



TangoChat 1.0 (LLM & RAG)

성명 조현우

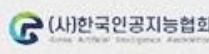
소속 한국전자통신연구원

SUBJECT

인공지능 기술의 대중화 (AI Democratization)를 위한
TANGO 커뮤니티 3회 컨퍼런스

주관 ETRI (한국전자통신연구원) 주최 과학기술정보통신부 IITP 정보통신기획평가원

후원



목 차

1

TANGO 프레임워크 버전 2

03

1. 패러다임 시프트
2. 프레임워크 구조의 변환

2

TangoChat 프로토타입

06

1. TASK
2. TARGET
3. AUTO NN
4. TANGO+CHAT
5. DOWNLOAD LLM
6. RUN
7. RAG
8. RETRIEVE URL
9. ANSWER

현재

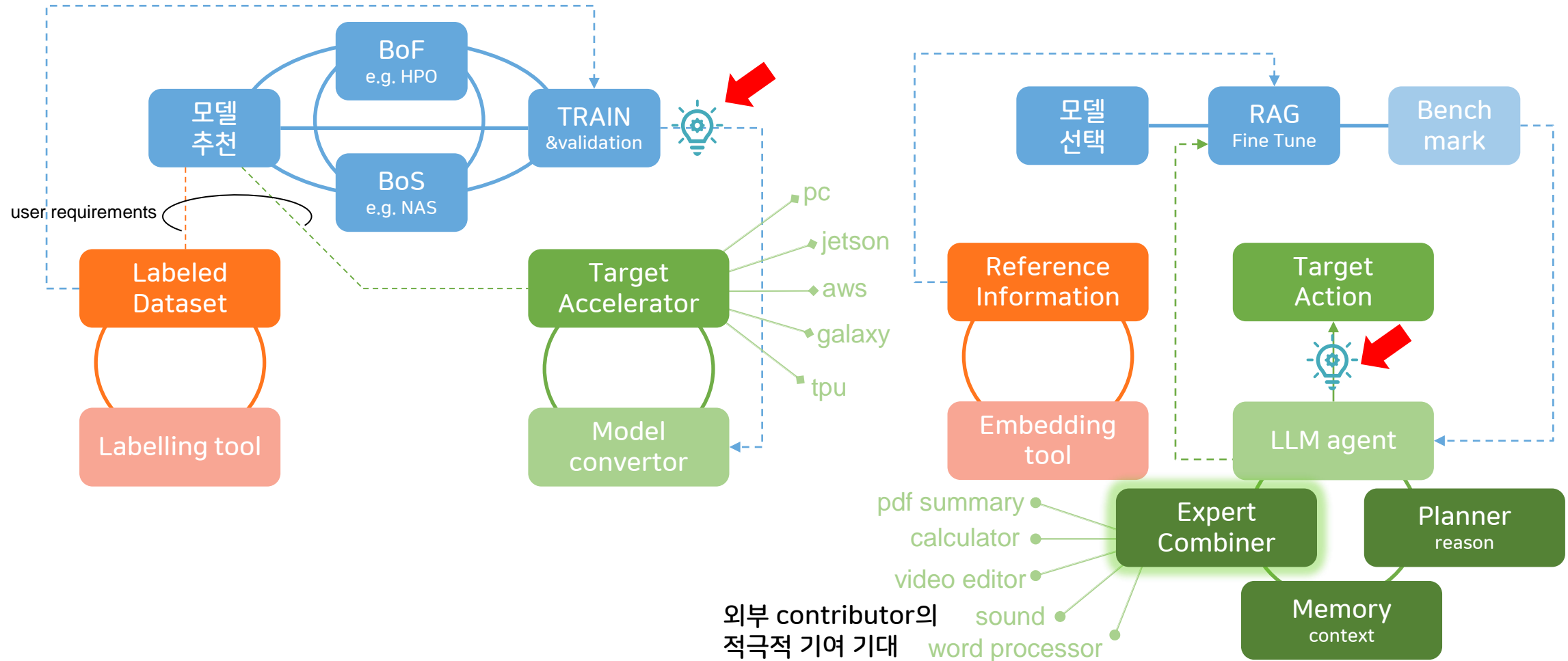
- 어떻게 신경망 모델을 학습할 것인가?
- 어떻게 타겟 디바이스의 가속기에 맞게 배포할 것인가?
- 얼마나 좋은 지 어떻게 객관적으로 평가할 것인가?
- 다음 번에 데이터와 클래스가 늘어나면 어떻게 할 것인가?
- 이 모든 파이프라인을 어떻게 자동화 할 것인가?
- 어떻게 비전문가가 이 도구를 쉽게 사용하게 할 것인가?

