J-pong Study: Vit Fine-Tuning

Week 1 (2025.12.30)

Group1

Prof. Jae-Hong Lee

Table of Contents

I Introduction

ViT 이론 + 기술적 구현 → 서비스

최서연	논문 study	Data Pipeline 구축, ViT Fine-tuning, Inference Logic 구현, Service Development	
장지수	논문 study		데모 서비스 기획 PPT 제작 Dev & Ops Presentation
서원덕	논문 study Literature Review Presentation		PPT 제작

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

Introduction & Related Work

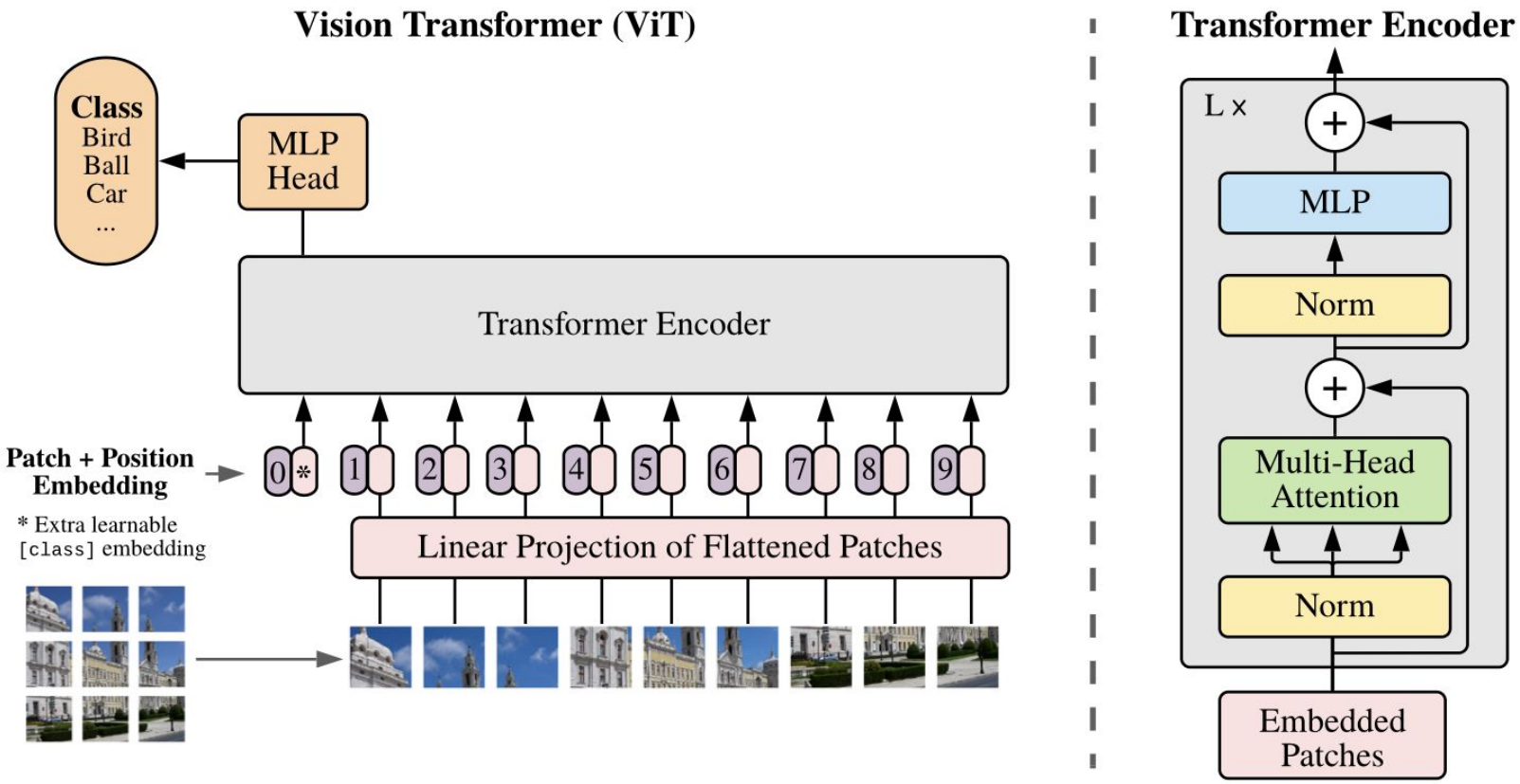
- Context
  - NLP → Transformer (GPT, BERT)
  - Computer Vision → CNN

- Limitation of Transformer on Computer Vision
  - Transformer: high scalability
  - CNN: limited scalability in large datasets

