik (1936~)



ChatGPT란?

• 자연어 처리 모델



컴퓨터 ⇄ 사람
의사 소통 연구

• 생성 모델

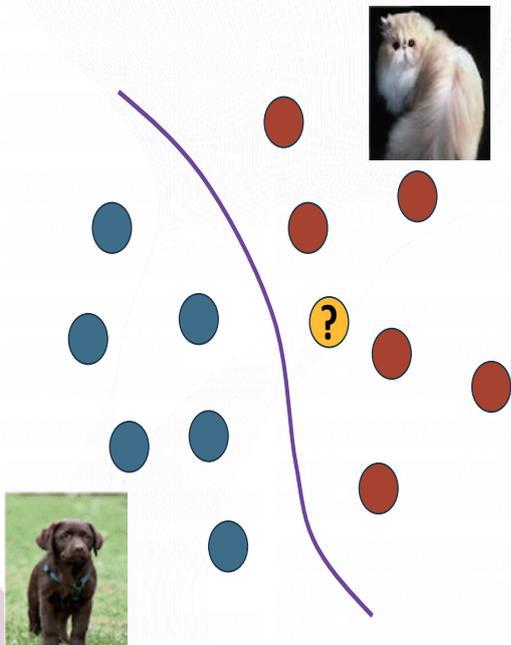


데이터의 패턴/구조 학습
→ 문장, 영상, 음악 등
새로운 데이터 생성

• OpenAI 

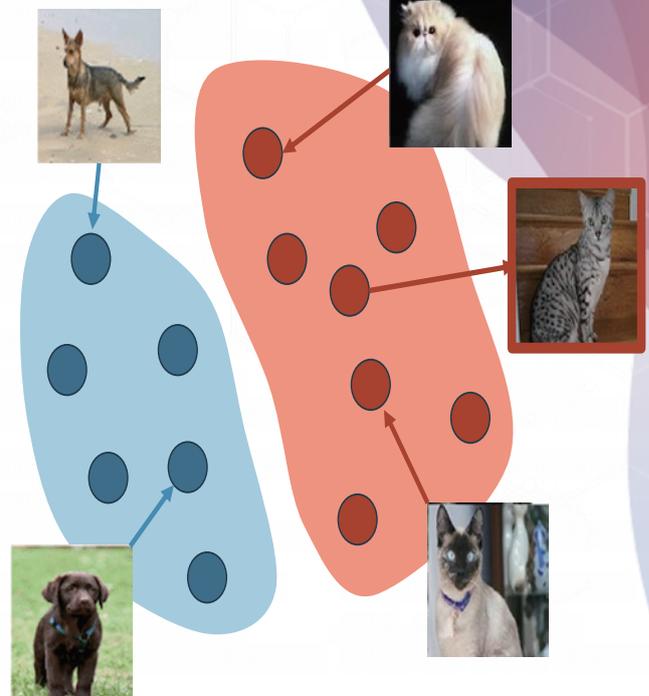


일론 머스크 등이 설립 (2015)
인간 친화적 AI 개발 목적
ChatGPT, Codex, DALL-E 등
개발



판별형(Discriminative) AI

클래스 판별을 위한 결정경계면 학습



생성형(Generative) AI

데이터 생성을 위한 확률분포 학습

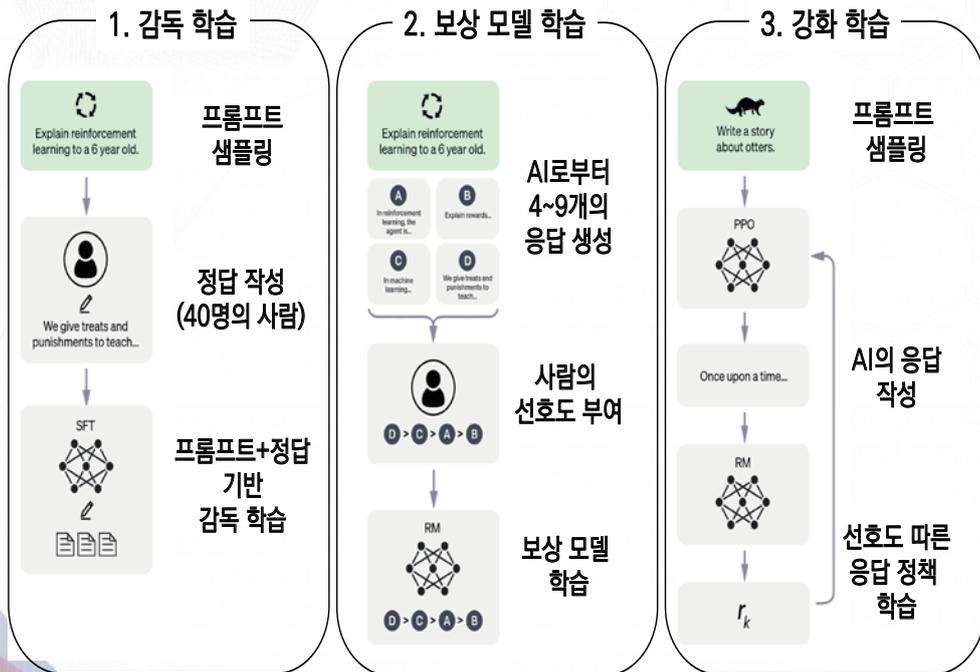
ChatGPT의 학습(1): 사전학습

- 다양한 데이터: 책, 웹페이지, 논문 등
- 대용량 데이터: 약 45TB (미국 의회 도서관 4.5배)
- 비감독 학습
- 언어 모델링
- GPU 기반 대규모 병렬 학습

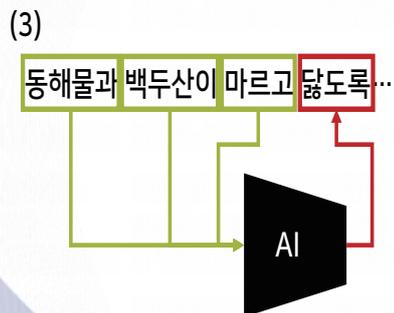
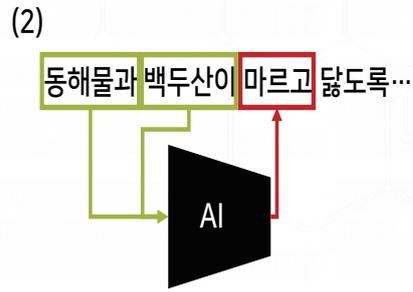
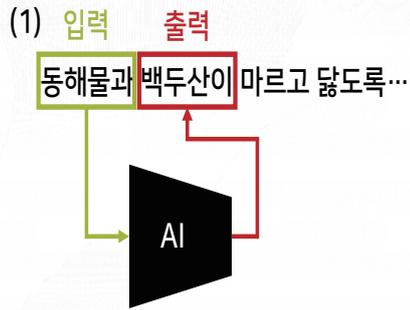
Training language models to follow instructions with human feedback, NeurIPS (2022)

ChatGPT의 학습(2): 미세 조정

Training language models to follow instructions with human feedback, NeurIPS (2022, InstructGPT)

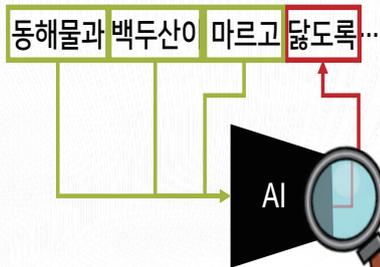


NLP AI: 언어 모델 학습

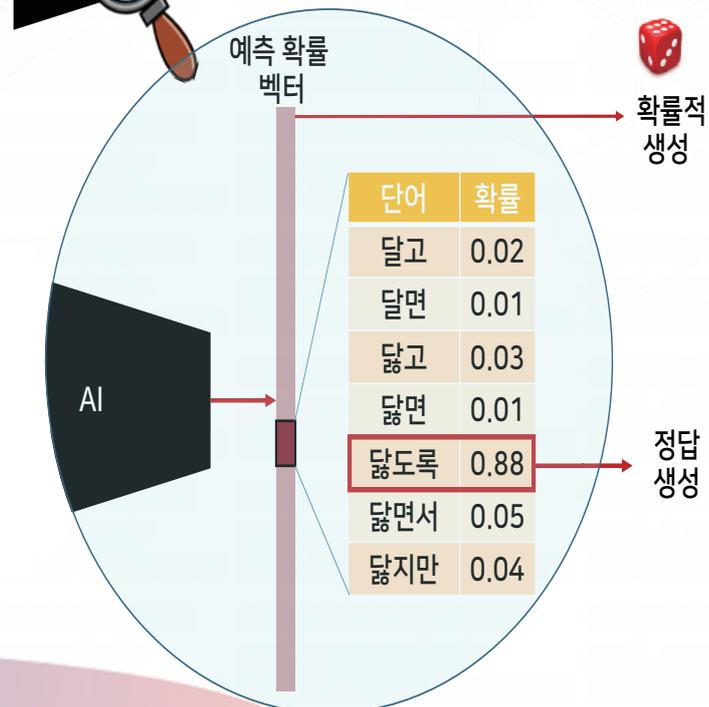


자연어 처리 작업

- 이전 단어 → 다음 단어
- 이전 구절 → 다음 구절
- 질문 → 답변
- 빈칸 채우기
- 키워드 자동완성
- 챗봇
- ...