미래

- 어디에 신경망 모델을 적용할 것인가?
- 어떻게 실제 동작과 연계할 것인가?
- 내 적용분야에 잘 맞는지 어떻게 평가할 것인가?
- 다음 번에 다른 분야에 적용하려면 어떻게 할 것인가?
- 이 모든 파이프라인을 어떻게 체계화 할 것인가?
- 이 도구를 얼마나 다양한 분야에 적용하게 할 것인가?

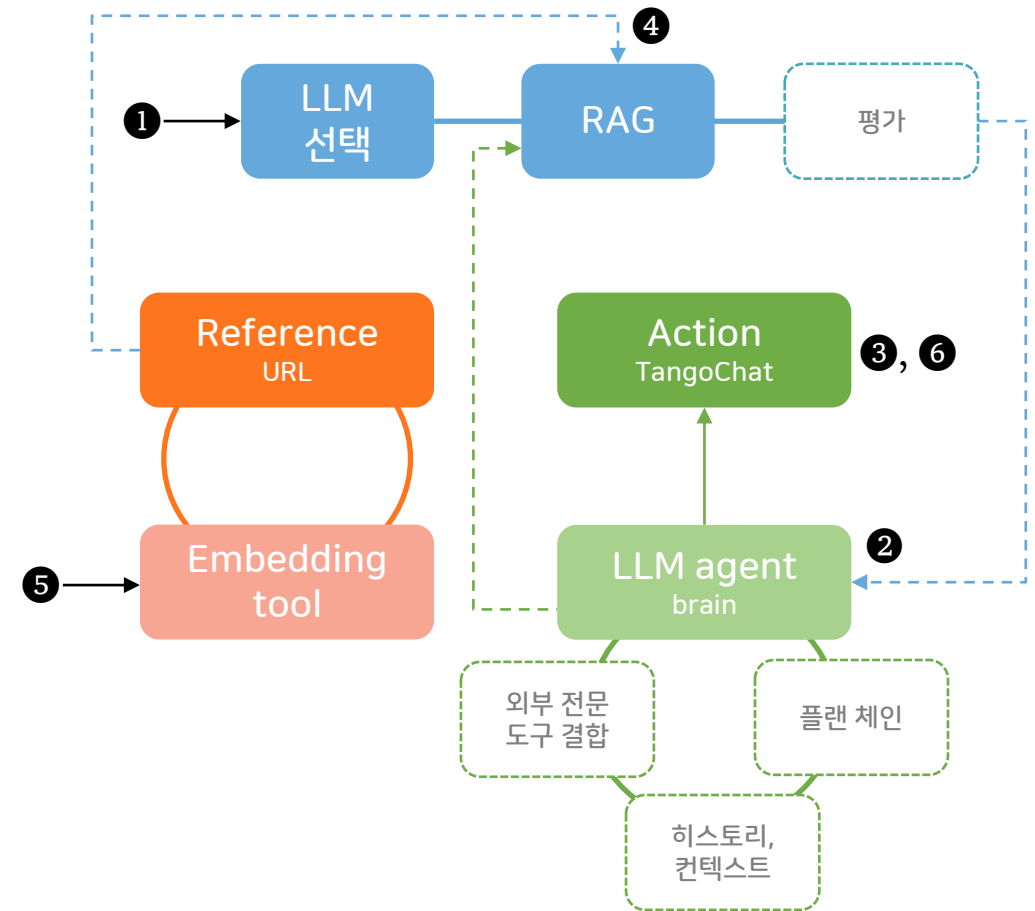
2. TANGO 프레임워크의 구조 변화 AutoNN(모델 생성)보다는 CodeGen(모델 적용)의 비중 확대

