

LREC-COLING 2024

# Release of Pre-Trained Models for the Japanese Language

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui,  
Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, Koh Mitsuda

Kentaro Mitsui  
rinna Co., Ltd.



## AI Democratization

- Aims to create a world where everyone can easily use AI
- Open source **codes, databases, pretrained models**, etc.
- Non-English-speaking regions lags behind English-speaking regions

## Pre-trained models

- **+ Self-supervised learning** enables to utilize massive data
- **+ Transformer architecture** facilitates efficient training
- **- Significant computational resources** are required

## Our contribution

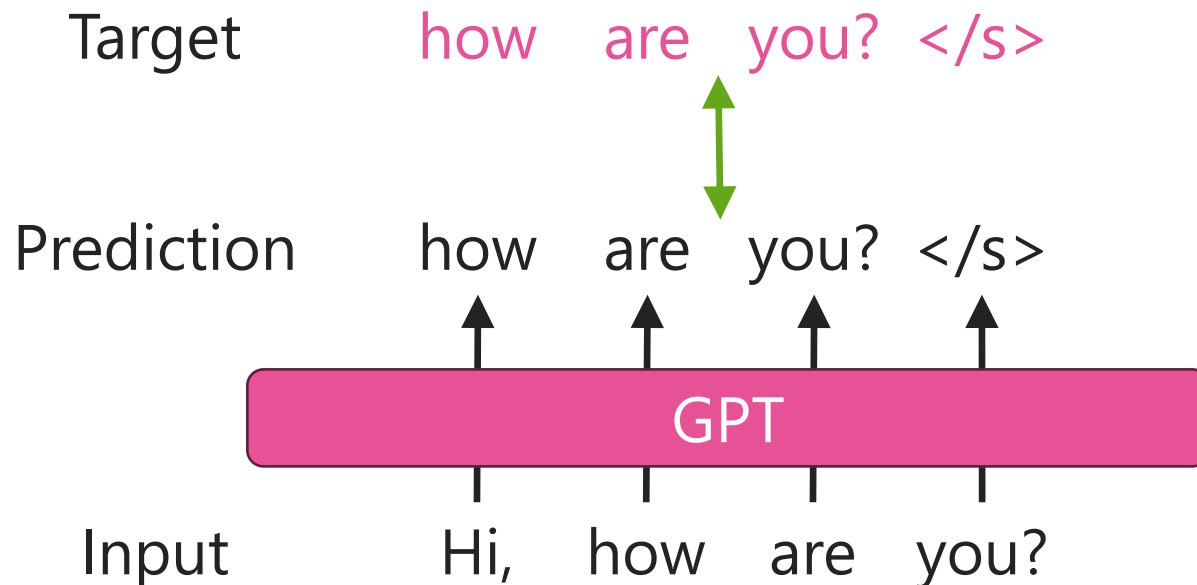
- Released **language, image, and speech models for Japanese**
- High performance and culturally-aware models

	Pre-trained model	Size	License	Date
<b>Language model</b>	rinna/japanese-gpt2-xsmall	37M	MIT	Aug. 2021
	rinna/japanese-gpt2-small	110M	MIT	Aug. 2021
	rinna/japanese-gpt2-medium	336M	MIT	Apr. 2021
	rinna/japanese-gpt-1b	1.3B	MIT	Jan. 2022
	rinna/japanese-gpt-neox-small	110M	MIT	Sep. 2022
	rinna/japanese-gpt-neox-3.6	3.6B	MIT	May 2023
	rinna/bilingual-gpt-neox-4b	4B	MIT	July 2023
<b>Language-image model</b>	rinna/japanese-clip-vit-b-16	197M	Apache 2.0	May 2022
	rinna/japanese-cloob-vit-b-16	197M	Apache 2.0	May 2022
	rinna/japanese-stable-diffusion	1.1B	CreativeML OpenRAIL-M	Sep. 2022
<b>Speech model</b>	rinna/japanese-hubert-base	95M	Apache 2.0	Apr. 2023

# Language Models: GPT

## Generative Pretrained Transformer (GPT)\*1

- Autoregressive language model
- Self-supervised learning through next token prediction
  - ◆ **High scalability:** can utilize massive data without annotation
  - ◆ **High training efficiency:** does not require sequential operations like RNN



Negative Log Likelihood  
 $\mathcal{L}_{\text{NLL}} = -\log p(\mathbf{x})$

\*1Radford et al., Improving Language Understanding by Generative Pre-Training, 2018