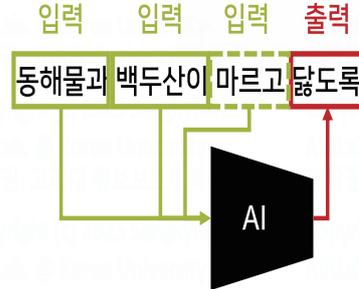
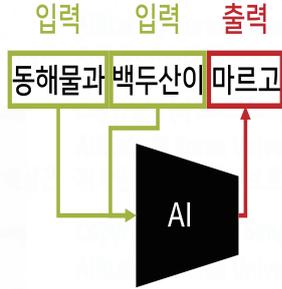


NLP AI: 추론

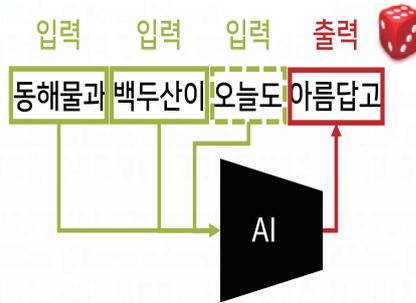
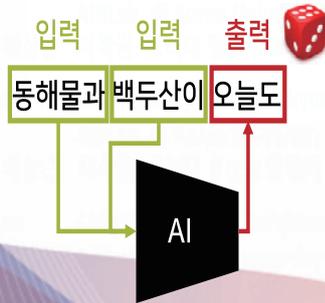


NLP AI: 출력 피드백

정답을 생성할 경우:

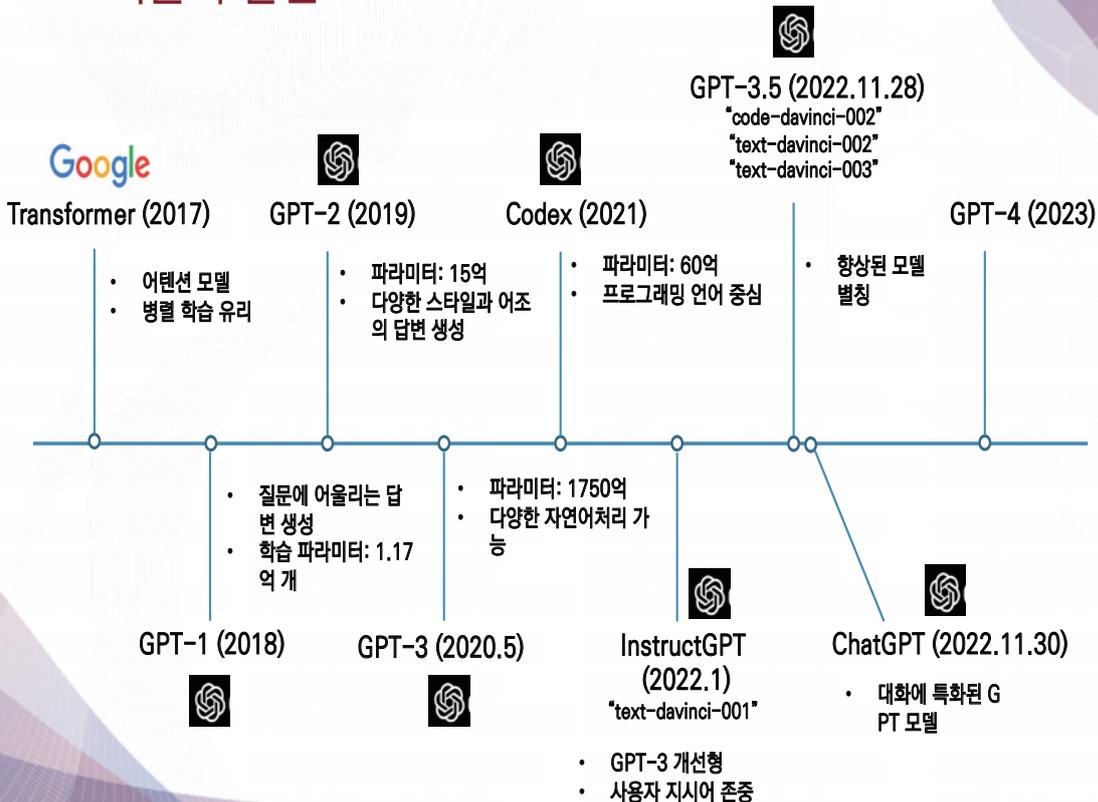


확률적으로 답을 생성할 경우:



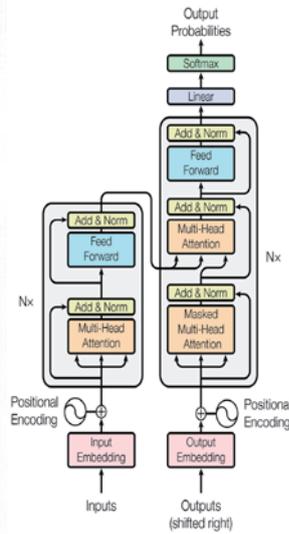
NLP AI 기술의 발전

※ OpenAI의 기술 중심으로 기재됨

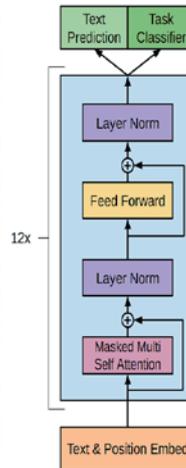


GPT의 심층 신경망 구조

Transformer 구조*



GPT (Generative Pre-trained Transformer) 구조**



어텐션 구조: 기존 NLP AI 모델에 비해 GPU를 이용한 병렬 학습에 유리

* Attention Is All You Need, NeurIPS (2017)

** Improving Language Understanding by Generative Pre-Training, OpenAI (2018, GPT-1)

Beginning of the Transformer

- Transformer [NeurIPS 2017]
- New SOTA on translation datasets at the time

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Brain
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

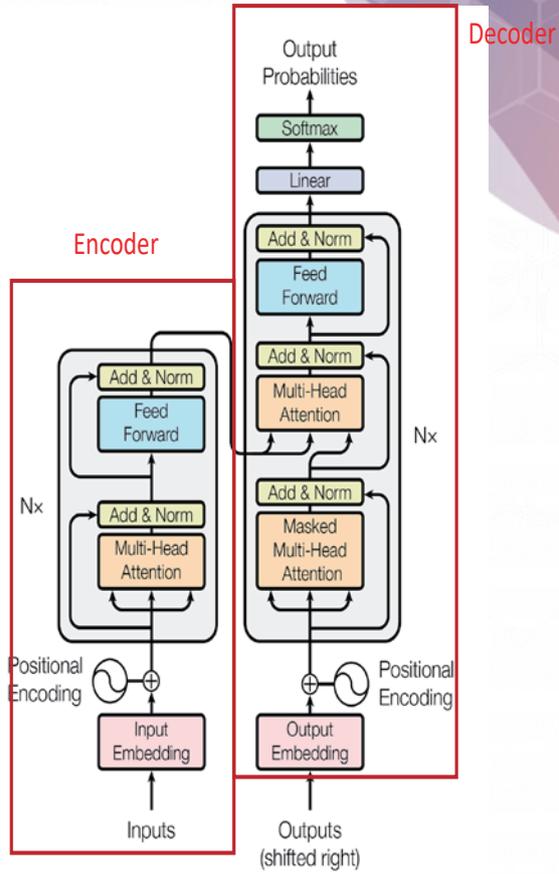
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

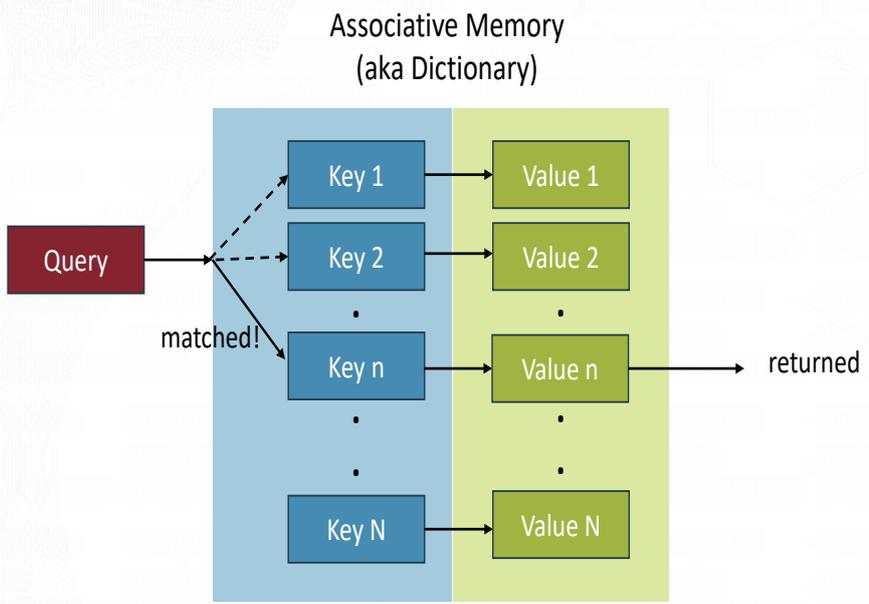
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-

Transformer Architecture

- Encoder-decoder architecture with only attention
 - No convolution or recurrence
- Encoder has:
 - (Masked) Self-attention
 - Positional encoding
 - Layer normalization