외부 액션 아이템을 결합할 수 있는 LLM 에이전트 중심으로 재구성



프로토타입 콘텐츠

- ❶ 공개 LLM 모델의 다운로드
- ❷ 다운로드 한 로컬 LLM 모델의 에이전트/브레인 적용
- ❸ 간단한 챗봇 결합: 서로 다른 LLM 모델 사이의 답변 비교
 - 가장 단순한 버전의 RAG 프로토타입
- ❹ 원하는 정보가 있는 URL 정보 제공
- ❺ URL 정보를 임베딩하기 위한 모델 다운로드 및 적용
- ❻ 더 정확한 답변 추출



1. TASK

태스크 선택 "Chat" (Learning Type 비활성화)

Task = Chat

Edit Project



✓ Project Info
Enter Project Info

2 Configuration
Select Configuration

3 Dataset
Select Dataset

4 Target
Select Target

Configuration

Task Type

☐ Classification ☐ Detection ☒ Chat

Learning Type

☒ Normal ☐ Transfer ☐ HPO

Input Source

Camera ID ▼

Camera ID (0~9)

0

Output Method

Screen Display ▼

Screen Display

0

Deploy Config

Light Weight Level

5

Precision Level

5

User Editing

No ▼

PREV NEXT

2. TARGET

타겟 선택 : 현재는 크게 의미 없음

Target = PC

Edit Project

✓ Project Info
Enter Project Info




✓ Configuration
Select Configuration

✓ Dataset
Select Dataset

4 Target
Select Target

Target

Selected Target: pc