- Mimicking NLP
  - NLP: Token
  - ViT: Patch
  - → State-Of-The-Art achieved

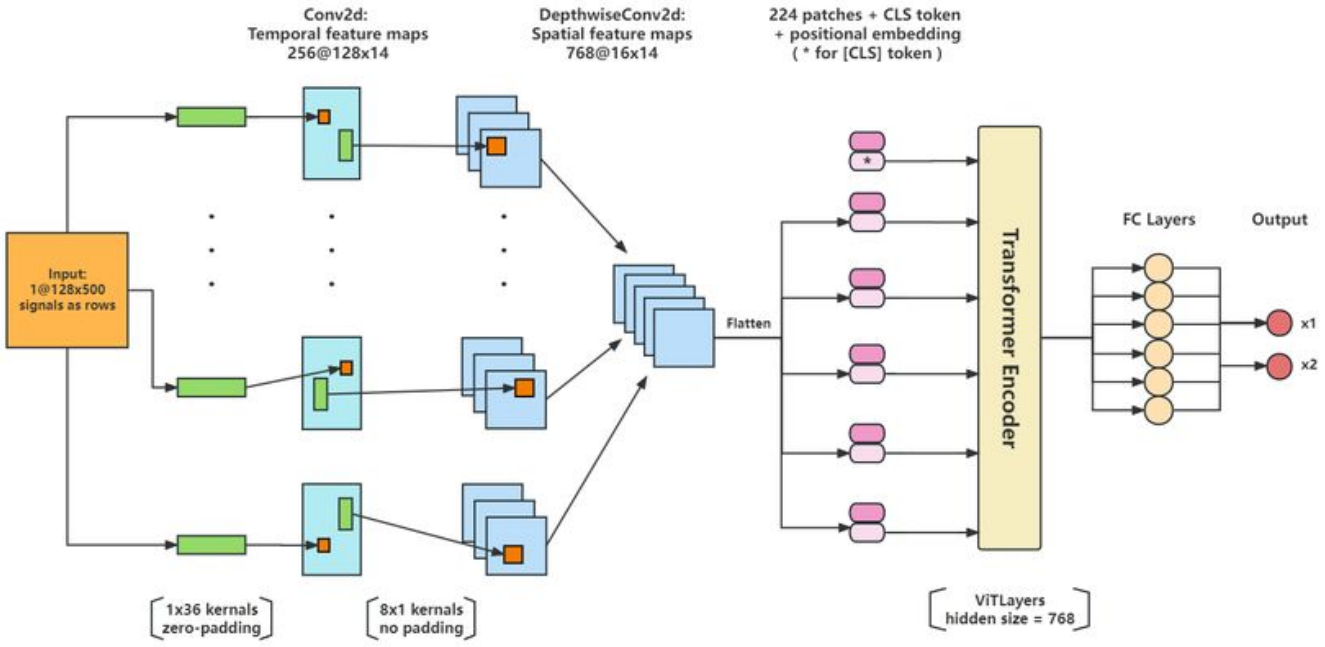
Structure of ViT

1. Patch partition
2. Linear projection
3. Classification Embedding
4. Position Embedding
5. Transformer Encoder
6. MLP Head & Fine-tuning



Structure of ViT

- Hybrid Architecture



- Inductive Bias: CNN vs ViT
  - CNN: 모든 layer에서 local, 평행이동 불변성
  - ViT: MLP 제외 모든 self-attention layer에서 global
    - 2D구조: patch partition, position embedding only

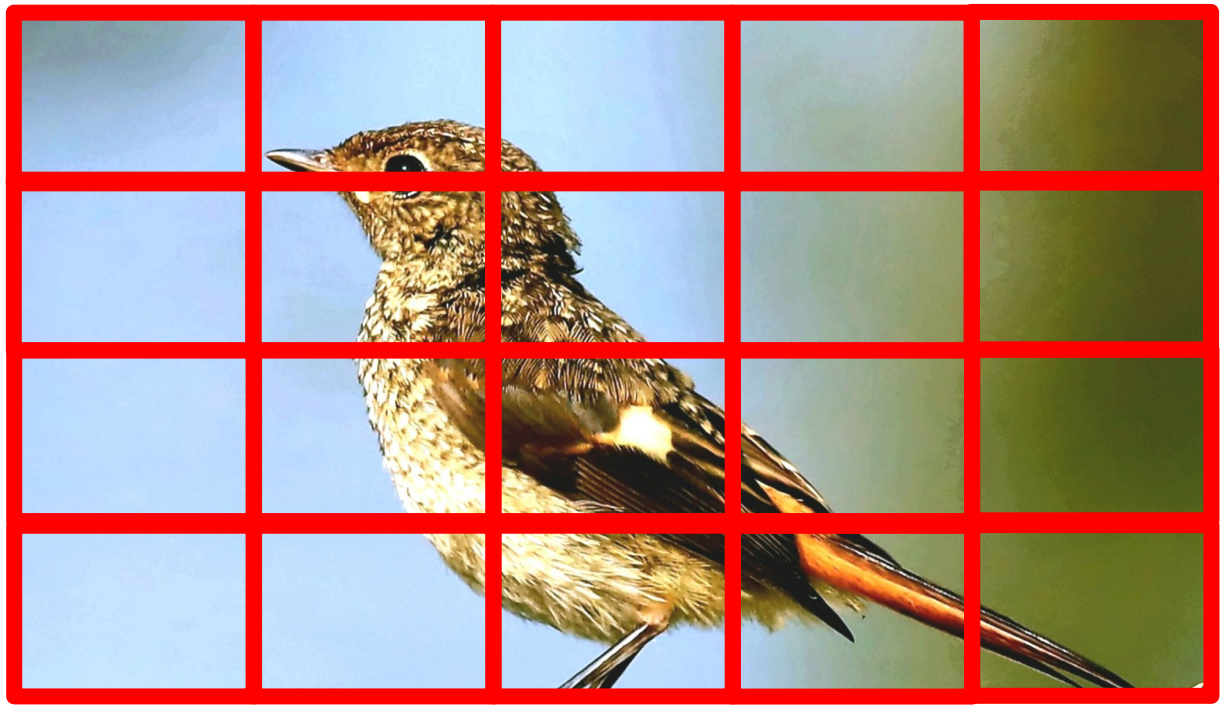
Fine-Tuning

pre-training → fine-tuning

- 기존 head 제거
- k개 class → k개 classification head
- higher resolution image
- 2D interpolation

화질구지

fixed patch size



안화질구지

Experiments - Setup

Model

	ViT (Base/Large/Huge)	**Hybrid**	**ResNet (CNN)**
핵심 구조	BERT-based (24 Layers)	CNN + Transformer	ResNet-v2 (Group Norm)
입력 형태	16 by 16	1 by 1 Feature Map Patches	Standard Convolutions
평가 목표	Scale-up 잠재력 확인	구조 결합의 효율성 검증	기존 모델 대비 벤치마킹

Experiments - Setup

실험 목적

- Transfer Learning Accuracy
- Model Scalability
- Pre-training Efficiency

Dataset

- pretraining
  - small: 1,300만 장 / 1,000개 class
  - medium: 1,400만 장 / 21,000개 class
  - large: 3억 300만 장 / 18,000개 class
- transfer/benchmark
  - VTAB (Visual Task Adaptation Benchmark): Natural/Specialized/Structured

Experiments - Setup

Pre-training and Fine-tuning

- Optimization Strategy

	Pre-training	Fine-tuning
Optimizer	Adam (빠른 초기 수렴)	SGD w/ Momentum (정교한 튜닝)
Key Tech	Large Batch Size (4096)	Polyak Averaging
Goal	대규모 데이터 표현력 학습	특정 Task 성능 극대화