# Language Models: GPT

## Our models

- Various model sizes (37M – 4B) for different use cases
- GPT-NeoX: Rotary embeddings<sup>\*2</sup> for out-of-context positions
- Instruction-following<sup>\*3</sup> variants
  - ◆ **Supervised Fine-Tuning (SFT)**
  - ◆ **Reinforcement Learning from Human Feedback (RLHF) via Proximal Policy Optimization (PPO)**

## Data

- Wikipedia, CC-100, mC4 (ja), Pile, Redpajama (en, for bilingual model)
- Anthropic HH, SHP, FLAN (translated into ja, for SFT and PPO)

<sup>\*2</sup>Su et al., Roformer: Enhanced transformer with rotary position embedding, 2021

<sup>\*3</sup>Ouyang et al., Training language models to follow instructions with human feedback, 2022

# Language Models: GPT

## Experiments

- Benchmark: JP Language Model Evaluation Harness

- ◆ Average of 8 tasks

- JCommonsenseQA
    - JNLI
    - MARC-ja
    - JSQuAD
    - Xwinograd
    - JAQKET v2
    - XLSum-ja
    - MGSM

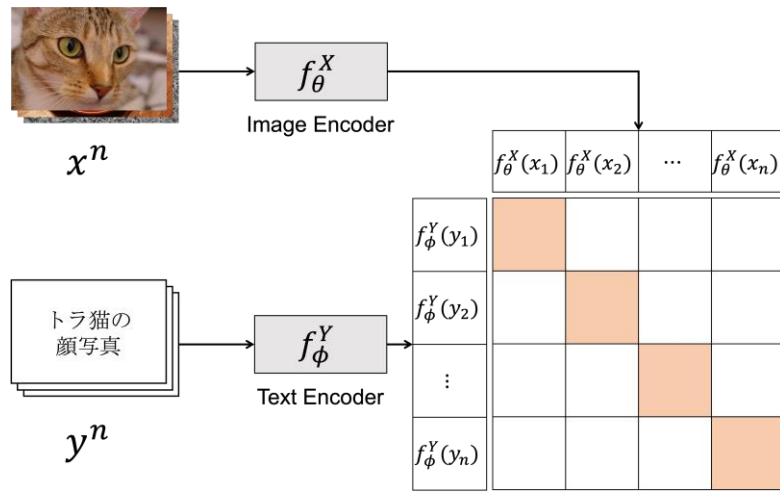
- Better performance w/ fewer parameters

Pre-trained model	Score
rinna/japanese-gpt2-xsmall	26.63
rinna/japanese-gpt2-small	27.33
rinna/japanese-gpt2-medium	28.33
rinna/japanese-gpt-1b	32.21
rinna/japanese-gpt-neox-small	30.11
rinna/japanese-gpt-neox-3.6b	36.60
rinna/bilingual-gpt-neox-4b	38.29
meta/llama-7b	33.28
meta/llama2-7b	42.97
meta/llama2-7b-chat	41.31
rinna/japanese-gpt-neox-3.6-sft	45.24
rinna/japanese-gpt-neox-3.6-ppo	46.37
rinna/bilingual-gpt-neox-4b-sft	47.65
rinna/bilingual-gpt-neox-4b-ppo	47.33

# Language-Image Models: CLIP

## Contrastive Language-Image Pre-training (CLIP)\*4

- Connects visual concepts with text in the embedding space
- Locked-image Tuning (LiT)\*5
  - ◆ Pretrained AugReg Vision Transformer image encoder was fixed
  - ◆ Randomly initialized BERT text encoder was updated solely



## Contrastive Loss

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} (\mathcal{L}_{\text{text}} + L_{\text{image}})$$

$$L_{\text{text}} = -\frac{1}{n} \sum_{i=1}^n \log \frac{\exp(f_\theta^X(x_i)^T f_\phi^Y(y_i))}{\sum_j \exp(f_\theta^X(x_j)^T f_\phi^Y(y_i))}$$

$$L_{\text{image}} = -\frac{1}{n} \sum_{i=1}^n \log \frac{\exp(f_\theta^X(x_i)^T f_\phi^Y(y_i))}{\sum_j \exp(f_\theta^X(x_i)^T f_\phi^Y(y_j))}$$

\*4Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

\*5Zhai et al., LiT: Zero-Shot Transfer with Locked-image text Tuning, 2022

# Language-Image Models: CLIP

## Data

- Conceptual 12M (CC12M), captions translated into Japanese
- Data augmentation using Bootstrapping Language-Image Pre-training (BLIP)<sup>\*6</sup>