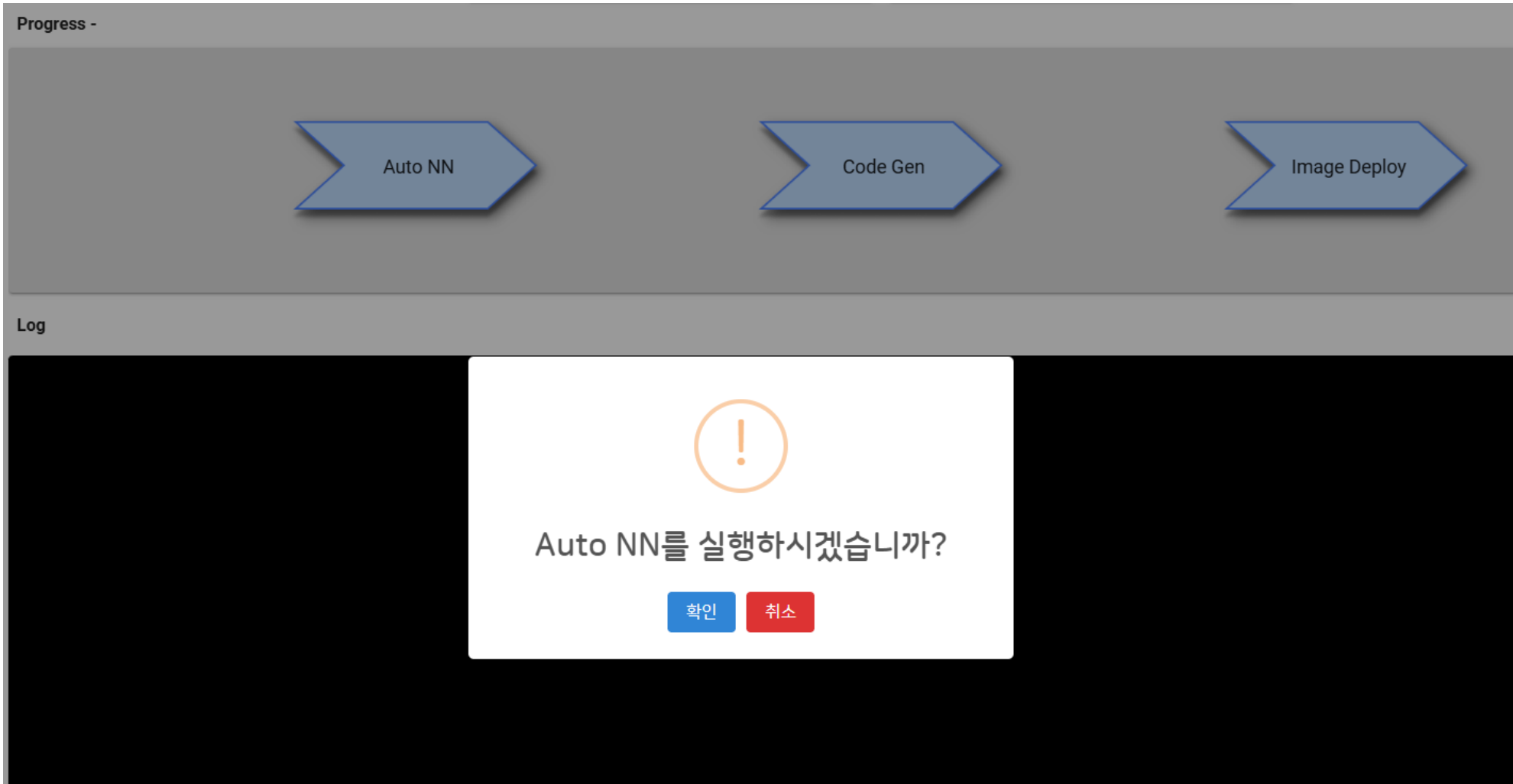
 Google Cloud Platform 2024-11-20	Info	Memory	OS	CPU	Accelerator	Engine
	GCP	2048 MB	ubuntu	x86	cpu	pytorch
 pc 2024-11-20	Info	Memory	OS	CPU	Accelerator	Engine
	PC	2048 MB	ubuntu	x86	cuda	pytorch
 Jetson-orin 2024-11-20	Info	Memory	OS	CPU	Accelerator	Engine
	Jetson_AGX_Orin	2048 MB	ubuntu	arm	cuda	tensorrt

PREV CREATE

3. AUTO NN

Auto NN 실행

Auto NN 클릭



4. TangoChat

TangoChat home

기존 TANGO에서 TANGO+CHAT으로 화면 전환

DOWNLOAD RUN DELETE RAG


Download a model

☐ **Llama 3.2** (3B) ☐ **Llama 3.1** (8B) ☐ **Phi 3.5** (4B) ☐ **Mistral 0.3** (7B) ☐ **Neural-Chat** (7B) ☐ **CodeLlama** (7B) ☐ **LLaVA 1.6** (8B) ☐ **Gemma 2** (9B) ☐ **Qwen 2.5** (7B) ☐ **EEVE Korean** (11B)
 Meta: small Meta: latest Microsoft Mistral AI Intel Meta: coding UW-Madison: vision Google Alibaba Yanolja: 한국어

Download

TangoChat  **v1.0**

S Select an AI model you want [DOWNLOAD > RUN > APPLY]

 Welcome! How can I help you?