Query, Key, Value in Associative Memory



Query, Key, Value

- Use inputs and learning parameters to build query, key, and value:

$$q_i = W_q x_i$$

$$k_i = W_k x_i$$

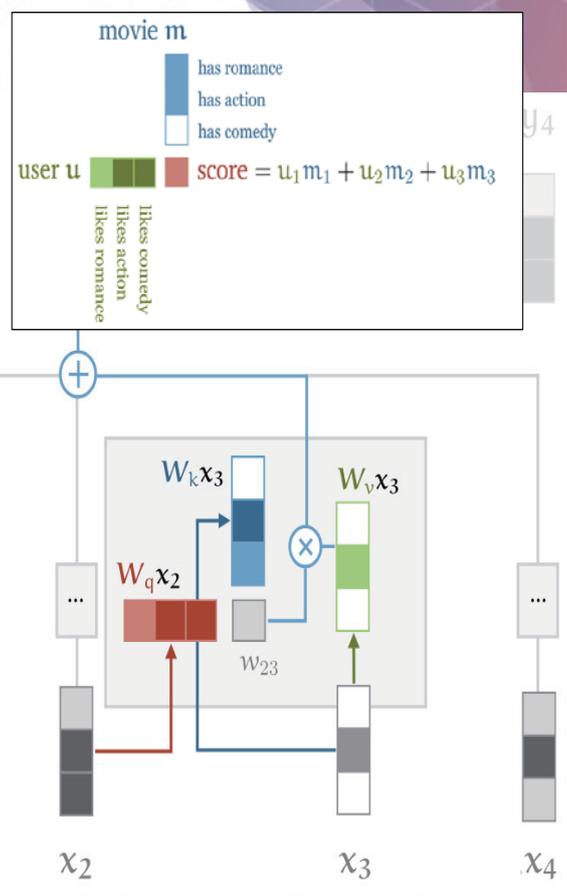
$$v_i = W_v x_i$$

Then, use them as follows:

$$w'_{ij} = q'_i k'_j$$

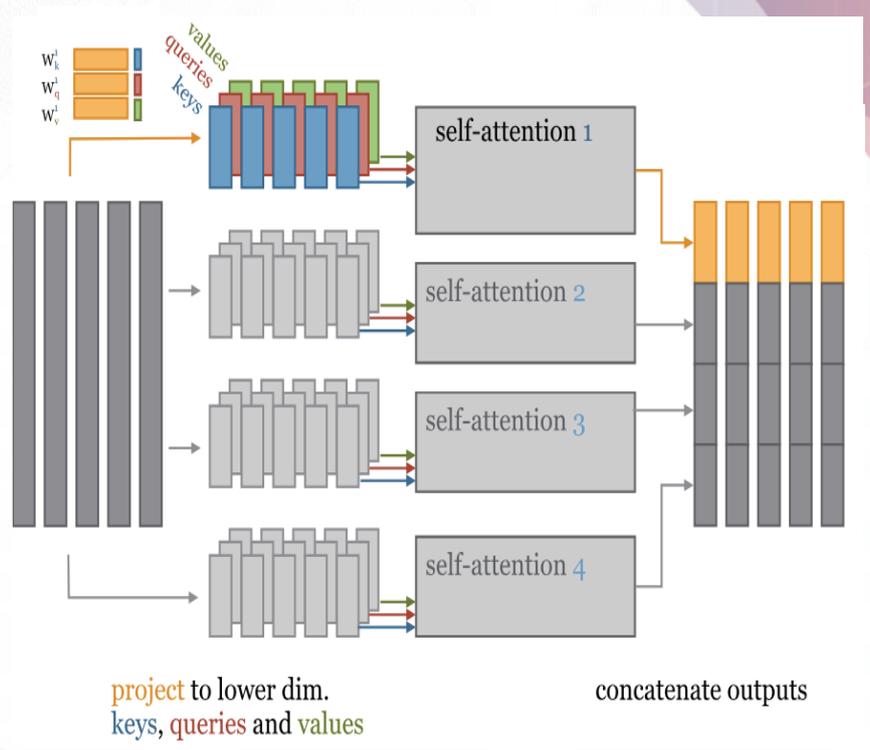
$$w_i = softmax(w'_i)$$

$$y_i = \sum_j w_{ij} v_j$$



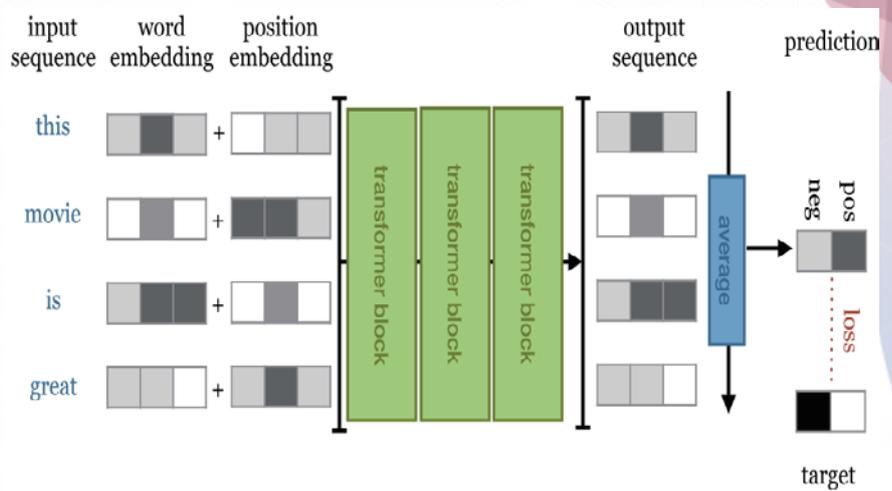
Multi-head attention

- Multiple "heads" = learning multiple sets of matrices W_q, W_k, W_v
- Similar motive to the use of multiple conv. filters in CNNs
- Leads to better prediction performance



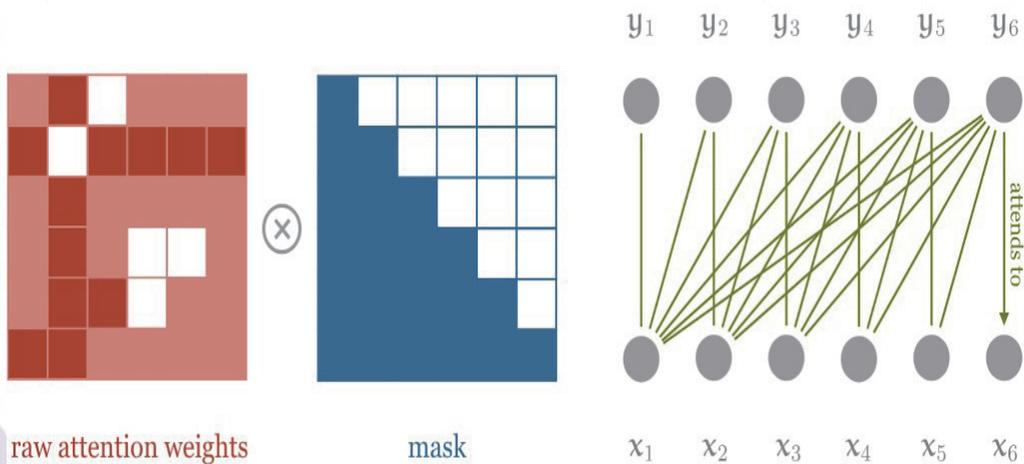
Position Encoding

- Solution: use an extra vector to provide positional information
- Position embedding
 - Learn positional vectors
 - Cons: need to see all lengths in training
- Position encoding
 - Use some function $f: \mathbb{N} \rightarrow \mathbb{R}^d$



Masking

- We need to mask future input tokens
- Otherwise, self-attention will see every token including the future

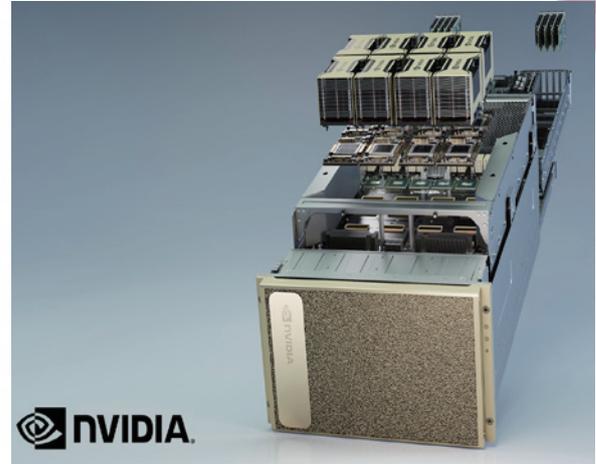


ChatGPT의 비용

- Fine-tuning을 위한 데이터 생성 비용
 - 사람에 의한 질문 또는 응답 작성, AI 응답에 대한 선호도 평가 비용

- 대규모 병렬 학습
 - 계산량: NVIDIA A100 GPU 1만개, 1개월
 - A100 GPU: 10,000\$/개
(총 GPU 비용: 약 1400억원)
 - H100 GPU: 44,000\$/개, x9 faster

 - 1회 학습: 지구 ⇄ 달 왕복 (약 70만km) 차량 주행 만큼 CO₂ 배출



NVIDIA DGX (A100 GPU x8)

응용사례: 지식 검색 및 요약

2차 전지의 종류, 소재, 개발 연도, 장단점들 환경 오염을 막기 위해 정부와 개인이 할 수 있는 일들을 대조식으로 표로 작성해 줘

종류	소재
납-산화물 (Pb-acid)	양극: 납(Pb), 음극: 산화 납(PbO2)
니켈-카드뮴 (Ni-Cd)	양극: 수산화 된 니켈 (Ni(OH)2), 음극: 카드뮴 (Cd)
니켈-금속수소 (Ni-MH)	양극: 수산화 된 금속 (MH), 음극: 수산화 된 니켈(Ni(OH)2)
리튬 폴리머 (Li-polymer)	양극, 음극 모두 : 플라스틱 (polymer)
리튬 이온(Li-ion)	양극: 리튬 코발트산화물 (LiCoO2), 음극: 석탄 (graphite)

이래는 정부와 개인이 환경 오염을 막기 위해 할 수 있는 일들을 대조식으로 나타낸 표입니다.