- Evaluation
  - Fine-tuning Accuracy
  - Few-shot Accuracy

Experiments

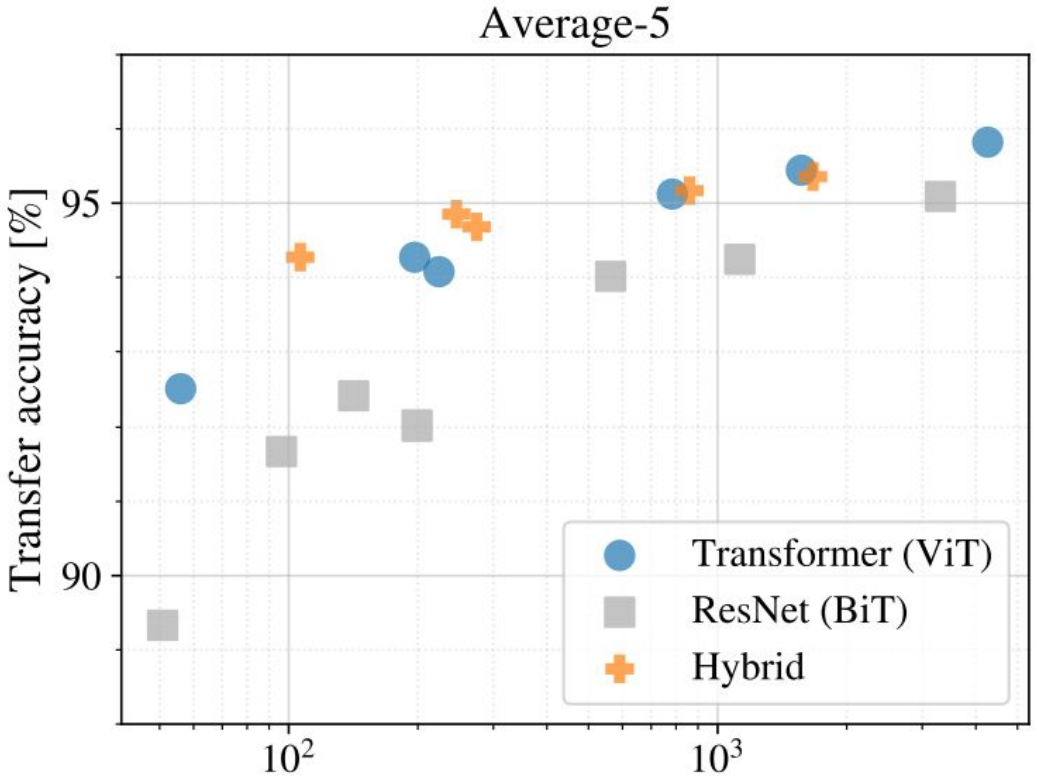
- **Comparison to State-of-the-Art**
- 기존 SOTA model
  - BiT (Big Transfer): CNN 기반 전이학습모델
  - Noisy Student: EfficientNet architecture 기반 image classification model
- New SOTA record (ViT-H/14)
- Efficiency over performance (ViT vs BiT)

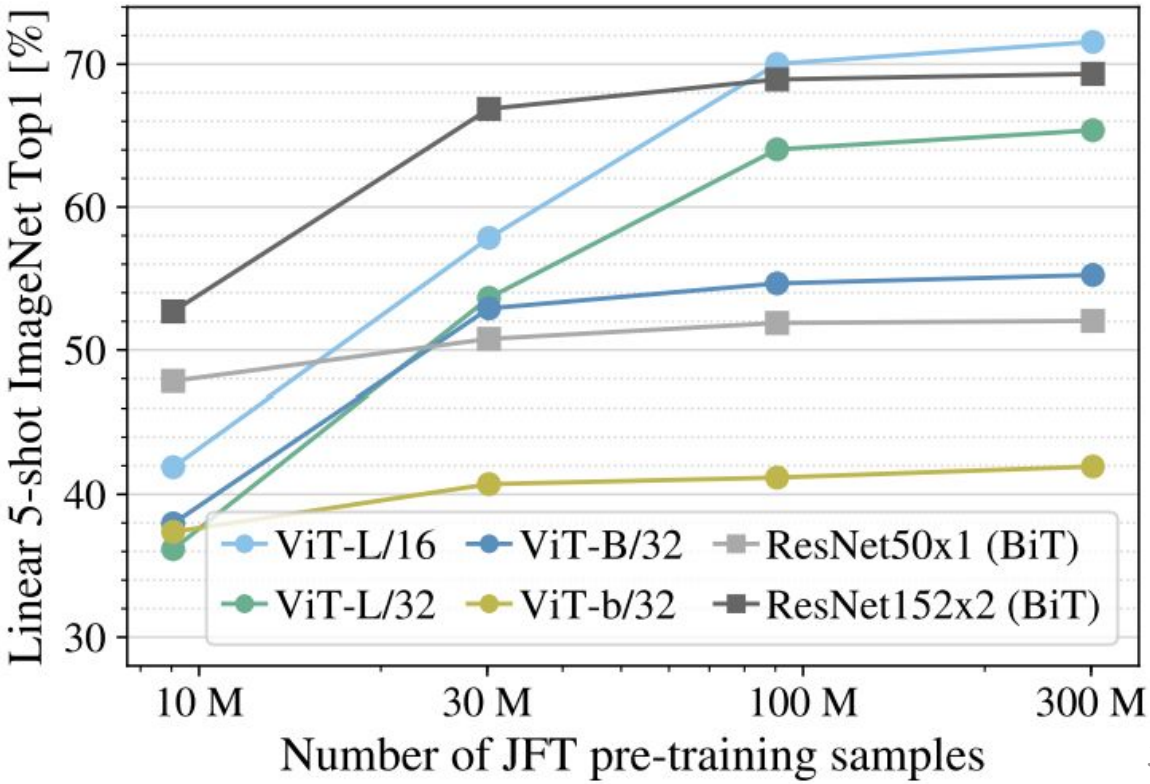
- **Pre-training Data Requirements**

small ~ medium scale	ImageNet-1k	ResNet(CNN) >> ViT
large scale	ImageNet-21k	ResNet >= ViT
huge scale	JFT-300M	ViT >>>>>넘사벽>>>>> ResNet