Your message



5. DOWNLOAD Download LLM

공개 LLM 모델을 로컬 PC로 다운로드


DOWNLOAD RUN DELETE RAG

Download a model

☒ **Llama 3.2 (3B)**
 ☐ Llama 3.1 (8B)
 ☐ Phi 3.5 (4B)
 ☐ Mistral 0.3 (7B)
 ☐ Neural-Chat (7B)
 ☐ CodeLlama (7B)
 ☐ LLaVA 1.6 (8B)
 ☐ Gemma 2 (9B)
 ☐ Qwen 2.5 (7B)
 ☐ EEVE Korean (11B)

Meta: small
 Meta: latest
 Microsoft
 Mistral AI
 Intel
 Meta: coding
 UW-Madison: vision
 Google
 Alibaba
 Yanolja: 한국어

Download

 Downloading...



DOWNLOAD RUN DELETE RAG

Download a model

☒ **Llama 3.2 (3B)**
 ☐ Llama 3.1 (8B)
 ☐ Phi 3.5 (4B)
 ☐ Mistral 0.3 (7B)
 ☐ Neural-Chat (7B)
 ☐ CodeLlama (7B)
 ☐ LLaVA 1.6 (8B)
 ☐ Gemma 2 (9B)
 ☐ Qwen 2.5 (7B)
 ☐ EEVE Korean (11B)

Meta: small
 Meta: latest
 Microsoft
 Mistral AI
 Intel
 Meta: coding
 UW-Madison: vision
 Google
 Alibaba
 Yanolja: 한국어

Download

✓ llama3.2 is successfully downloaded.

Done(34.16 sec).

6. RUN

로컬 모델 확인 → TangoChat 브레인 교체 → Chat

RUN 탭 → APPLY 클릭 → Your message 작성

DOWNLOAD **RUN** DELETE RAG

Make TangoChat smart!

- ☒ Llama 3.2 (3B)
☐ EEVE Korean (11B)

Apply

TangoChat v1.0

S Select an AI model you want [DOWNLOAD > RUN > APPLY]

Welcome! How can I help you?

Your message

DOWNLOAD **RUN** DELETE RAG

Make TangoChat smart!

- ☐ Llama 3.2 (3B)
☒ EEVE Korean (11B) ① 선택

Apply

② 클릭

TangoChat v1.0

S Let's talk to EEVE KOREAN !! ② 적용 확인

Welcome! How can I help you?

Your message

TangoChat v1.0

S Let's talk to EEVE KOREAN !!

Welcome! How can I help you?

ML-TANGO가 뭐니?

✓ Done, averaged 18.14 tokens/second

안녕하세요! ML-Tango에 대해 설명해 드리겠습니다. ML-Tango는 인간과 대화할 수 있는 AI 어시스턴트입니다. 머신 러닝 기술을 사용하여 자연어 처리(NLP) 능력을 향상시켜 더 정확하고 관련성 높은 응답을 제공할 수 있습니다.

ML-Tango의 주요 기능 중 하나는 대화의 맥락을 이해하고 그에 맞게 반응하는 능력입니다. 이를 통해 대화 흐름을 따라가며 일관된 방식으로 소통할 수 있습니다. 또한, ML-Tango는 다양한 주제에 대해 정보를 제공할 수 있으며 사용자의 질문에 도움을 주기 위해 지식 기반을 갖

Your message

7. RAG

URL 기반 간단한 RAG 프로토타입

RUN 탭 → APPLY 클릭 → Your message 작성

DOWNLOAD RUN DELETE **RAG**

Large Language Model

TangoChat is now running by

EEVE Korean (11B)

①
현재 TangoChat
적용 브레인 확인

TangoChat v1.0

S Let's talk to EEVE KOREAN !!

Welcome! How can I help you?

ML-TANGO가 뭐니?

Your message

② 임베딩 모델 선택

Embedding Model

VECTOR to LLM ⇄ TEXT from URL

- ☒ MXBai-embed-large (334M)
- ☐ Nomic-embed-text (137M)
- ☐ All-miniLM (23M)

PULL & APPLY

Retrieve Source

URL

e.g. <https://github.com/ML-TANGO/TANGO>

Retrieve

Embedding Model

VECTOR to LLM ⇄ TEXT from URL

- ☒ MXBai-embed-large (334M)
- ☐ Nomic-embed-text (137M)
- ☐ All-miniLM (23M)

PULL & APPLY **②** 클릭

✓ mxbai-embed-large is successfully applied.