	정부	개인
에너지	- 신재생 에너지 개발 및 보급 - 에너지 절약을 위한 법규 제정 및 시행 - 대중교통 및 자전거 이용 촉진	- 에너지 절약을 위한 습관 실천 - 에너지 효율 높은 가전제품 선택 및 사용 - 자동차 대신 대중교통 또는 자전거 이용
폐기물	- 폐기물 처리 시설 건립 및 운영 - 폐기물 처리 법규 제정 및 시행 - 폐기물 발생량 감축을 위한 캠페인 및 교육	- 분리수거 실천 - 일회용품 대신 지속 가능한 제품 선택 및 사용 - 쓰레기 줄이기 캠페인에 참여
수질	- 수질 오염원 제거 및 관리 - 수질 오염 사건 발생 시 대처 - 수질 오염 감시 및 조사	- 쓰레기를 바다나 강 등에 버리지 않음 - 불필요한 화학물질 사용 줄이기 - 수돗물 대신 정수기 사용
대기	- 대기 오염원 제거 및 관리 - 대기 오염 사건 발생 시 대처 - 대기 오염 감시 및 조사	- 대중교통 이용 또는 공유 자동차 이용 - 화학물질 사용 줄이기 - 실외 배배큐나 화물차 운전 등 대기 오염 활동 줄이기

위와 같이 정부와 개인이 각자 할 수 있는 일들이 있습니다. 정부는 대규모의 에너지 인프라 구축 및 법규 제정과 시행을 통해 대규모의 환경 문제에 대응할 수 있으며, 개인은 소규모의 노력을 통해 일상적인 환경 문제에 대응할 수 있습니다. 결국, 정부와 개인이 함께 노력하여 환경 오염을 막을 수 있을 것

응용사례: 자동 코딩

■ 코드 작성

Make c codes for a simple server that listens on the local port number 8080 and sends "Greetings!" whenever the client sends the characters "Hi!".

Sure! Here's an example C code for a simple server that listens on the local port number 8080 and responds with "Greetings!" when the client sends the characters "Hi!":

```

c
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <sys/socket.h>
#include <arpa/inet.h>
#include <unistd.h>

#define PORT 8080

int main(int argc, char const *argv[]) {
    int server_fd, new_socket, valread;
    struct sockaddr_in address;
    int addrlen = sizeof(address);
    char buffer[1024] = {0};
    const char* greeting = "Greetings!\n";

    // Creating socket file descriptor
    if ((server_fd = socket(AF_INET, SOCK_STREAM, 0)) == 0) {
        perror("socket failed");
        exit(EXIT_FAILURE);
    }

    // Forcefully attaching socket to the port 8080
    if (setsockopt(server_fd, SOL_SOCKET, SO_REUSEADDR | SO_REUSEPORT,
        &opt, sizeof(opt))) {
        perror("setsockopt");
        exit(EXIT_FAILURE);
    }
}

```

```

address.sin_family = AF_INET;
address.sin_addr.s_addr = INADDR_ANY;
address.sin_port = htons( PORT );

// Forcefully attaching socket to the port 8080
if (bind(server_fd, (struct sockaddr *)&address,
    sizeof(address))<0) {
    perror("bind failed");
    exit(EXIT_FAILURE);
}

if (listen(server_fd, 3) < 0) {
    perror("listen");
    exit(EXIT_FAILURE);
}

while (1) {
    printf("Waiting for incoming connection...\n");
    if ((new_socket = accept(server_fd, (struct sockaddr *)&address,
        (socklen_t*)&addrlen))<0) {
        perror("accept");
        exit(EXIT_FAILURE);
    }

    printf("New connection established!\n");

    while ((valread = read(new_socket, buffer, 1024)) > 0) {
        printf("Received message: %s\n", buffer);

        if (strcmp(buffer, "Hi!\n") == 0) {
            send(new_socket, greeting, strlen(greeting), 0);
        }

        memset(buffer, 0, sizeof(buffer));
    }
}

return 0;
}

```

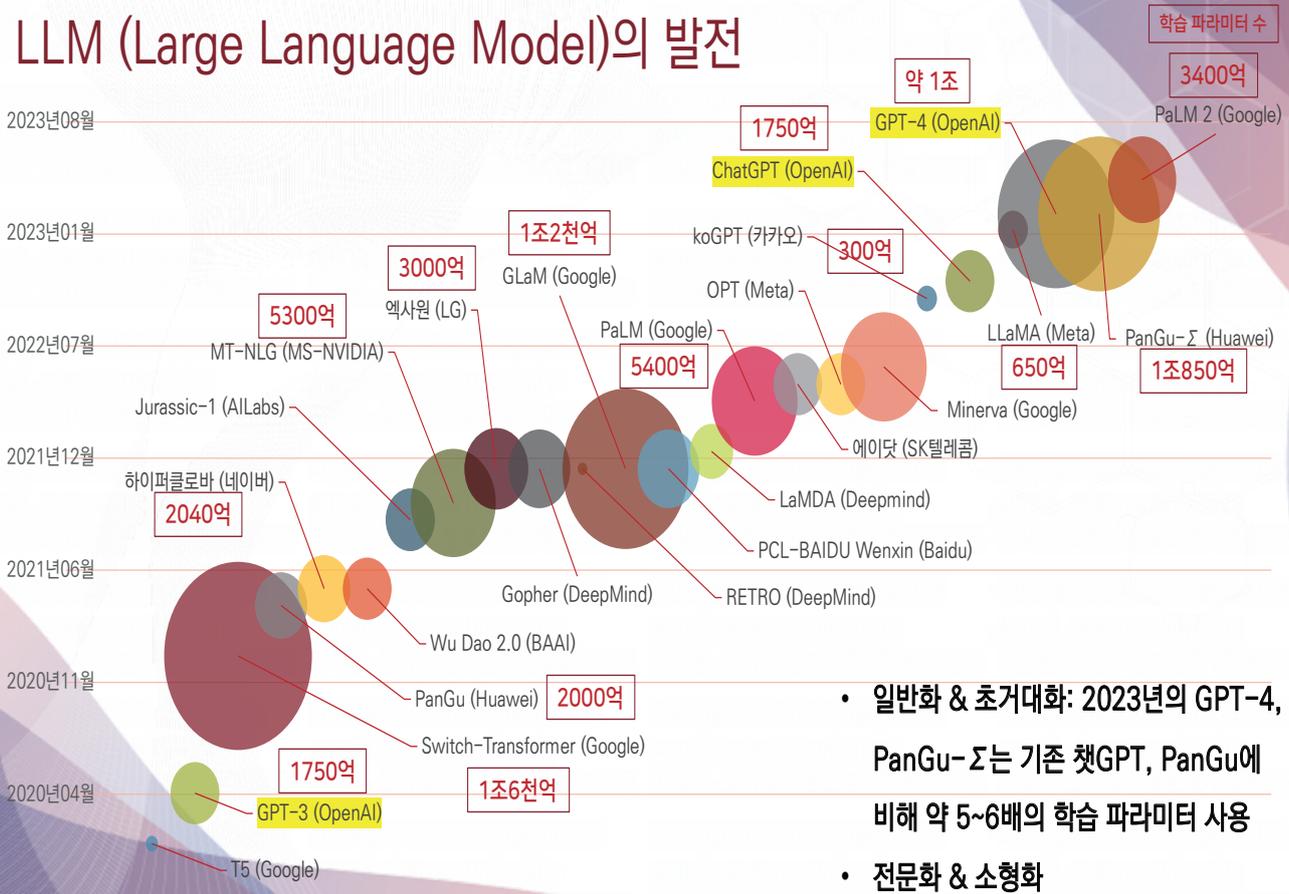
응용사례: 스토리 생성

스타워즈 1편을 10페이지 분량의 5세용 동화책으로 만들고 싶어. 각 페이지에 표시할 문장을 표로 만들어 줘. 등장인물의 이름은 쉬운 이름으로 새로 만들어 줘.

좋아요! 아래는 각 페이지에 표시될 문장을 표로 만든 것입니다. 등장인물의 이름은 쉬운 이름으로 새로 만들었습니다.