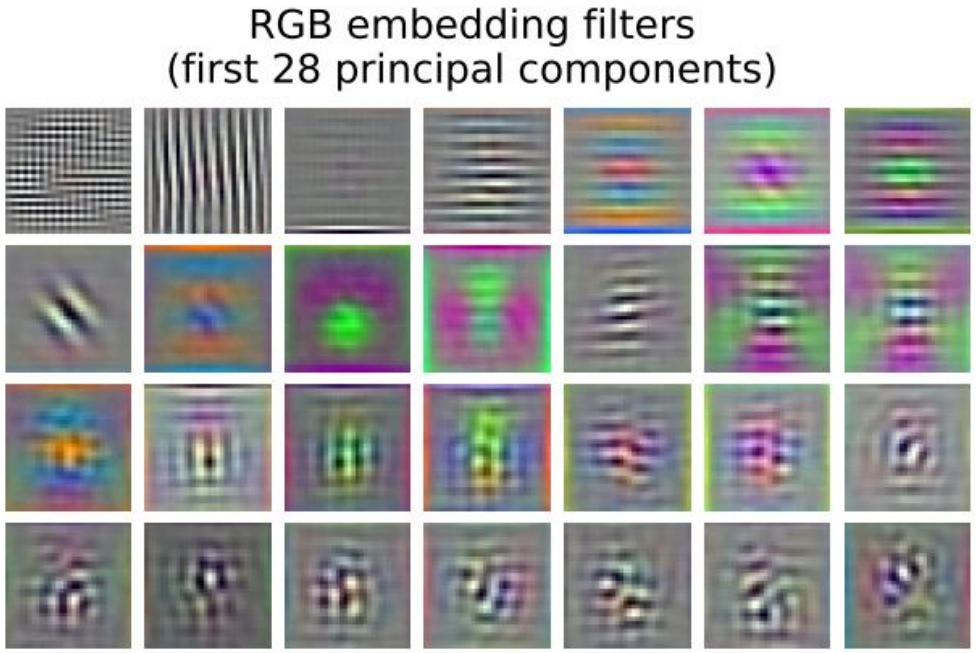
Experiments - Scaling Study

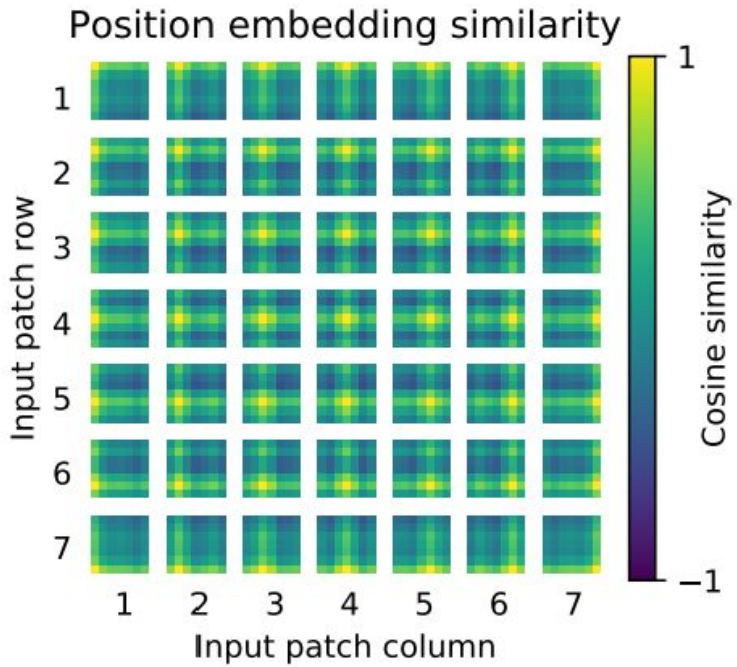
1. Compute Efficiency in same compute budget: ViT > ResNet (accuracy)
2. Hybrid vs ViT: same in large scale
3. No Saturation