Done(1.72 sec).

TangoChat is now using mxbai-embed-large.

③ 임베딩 모델 다운로드 및 적용 확인

8. RETRIEVE

URL 정보 제공 후 Retrieve

Retrieve URL

DOWNLOAD RUN DELETE **RAG**

Large Language Model

TangoChat is now running by

EEVE Korean (11B)

Embedding Model

VECTOR to LLM ⇐ TEXT from URL

- ☒ MXBai-embed-large (334M)
- ☐ Nomic-embed-text (137M)
- ☐ All-miniLM (23M)

PULL & APPLY

TangoChat is now using mxbai-embed-large.

Retrieve Source

URL

<https://github.com/ML-TANGO/TANGO>

Retrieve

① URL 제공

Retrieve Source

URL

<https://github.com/ML-TANGO/TANGO>

Retrieve

② 클릭

✓ successfully retrieved.


TangoChat retrieves from

<https://github.com/ML-TANGO/TANGO>

③ URL Retrieve 확인

TangoChat  **v1.0**

S Let's talk to EEVE KOREAN !!

 Welcome! How can I help you?

 ML-TANGO가 뭐니?

Your message



9. ANSWER with your Context RAG 후 CHAT

같은 질문에 대한 다른 답 확인

그럴듯한 답변



ML-TANGO가 뭐니?



✓ Done, averaged 18.14 tokens/second

안녕하세요! ML-Tango에 대해 설명해 드리겠습니다. ML-Tango는 인간과 대화할 수 있는 AI 어시스턴트입니다. 머신 러닝 기술을 사용하여 자연어 처리(NLP) 능력을 향상시켜 더 정확하고 관련성 높은 응답을 제공할 수 있습니다.

ML-Tango의 주요 기능 중 하나는 대화의 맥락을 이해하고 그에 맞게 반응하는 능력입니다. 이를 통해 대화 흐름을 따라가며 일관된 방식으로 소통할 수 있습니다. 또한, ML-Tango는 다양한 주제에 대해 정보를 제공할 수 있으며 사용자의 질문에 도움을 주기 위해 지식 기반을 갖

Your message



더 정답에 가까운 답변

Main Feature of TANGO The TANGO framework aims to deploy and load ready-to-use deep learning models for the specific vision task (classification, object detection, or instance segmentation) onto the user's target devices by automatically constructing and training deep learning models without the help of experts or with minimal knowledge on usage on TANGO. To this end, data preparation, neural network model creation, and optimization for target device can be accomplished within TANGO framework. Data Preparation



✓ Done, averaged 30.52 tokens/second




TA NGO (Target Aware No-code Neural Network Generation and Operaatiion Framework)는 초보 사용자도 코드 작업 없이 신경망 어플리케이션을 개발하고 대상 기기에 쉽게 배포할 수 있는 자동 신경망 생성 및 운영 프레임워크를 개발하는 프로젝트의 코드명입니다.

기계 학습 통합 프레임워크용 TANGO는 노련하지 않은 사용자들이 적은 노력이나 가능하면 코드를 전혀 사용하지 않고도 신경망 어플리케이션을 쉽게 개발하고 대상 기기에 배포할 수 있도록 자동 신경망 생성 및 배치 프레임워크를 개발하는 것을 목표로 합니다. 이를 위해 데이터 준비, 신경망 모델 선택 및 타겟 장치에 대한 최적화 작업을 TANGO 프레임워크 내에서 수행할 수 있습니다.

ML-TANGO가 뭐니?



감사합니다.

주관 ETRI () 주최  과학기술정보통신부  정보통신기획평가원

후원