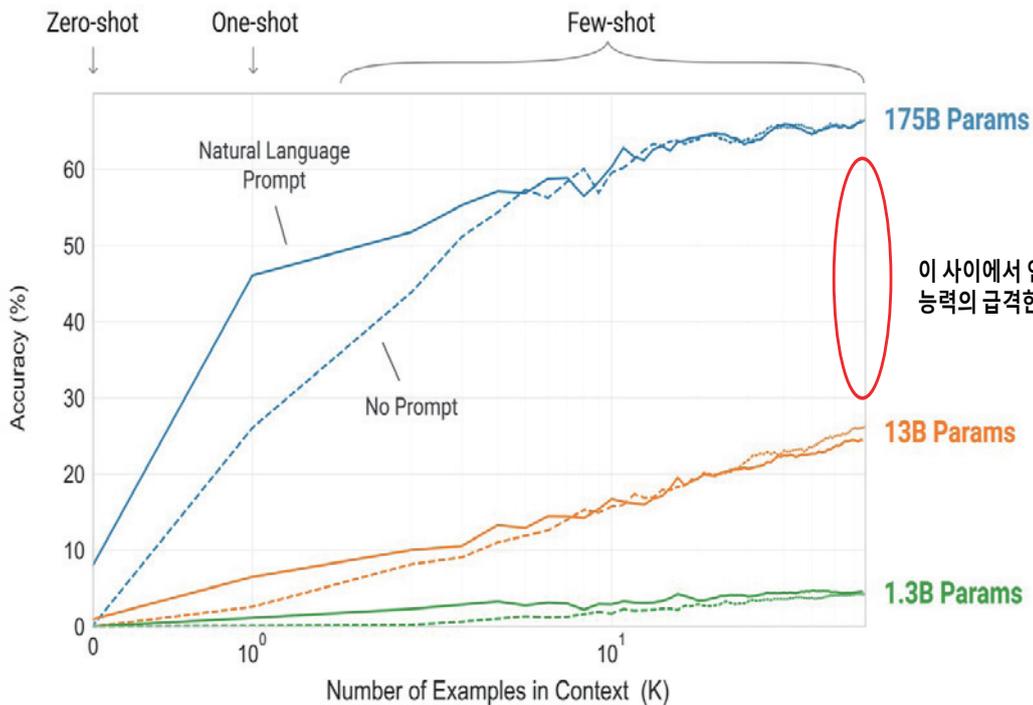
페이지	문장
1	한때, 멀리 떨어진 우주에는 왕국이 있었습니다.
2	하지만 왕국은 무자비한 침략을 받아 전쟁이 일어났습니다.
3	그러던 중 공주 벨라는 적에게 잡혔습니다.
4	하지만 벨라는 매우 용감하고 지혜로웠습니다.
5	그녀는 자신을 구해줄 사람을 찾기 위해 코리아 스타의 요원을 보냈습니다.
6	요원은 헨리와 제다이 기사 존이 나타났습니다.
7	헨리는 매우 똑똑하고 용감한 소년입니다.
8	그는 벨라를 구하기 위해 모든 것을 걸었습니다.
9	마침내 헨리와 존은 벨라를 구출했습니다.
10	그리고 모두 함께 미래를 향해 모험을 떠났습니다.

LLM (Large Language Model)의 발전



- 일반화 & 초거대화: 2023년의 GPT-4, PanGu- Σ 는 기존 챗GPT, PanGu에 비해 약 5~6배의 학습 파라미터 사용
- 전문화 & 소형화

LLM 모델 크기와 추론 능력



Brown et al., Language Models are Few-Shot Learners, NeurIPS 2020 (GPT-3)

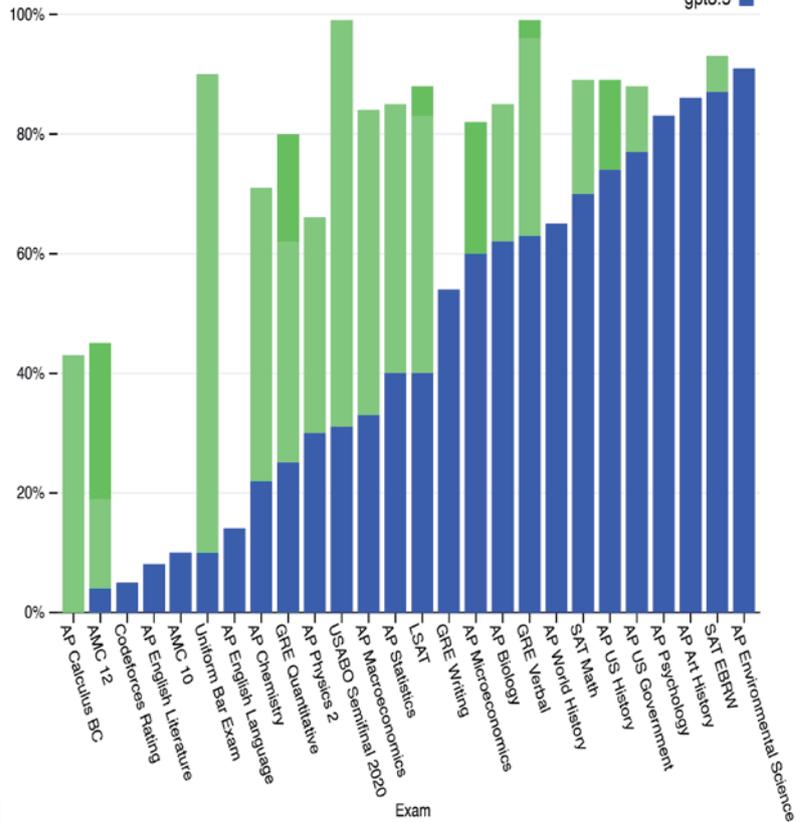
GPT3.5 vs 4

GPT-4

- 다양한 작업 및 언어에서 성능 향상

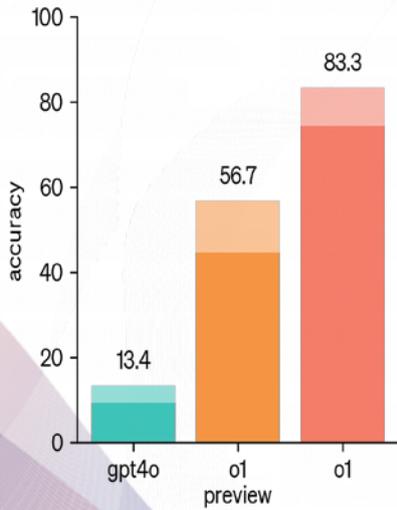
Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)

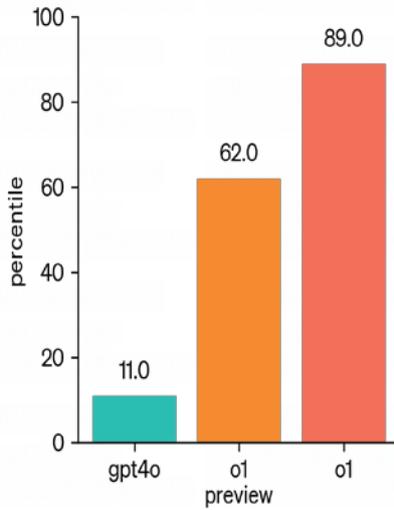


GPT o1 (2024.9.12)

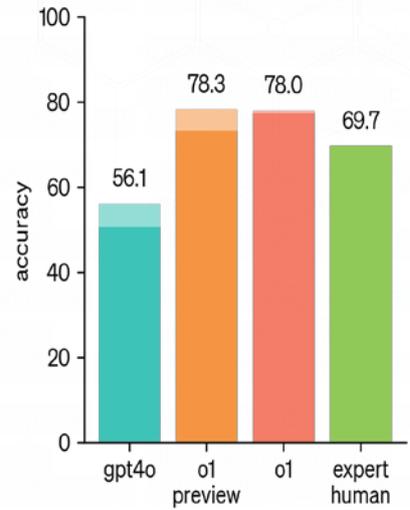
Competition Math (AIME 2024)



Competition Code (Codeforces)



PhD-Level Science Questions (GPQA Diamond)



챗GPT 기술 전망

- 전 세계 AI 시장 2026년도 900조원 규모 전망 (IDC)
- 거대언어 모델 (LLM)
 - 소형화, 전문화
 - 외부 DB를 통한 최신 자료 접근
 - REALM (Google), RETRO (DeepMind), LLM-Augmenter (MS), RA-CM3 (Meta)
 - 신뢰성 확보
 - Bing Copilot (MS): 근거 자료 제시
- 데이터베이스, 클라우드, 통신, 반도체 등 관련 기술 개발 필요

Data Privacy 이슈

- OpenAI Privacy Policy (June 23, 2023, <https://openai.com/policies/privacy-policy>)
- 수집되는 정보
 - 사용자 제공 개인 정보 (Account, User content (input, file upload, file upload, social media info))
 - 서비스 이용시 자동 수집되는 개인 정보 (웹 브라우저, usage (type of device, time zone, country, date/time), device 정보, cookie, analytics)
- 데이터 사용처
 - 서비스 제공, 관리, 유지/보수, 분석
 - 서비스 개선 및 연구
 - 새 서비스 개발
 - 사기나 범죄 행위 억제, 자사 IT 시스템/아키텍처/네트워크 보호,
 - 사업 이전
 - 프라이버시, 안전, 소유권 등 보호를 위한 법적 필요

As noted above, we may use Content you provide us to improve our Services, models that power ChatGPT. See [here](#) for instructions on how you can opt out to train our models.

User Content Opt Out Request

One of the most useful and promising features of AI models is that they can improve over time. We continuously improve the models that power our services, such as ChatGPT and DALL-E, via scientific and engineering breakthroughs as well as exposure to real world problems and data.

As part of this continuous improvement, when you use ChatGPT or DALL-E, we may use the data you provide us to improve our models. Not only does this help our models become more accurate and better at solving your specific problem, it also helps improve their general capabilities and safety.

We know that data privacy and security are critical for our customers. We take great care to use appropriate technical and process controls to secure your data. We remove any personally identifiable information from data we intend to use to improve model performance.

We understand that in some cases you may not want your data used to improve model performance. You can opt out of having your data used to improve our models by filling out this form. Please note that in some cases this will limit the ability of our models to better address your specific use case.

For details on our data policy, please see our [Privacy Policy](#) and [Terms of Use](#) documents.

*Please ensure the email you provide is associated with your account, and that the Organization ID is of the format "org-xKam3pleOr9gID" otherwise we will not be able to process your request.

sanglee77@gmail.com [Switch account](#)

* Indicates required question

Email *

Your email

Organization ID (found on [Account Org Settings](#)) *

Your answer

Organization name (found on [Account Org Settings](#))

Your answer

A copy of your responses will be emailed to the address you provided.