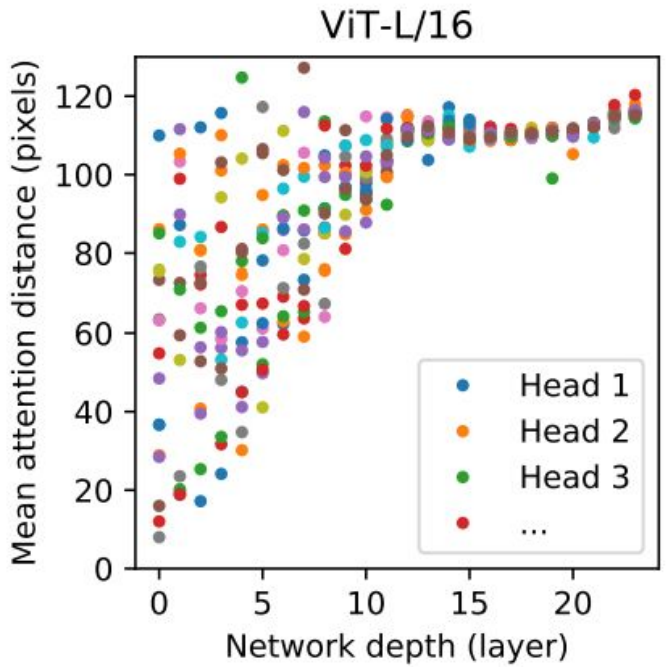




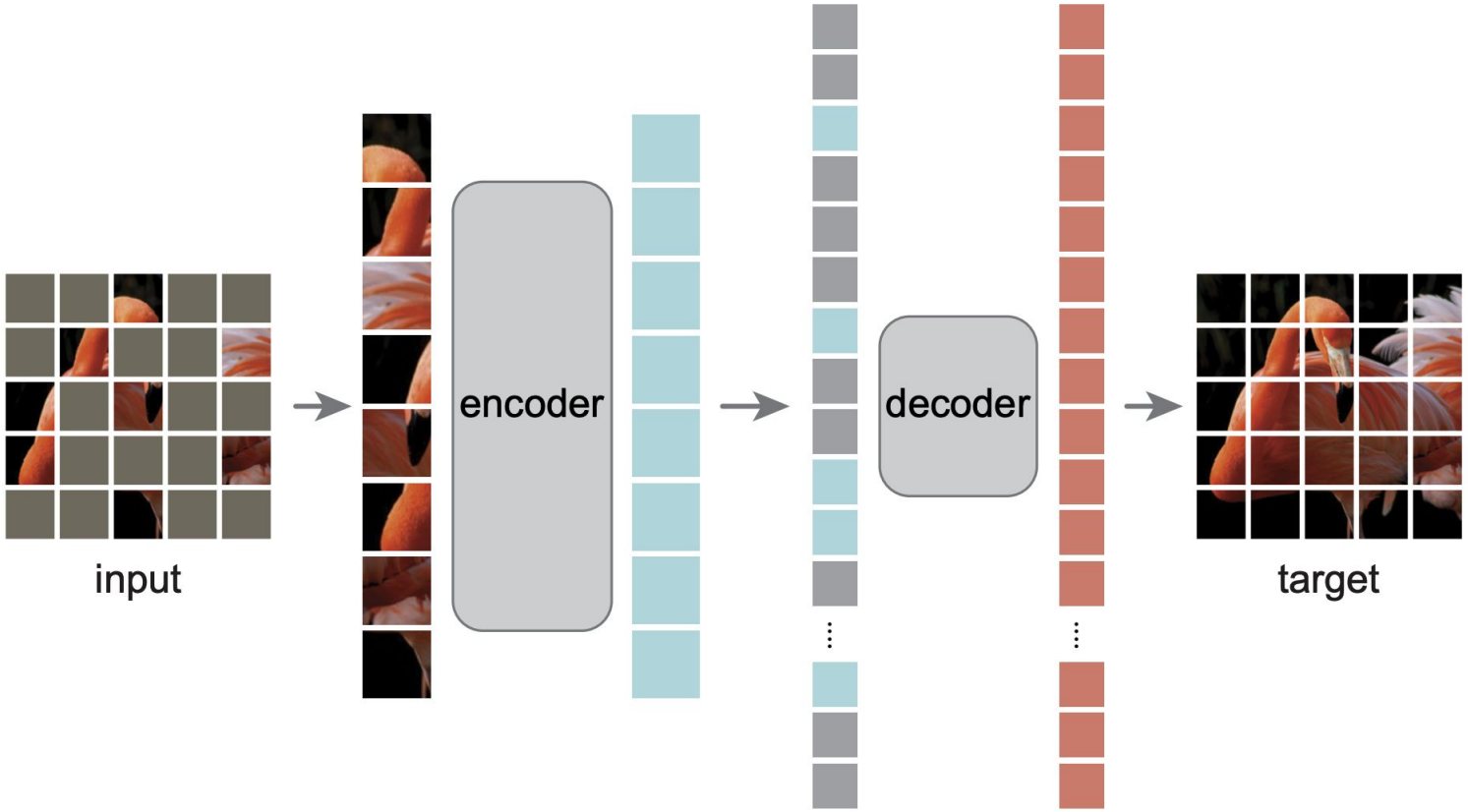
Experiments - Inspecting Vision Transformer