## Experiments

- Higher top-1 accuracy in **Japanese ImageNet zero-shot classification**

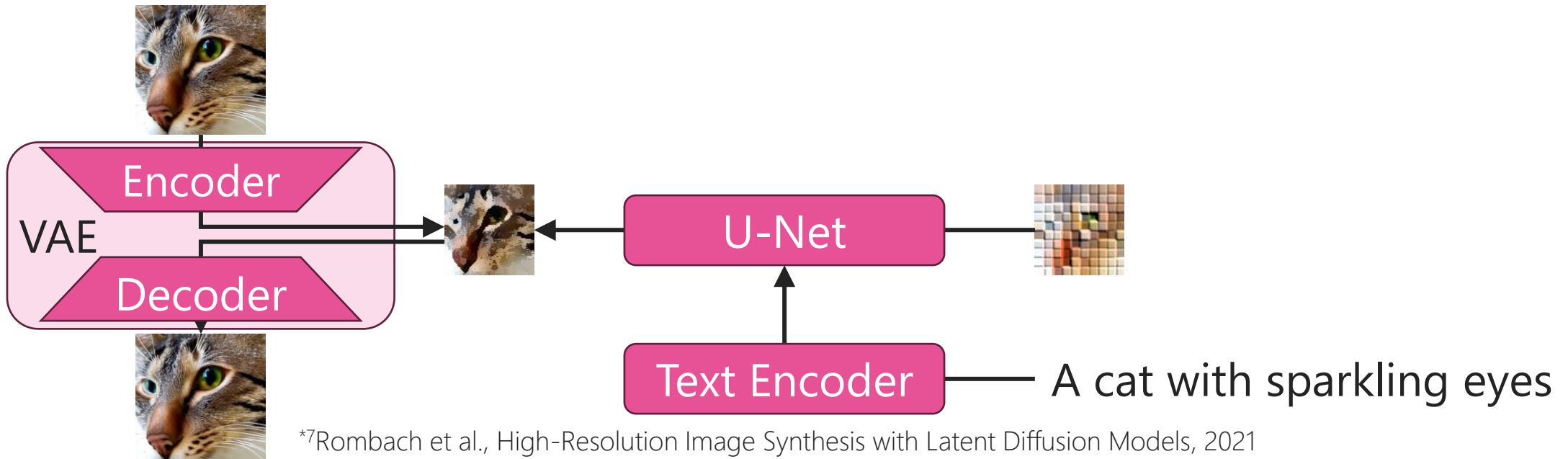
Pre-trained model	Score
laion/clip-base	38.00
laion/clip-large	53.09
rinna/japanese-clip-vit-b-16	50.69
rinna/japanese-cloob-vit-b-16	54.64

<sup>\*6</sup>Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022

# Language-Image Models: Stable Diffusion

## Stable Diffusion (SD)<sup>\*7</sup>

- Text-to-image model comprising VAE, text encoder, and U-Net
- Two-stage fine-tuning from compvis/stable-diffusion-v1-4
  - ◆ Stage 1: Text encoder was trained from scratch w/ other modules fixed
  - ◆ Stage 2: Text encoder and U-Net were jointly trained



<sup>\*7</sup>Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, 2021

# Language-Image Models: Stable Diffusion

## Data

- 100M images including Japanese subset of LAION-5B

## Experiments

- **Japanglish** (“salary man”) and **Japanese-specific culture** (“Ukiyo-e”) were well captured

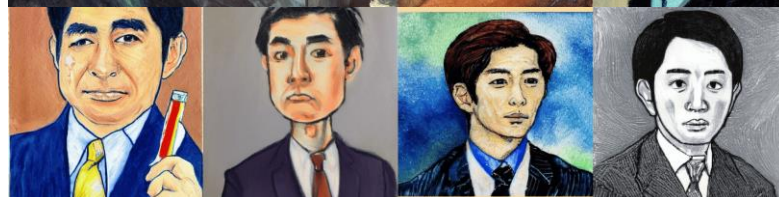
SD



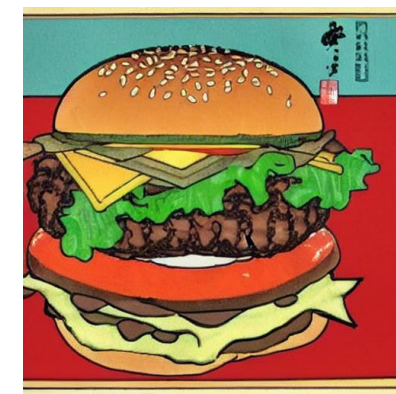
Ours (1<sup>st</sup> stage)



Ours (2<sup>nd</sup> stage)