Submit Clear form

사용 예시: 웹사이트 요약

<https://openai.com/policies/privacy-policy/> 를 읽고, chatgpt의 사용자로서 데이터 프라이버시 측면에서 주의해야 할 점을 개괄식으로 요약해 줘.



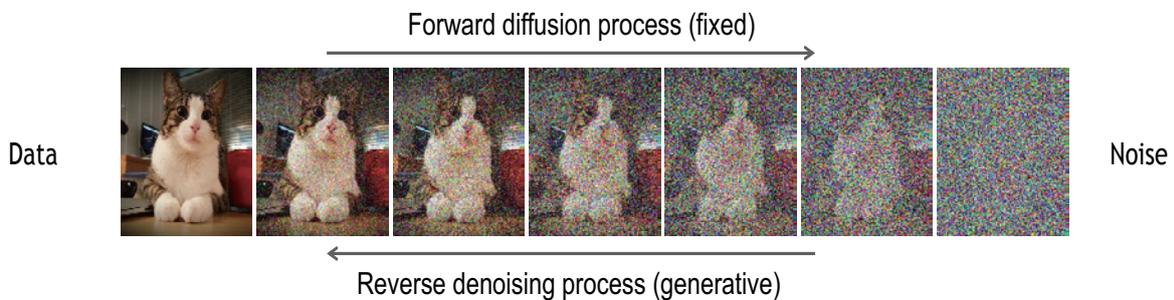
ChatGPT 사용자로서 데이터 프라이버시 측면에서 주의해야 할 점은 다음과 같습니다:

1. 개인 정보 수집: OpenAI는 사용자의 계정 정보, 대화 내용, IP 주소, 쿠키 등의 기술적 정보를 수집합니다.
2. 데이터 활용: 이 정보는 서비스 개선, 보안 강화, 법적 의무 이행 등을 위해 사용될 수 있습니다.
3. 제3자와의 데이터 공유: 특정 상황에서 제3자 서비스 제공자, 관련 기업, 또는 법적 기관과 정보를 공유할 수 있습니다.
4. 옵트아웃 옵션: 사용자는 데이터가 모델 학습에 사용되는 것을 거부할 수 있습니다.

더 자세한 사항은 [OpenAI의 프라이버시 정책](#)을 참고하세요.

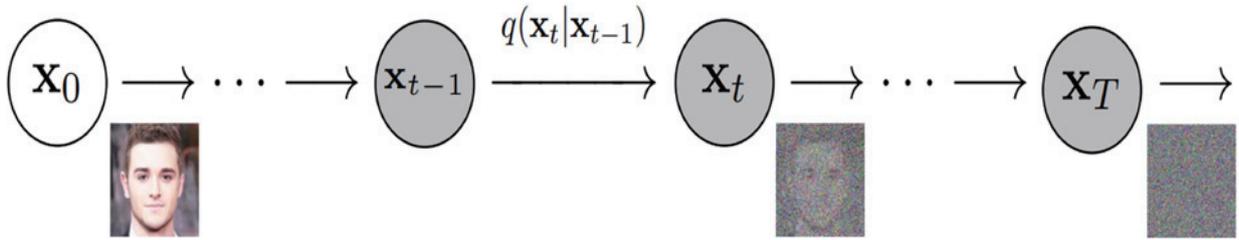
이미지 생성형 AI: Denoising Diffusion Models

- 두 가지 프로세스로 구성
 - Forward process: 원본 입력에 점증적으로 노이즈를 추가하는 과정
 - Reverse process: 역으로, 노이즈를 점증적으로 제거하여 원본 이미지를 복원하는 과정



- Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015
- Ho et al., Denoising Diffusion Probabilistic Models (DDPM), NeurIPS 2020

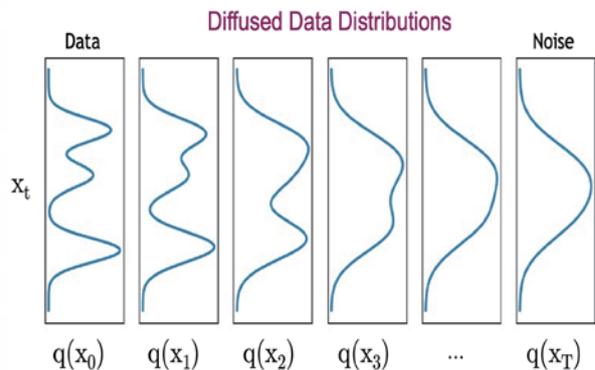
Forward Process



- $x_{(t-1)} \rightarrow x_t$ 의 조건부 확률을 가우시안 분포로 정의: $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$
- Noise variance schedule β_t : 추가 노이즈량 조정 파라미터 $\{\beta_t \in (0, 1)\}_{t=1}^T$
- Markov chain: $x_{(t-1)}$ 가 주어진 경우, x_t 의 생성은 $(t - 1)$ 이전에 영향 받지 않음 $q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})$

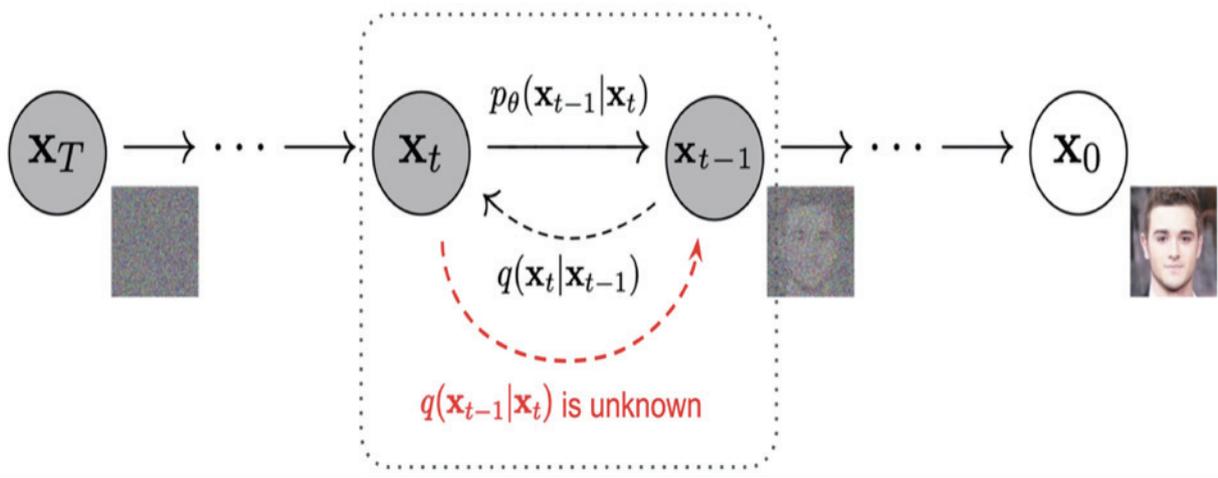
Forward Process로 인한 분포의 변화

$$q(\mathbf{x}_t) = \int \underbrace{q(\mathbf{x}_0, \mathbf{x}_t)}_{\text{Joint dist.}} d\mathbf{x}_0 = \int \underbrace{q(\mathbf{x}_0)}_{\text{Input data dist.}} \underbrace{q(\mathbf{x}_t | \mathbf{x}_0)}_{\text{Diffusion kernel}} d\mathbf{x}_0$$



- $q(x_t)$ 는 $q(x_0)$ 에 가한 Gaussian convolution의 결과로 생각할 수 있음

Reverse Process



- 디퓨전 모델의 목표는 $q(x_{t-1}|x_t)$ 를 학습하여, 완전한 노이즈로부터 점증적으로 노이즈를 제거하여 원본 이미지를 얻는데 있음

Sampling from the reverse process

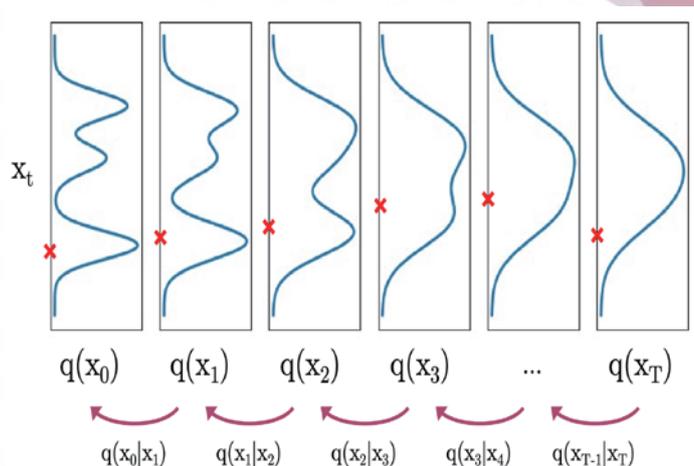
- Sample $x_T \sim \mathcal{N}(0, \mathbf{I})$
- Iteratively sample $x_{t-1} \sim q(x_{t-1}|x_t)$

true denoising dist.