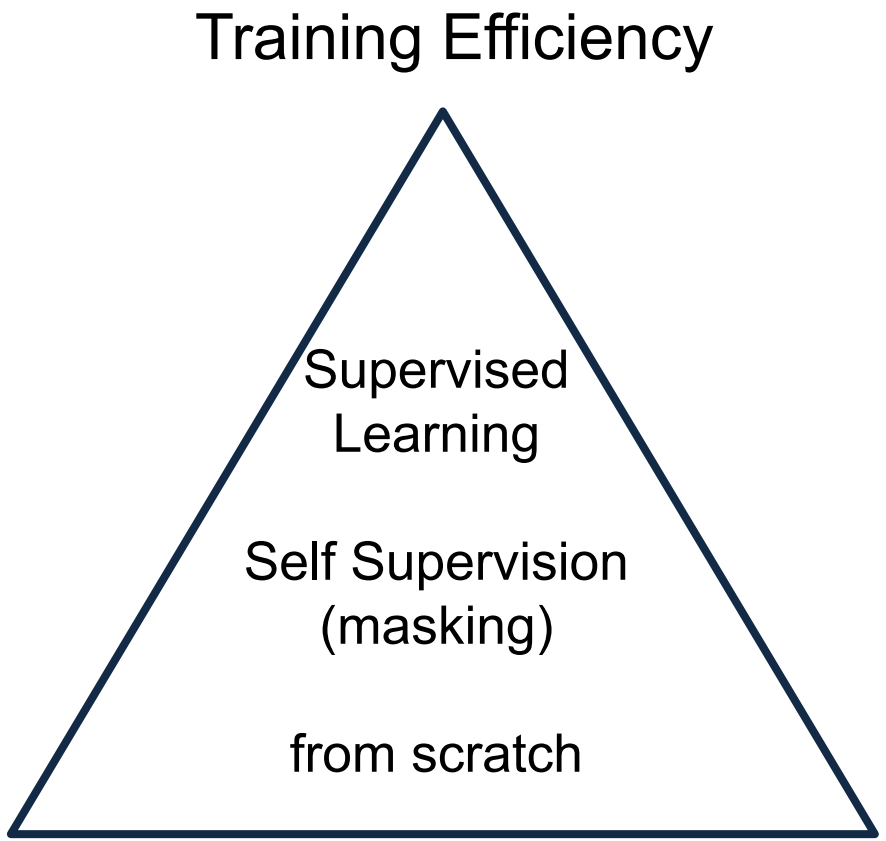






Experiments - Self Supervision





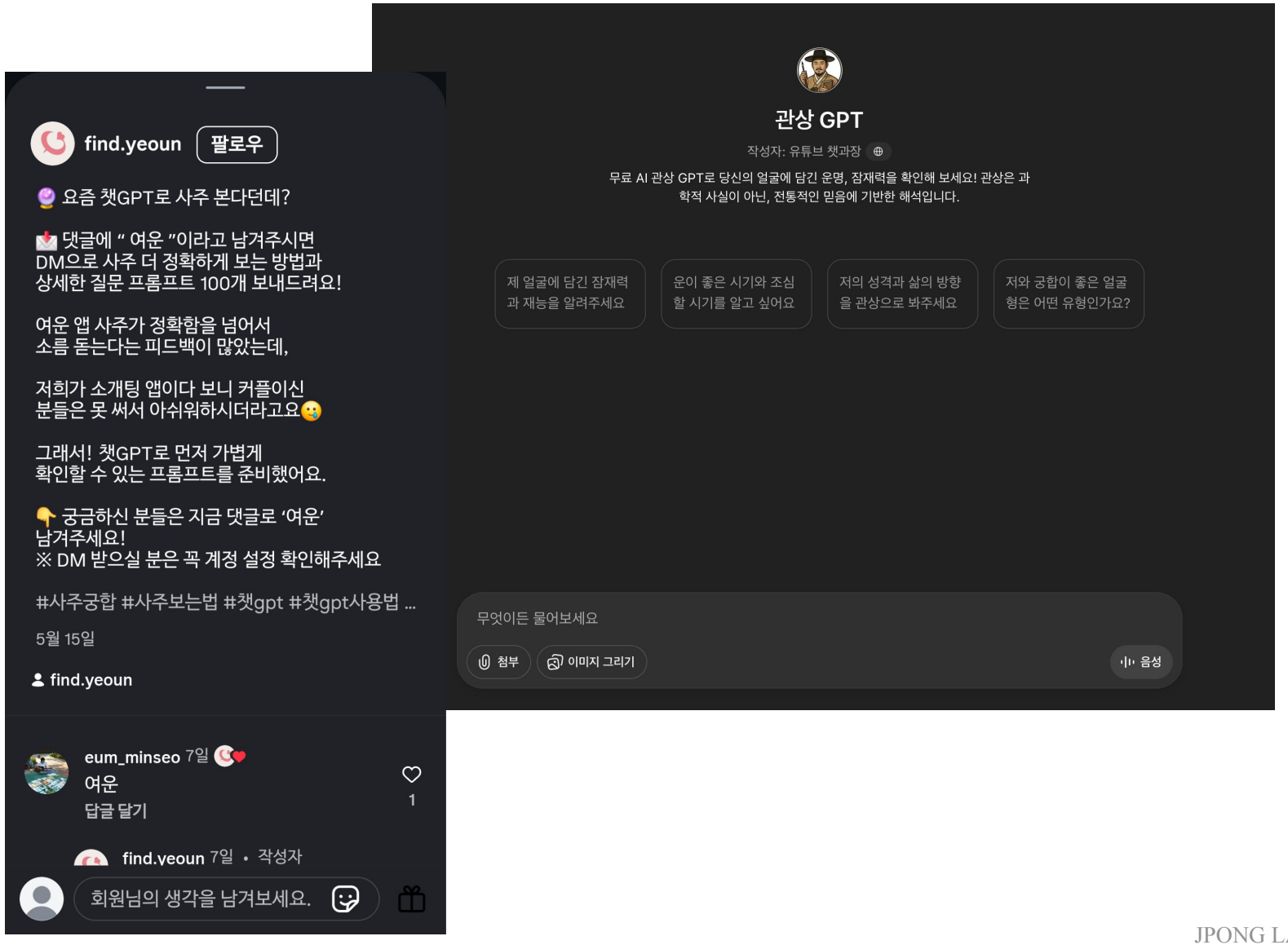
Conclusion

1. Summary of Contributions
   - Generality of Transformers
   - Inductive Bias Mimalization
   - Superior Scalability
   - Computational Efficiency