Summer festival at  
shrine in evening,  
Watercolor painting

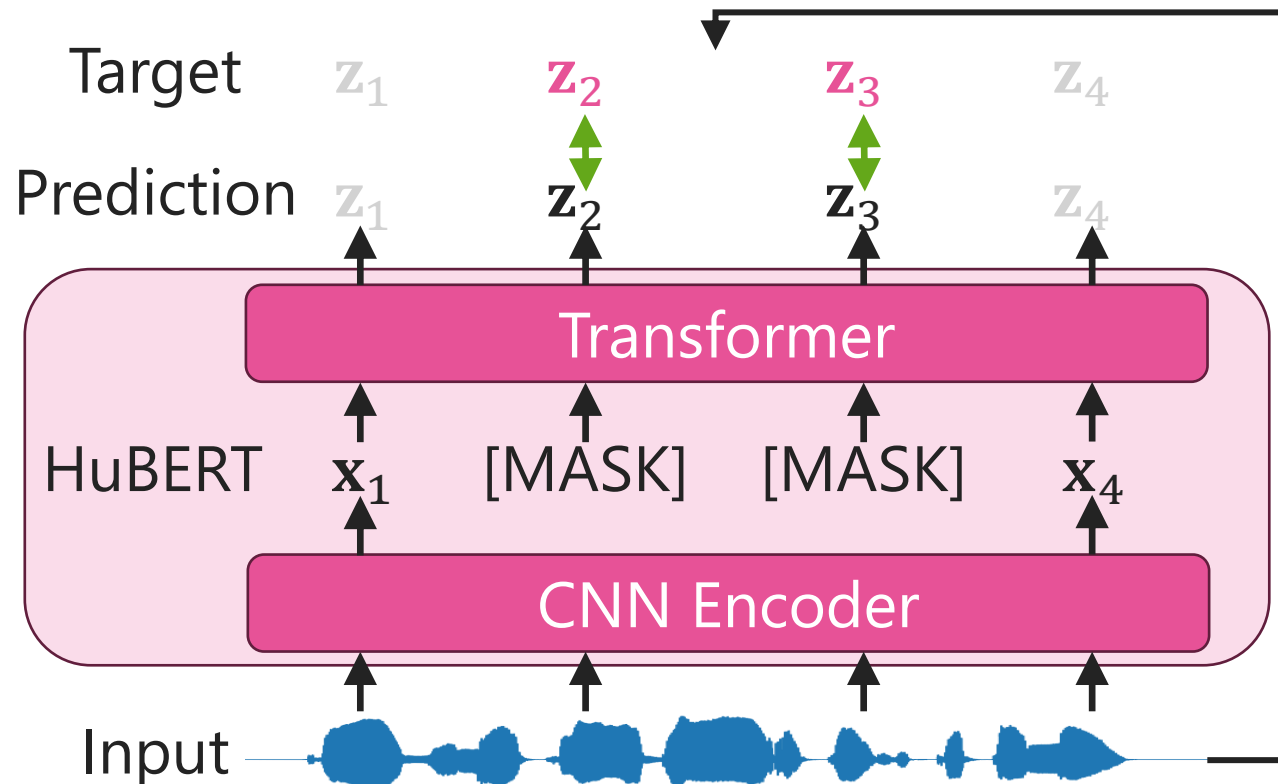


Hamburger,  
Ukiyo-e

# Speech Models: HuBERT

## Hidden-Unit BERT (HuBERT)\*<sup>8</sup>

- Self-supervised speech representation model (**no need for text**)
- Trained with **BERT-like masked language modeling objective**



Negative Log Likelihood

$$\mathcal{L}_{\text{NLL}} = - \sum_{t \in M} \log p(z_t | \tilde{\mathbf{X}}, t)$$

\* $M$ : set of masked indices

\* $\tilde{\mathbf{X}}$ :  $\mathbf{X}$  masked

K-means clustering on MFCC (1<sup>st</sup> stage) or 6<sup>th</sup> layer feature (2<sup>nd</sup> stage)

\*<sup>8</sup>Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, 2021

# Speech Models: HuBERT

## Data

- ReazonSpeech corpus v1 (19k hours of speech from TV programs)

## Experiments

- **Better WER in CTC-based ASR fine-tuning** on the CSJ corpus (552 hours of spontaneous speech)

Pre-trained model	Eval1	Eval2	Eval3
meta/hubert-base			
32-hour labeled	13.12	10.33	10.66
552-hour labeled	7.88	5.66	6.48
rinna/Japanese-hubert-base			
32-hour labeled	9.30	7.07	6.87
552-hour labeled	5.72	4.45	4.73

# Conclusions

- We have released (and will release) various pretrained models
  - ◆ Accessible at **Hugging Face** and **GitHub**
  - ◆ Released under **easy-to-use licenses**
  - ◆ Insights may be useful for other languages as well



<https://huggingface.co/rinna>



<https://github.com/rinnakk>