$$q(x_t) = \int q(x_t|x_{t-1})q(x_{t-1})dx_{t-1}$$

intractable to compute

$$q(x_{t-1} | x_t) = q(x_t | x_{t-1}) \frac{q(x_{t-1})}{q(x_t)}$$



- β_t 가 충분히 작은 경우, $q(x_{t-1}|x_t)$ 를 Normal 분포로 approximate 할 수 있음

Training & Sampling

Algorithm 1 Training

- 1: repeat
- 2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
- 3: $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 5: Take gradient descent step on

$$\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2$$
- 6: until converged

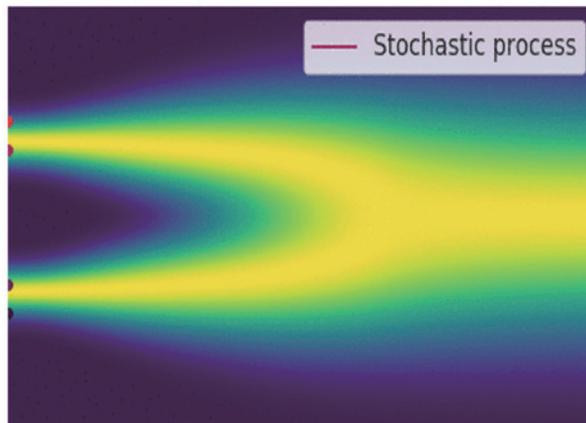
L_{t-1}

minimize L_0, L_1, \dots, L_{T-1}

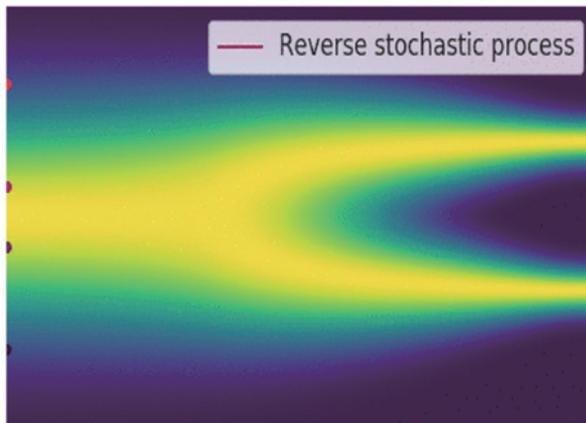
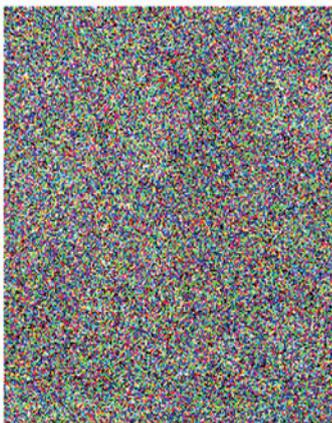
Algorithm 2 Sampling

- 1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ full random image
- 2: for $t = T, \dots, 1$ do reverse steps
- 3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
- 4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
- 5: end for subtract noise estimate from the current x_t
- 6: return \mathbf{x}_0

- Add randomness since reverse steps follow normal distribution.
- Do not add randomness when $t=1$ (when we create x_0)

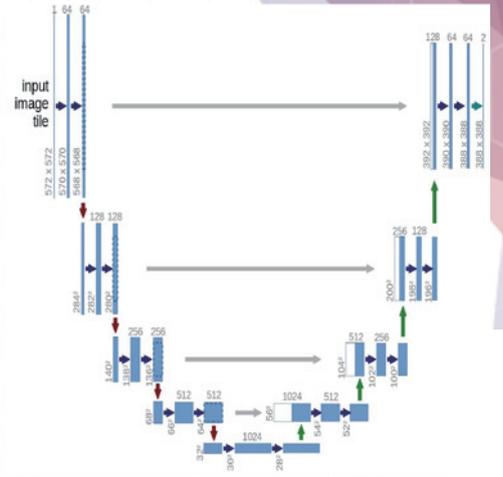
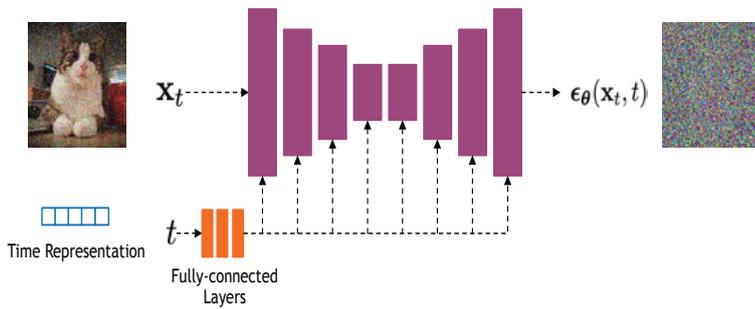


Forward



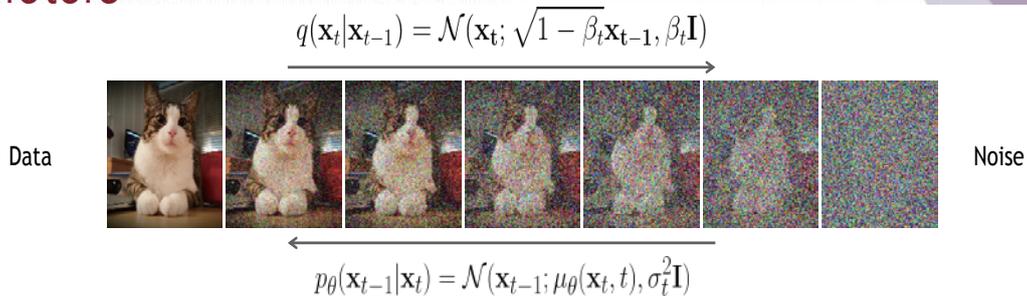
Reverse

Neural Architecture for $\epsilon_{\theta}(x_t, t)$



- 많은 diffusion model이 U-Net 구조를 사용
 - U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger, Fischer, and Brox, Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241, 2015
- 시간 값은 transformer와 유사한 positional embedding 방식으로 입력
- Self-attention, residual block 등 다양한 기법을 적용

Parameters



- Linear schedule (DDPM): increase $\beta_1 = 10^{-4} \rightarrow \beta_T = 0.02$, set $\sigma_t^2 = \beta_t$
- 하지만, 너무 빠르게 노이즈가 유입되어 유의미한 학습이 어려운 경향을 보임
- Cosine schedule [Nichol & Dhariwal, ICML 2021]

Cosine
Linear



Classifier Guidance

Algorithm 1 Classifier guided diffusion sampling, given a diffusion model $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, and gradient scale s .

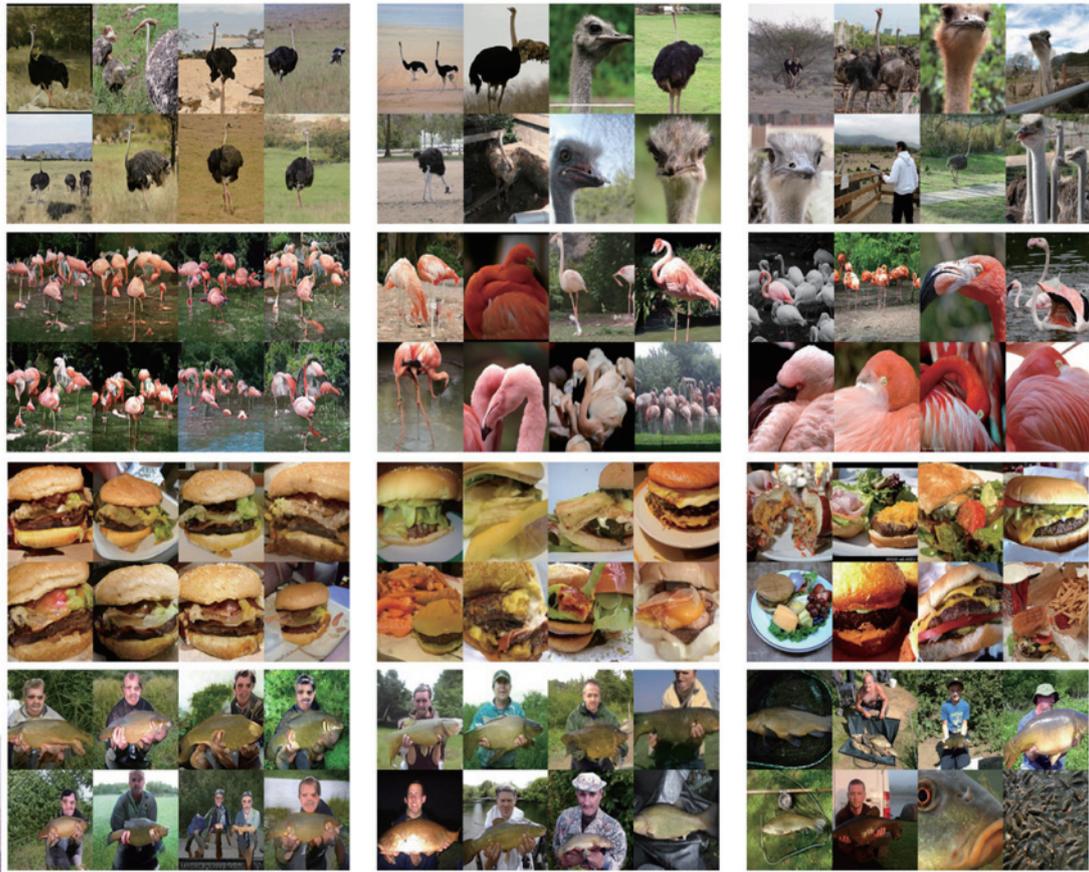
```
Input: class label  $y$ , gradient scale  $s$ 
 $x_T \leftarrow$  sample from  $\mathcal{N}(0, \mathbf{I})$ 
for all  $t$  from  $T$  to 1 do
     $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$ 
     $x_{t-1} \leftarrow$  sample from  $\mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$ 
end for
return  $x_0$ 
```

- Dhariwal and Nichol, "Diffusion models beat GANs on image synthesis", NeurIPS 2021
- Pretrain된 별도의 classifier $p_\phi(y|x)$ 의 그래디언트를 원하는 클래스의 특징을 갖는 이미지 생성을 위해 사용
- 해당 classifier는 noise가 포함된 입력에 대해서도 학습되어야 하므로, 기존 모델을 사용하기 어려움

Classifier Guidance



Figure 2: Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.