2. Future Directions
   - Task Expansion
   - Self-supervised Learning
   - Further Scaling

Using AI…

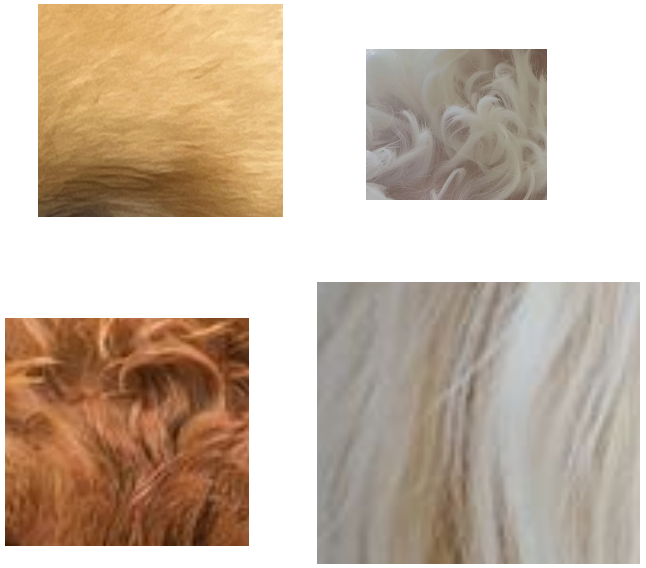
사주, 타로, 관상
…



강아지 관상

III Dev & Ops: 1. Market Opportunity

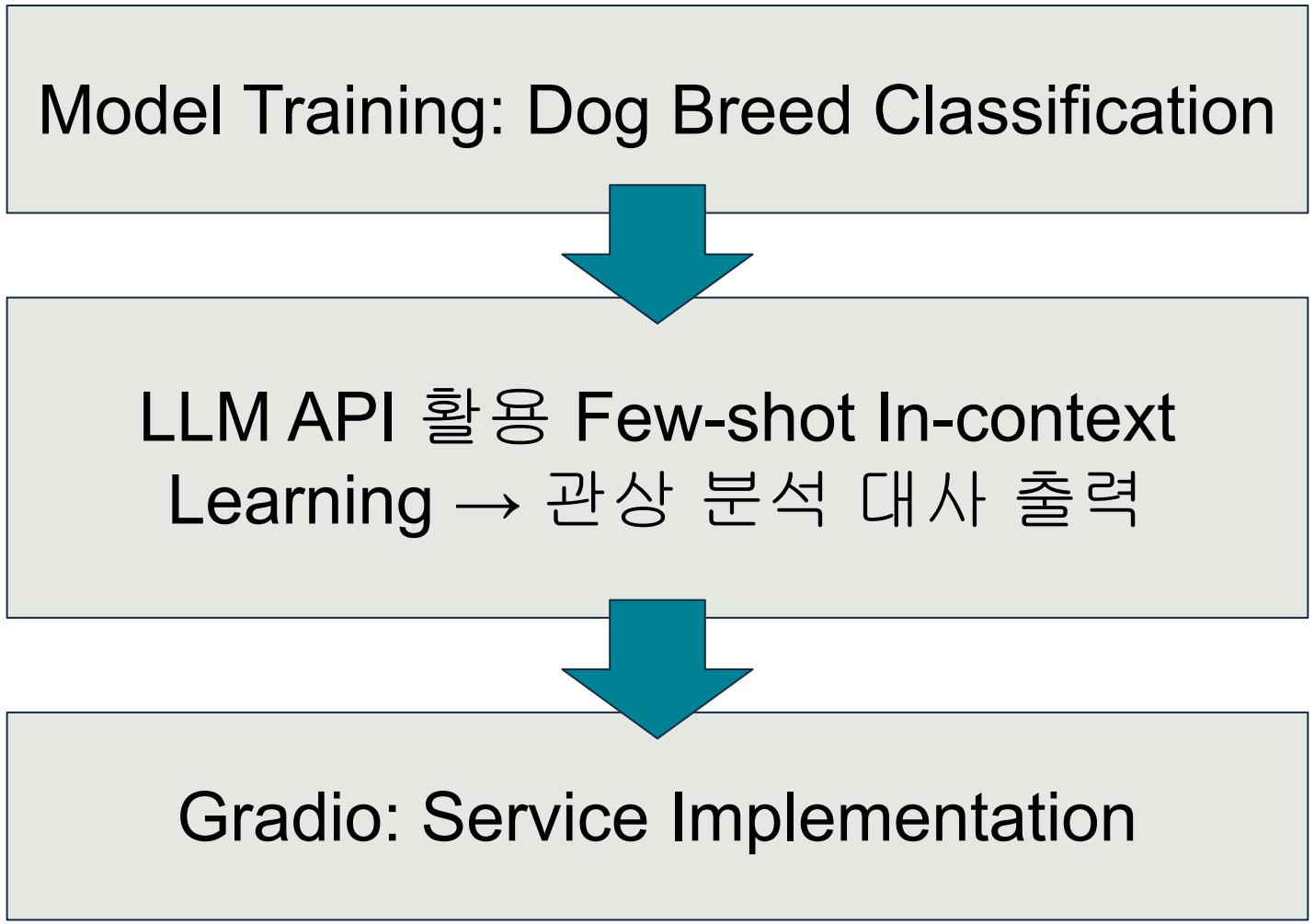
기술적 지향점



visual feature

소비자 트렌드

III Dev & Ops: 2. Strategic Direction



Dataset Overview

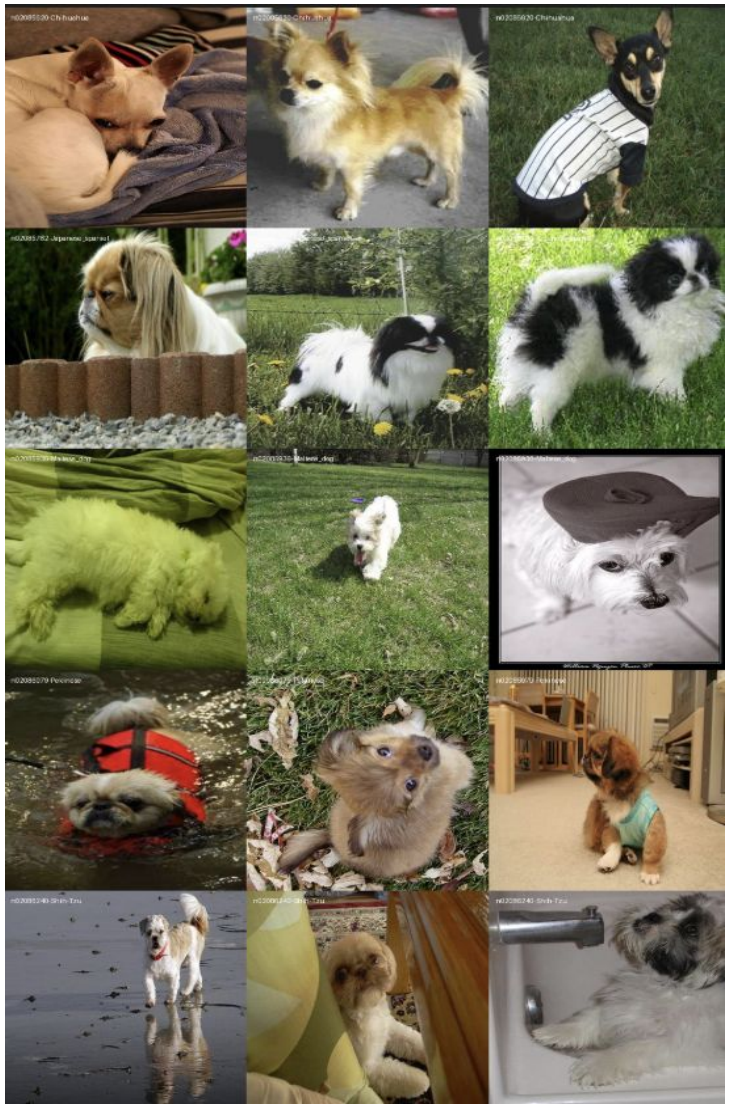
- Target : Stanford Dogs Dataset (120 Breeds)
- Train set 12,000, Test set 8,580 (6:4)
- Task : 미세 견종 분류
- 이미지 수: 총 20,580장

Data Pre-processing

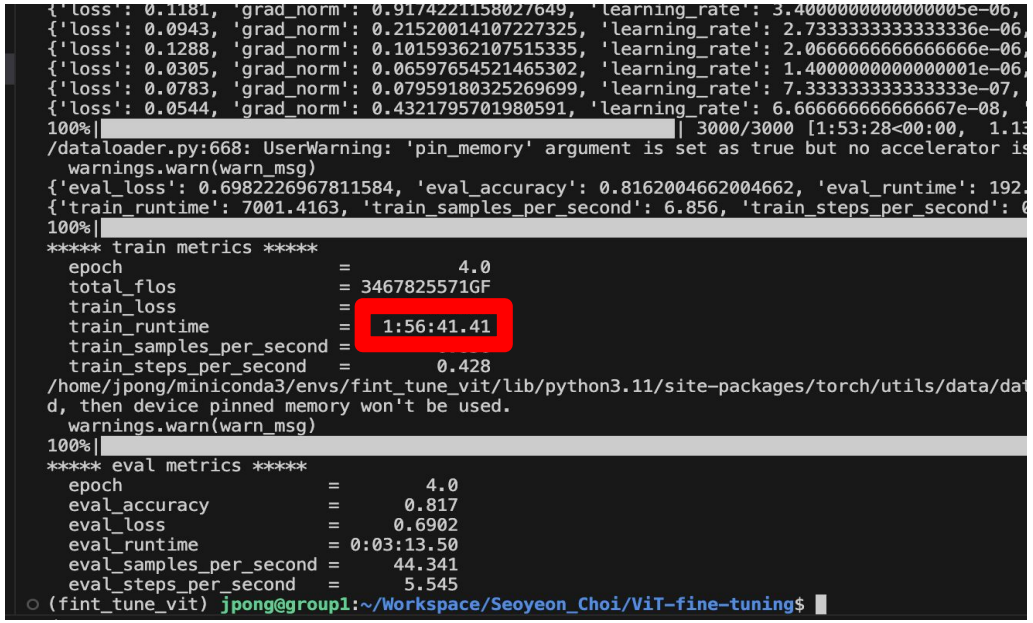
- Input Size Modification : 224 x 224 (ViT 패치 규격에 최적화)
- Normalization: ImageNet 기준 통계량 정규화 적용