BigGAN

Diffusion (w/ classifier guidance)

Training Set

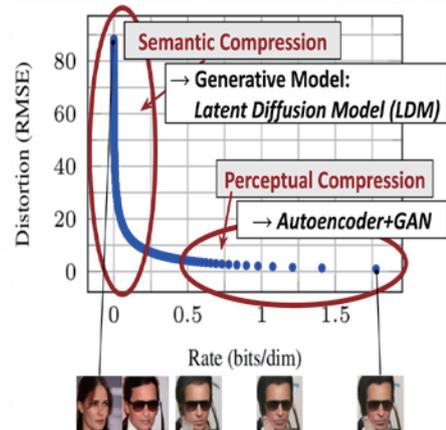
Latent Diffusion Model

- 앞에서 언급한 Diffusion model의 단점
 - Input space에서 diffusion 연산 → 고해상도로 갈수록 증가하는 계산 비용
 - GAN에 비해 높은 계산 비용

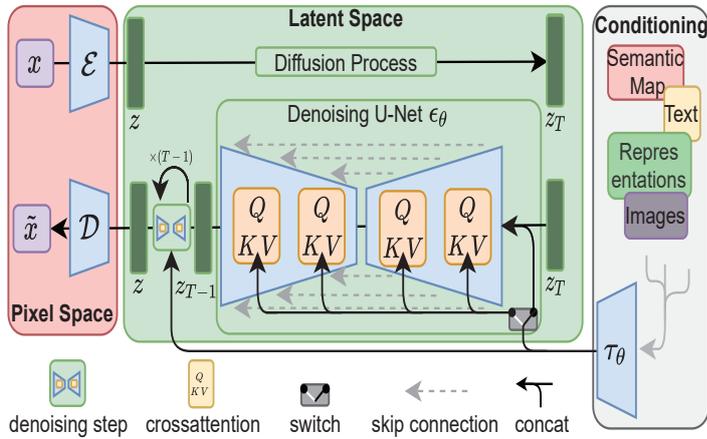
- Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
 - 의미(내용)을 재현하는 LDM과
 - 디테일을 재현하는 en/decoder로 문제를 분할

- Key idea: 저차원의 Latent space에서 diffusion 수행

- Basis of Stable Diffusion (2022-08)



Latent Diffusion Model



$$L_{LDM} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right]$$

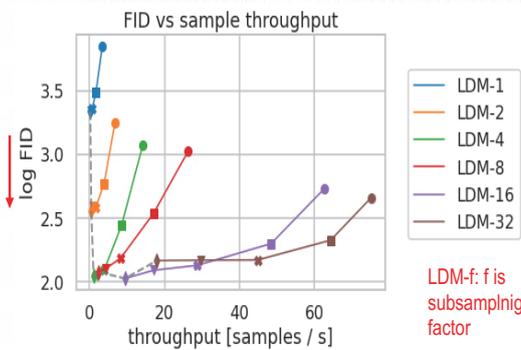
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2 \right]$$

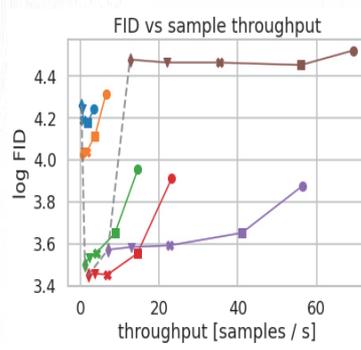
(conditional LDM)

- Latent 공간에서 diffusion forward/reverse 수행
- Attention이 추가된 형태의 U-Net 구조 사용
- 텍스트, 이미지 등의 입력을 받아 conditional LDM 학습 가능
- Encoder/Decoder를 사용하여 원래 이미지 공간에서/으로 변환

CelebA-HQ



ImageNet



Latent Diffusion Model

FID (Frechet inception distance)

- 생성 이미지와 실제 이미지의 분포를 비교 (작을수록 좋음)
- Martin et al., NeurIPS, 2017



Figure 4. Samples from LDMs trained on CelebAHQ [38], FFHQ [40], LSUN-Churches [98], LSUN-Bedrooms [98] and class-conditional ImageNet [12], each with a resolution of 256×256 . Best viewed when zoomed in. For more samples cf. the supplement.

Latent Diffusion Model

Layout-to-image



Text-to-image

*"A street sign that reads
'Latent Diffusion'"*



*"An oil painting
of a space shuttle"*

OpenAI: DALL-E2 (Text-to-Image)

<https://openai.com/dall-e-2/>



Teddy bears mixing sparkling chemicals as mad scientists



An astronaut riding a horse in a photorealistic style



A bowl of soup as a planet in the universe

Google: Imagen (Text-to-Image)

<https://imagen.research.google/>



A cute corgi lives in a house made of sushi



A majestic oil painting of a raccoon Queen wearing red French royal gown.



A robot couple fine-dining with the Eiffel Tower in the background

DALL·E 2

“a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese”



IMAGEN

“A photo of a raccoon wearing an astronaut helmet, looking out of the window at night.”



DALLE-E2 (OpenAI): Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv 2022.

IMAGEN (Google): Saharia et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, arXiv 2022.

Meta AI: Make-A-Video (Text-to-Video)



An artist's brush painting on a canvas close up



A young couple walking in heavy rain



Horse drinking water

https://makeavideo.studio/?fbclid=IwAR30C-wvfvMfaLH-5uWN3i39Settifi7SCu_L-Kffmm7iI9F4LFGHUSA

Google: Imagen Video (Text-to-Video)



SORA (OpenAI)



Copyright (c) 2025 Sangkyun Lee

57

이미지 생성형 모델

- Diffusion 모델: stochastic process에 기반한 이미지 생성 AI 모델
- 많은 서비스 등장
 - Stable Diffusion, DALL-E2, Imagen, Midjourney, Leonardo.Ai, Bing, Adobe Firefly, InstantArt, NightCafe, Lexica, Playground AI, OpenAI Sora
- 생성 비용 절감, 이미지 퀄리티 향상, 쉬운 프롬프팅, 다양한 입력 처리 등 다양한 측면에서 발전될 것으로 전망됨

Copyright (c) 2025 Sangkyun Lee

58

정보 진위 판별 이슈



펜타곤 폭발 가짜 뉴스와 S&P 500 지수 -30p 변동
(May 22, www.livemint.com)

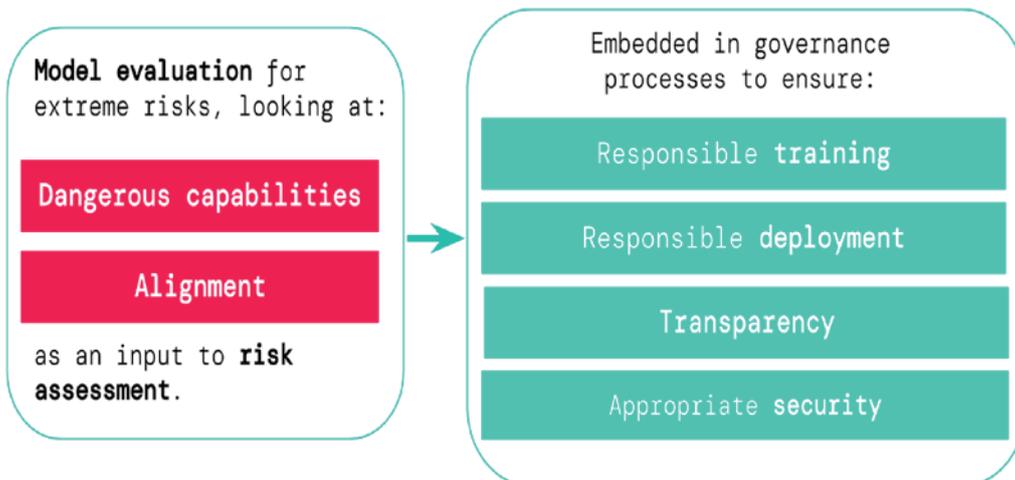
미 정치인 발언 영상, (CBS News, June 4)

Copyright (c) 2025 Sangkyun Lee

59

초거대 AI의 잠재 위험 평가 필요성

- AI의 잠재적 기능이 인류의 가치 기준에 부합하는지 평가 필요성 강조
- 초거대 AI가 가질 수 있는 잠재적 위험 기능: 새로운 사이버 공격 제작, 기만, 설득 및 조종, 정치 선동, 무기 탈취 및 제조, 장기 계획, AI에 의한 AI 제작, 상황 인식, 자기 복제



Model evaluations for extreme risks, DeepMind et al., 2023.5

Copyright (c) 2025 Sangkyun Lee

60

Safe, Secure and Trustworthy AI

바이든 행정명령 (10.30, 2023)



OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM • PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

My Administration places the highest urgency on governing the development and use of AI safely and responsibly, and is therefore advancing a coordinated, Federal Government-wide approach to doing so. The rapid speed at which AI capabilities are advancing compels the United States to lead in this moment for the sake of our security, economy, and society.

In the end, AI reflects the principles of the people who build it, the people who use it, and the data upon which it is built. I firmly believe that the power of our ideals; the foundations of our society; and the creativity, diversity, and decency of our people are the reasons that America thrived in past eras of rapid change. They are the reasons we will succeed again in this moment. We are more than capable of harnessing AI for justice, security, and opportunity for all.



President Joe Biden signs an executive on artificial intelligence in the East Room of the White House, Monday, Oct. 30, 2023, in Washington. Vice President Kamala Harris looks on at right. (AP Photo/Evan Vucci, File)

감사합니다