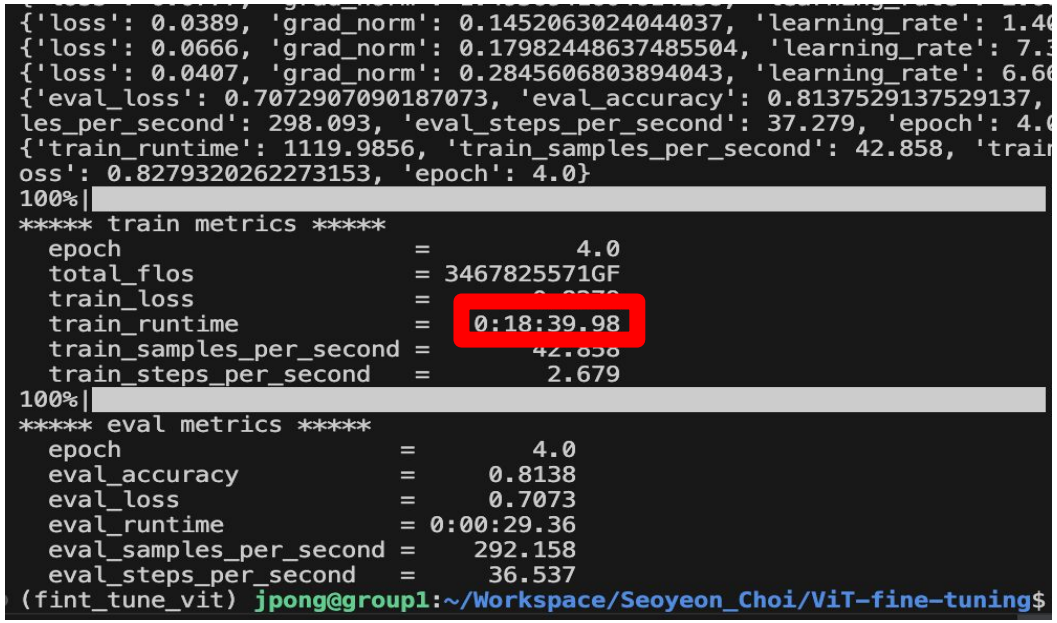
Generalization

- Augmentation : Random Crop, Flip, Rotation 적용 (train set)
- Label Smoothing: label_smoothing_factor = 0.1
- Epoch: epoch = 10



III Dev & Ops: 3. ViT Fine-Tuning: Training



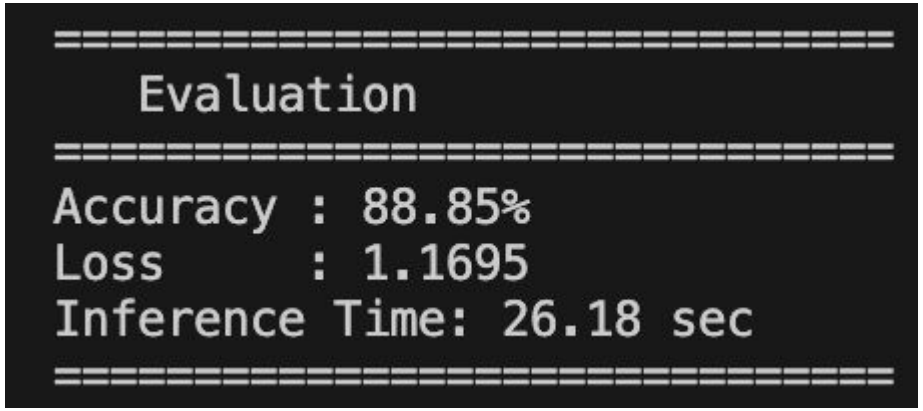


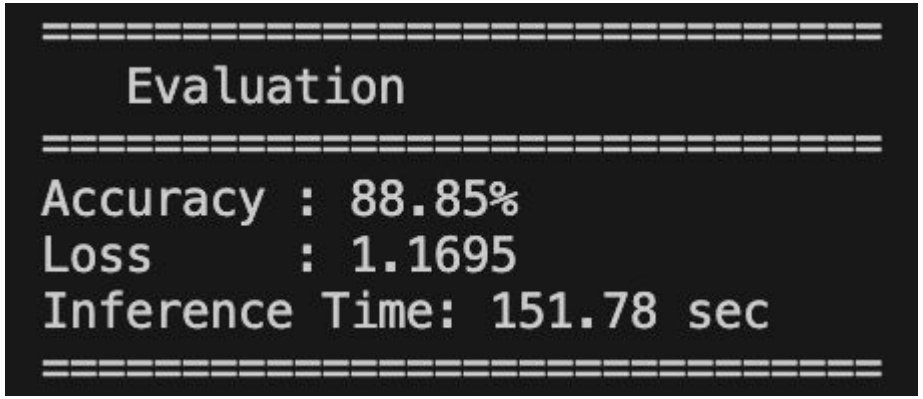
Hardware Acceleration(CPU vs GPU)

	CPU (Intel(R) Xeon(R) w7-3465X)	**GPU** (NVIDIA GeForce RTX 5090)
Runtime	1시간 56분 41초	18분 39초
Throughput	6.86 samples/s(초당 처리량 저조)	42.86 samples/s(처리 속도 향상)

1) Quantitative Performance

- **Validation Accuracy : 88.85%**
  → 120종의 미세 견종 분류 작업에서 약 90%의 정확도 확보
- **Validation Loss : 1.1695**
  → Normalization 및 Augmentation을 통해 Overfitting을 효과적으로 억제하며 모델이 안정적으로 수렴





2) Hardware Acceleration(CPU vs GPU)

GPU : 26.18초
CPU : 151.78초

1) AI Inference Pipeline

- ML Model : Vision Transformer 기반 fine-tuning(224 x 224 이미지 classification)
- LLM api 사용 (Gemini) : ViT 결과를 input으로 받아서 관상 관련 대사를 출력
- Service: Gradio 프레임워크 → 구축, Hugging Face Spaces → 웹 서비스 배포

2) LLM api 기반 대사 출력 : few-shot in-context learning

- input: 견종

- 관상가 말투
- 견종의 신체적 특징 기반 관상 용어 연결
- 주인에게 어떤 복을 가져다 줄 지
- 분량 제한 200자

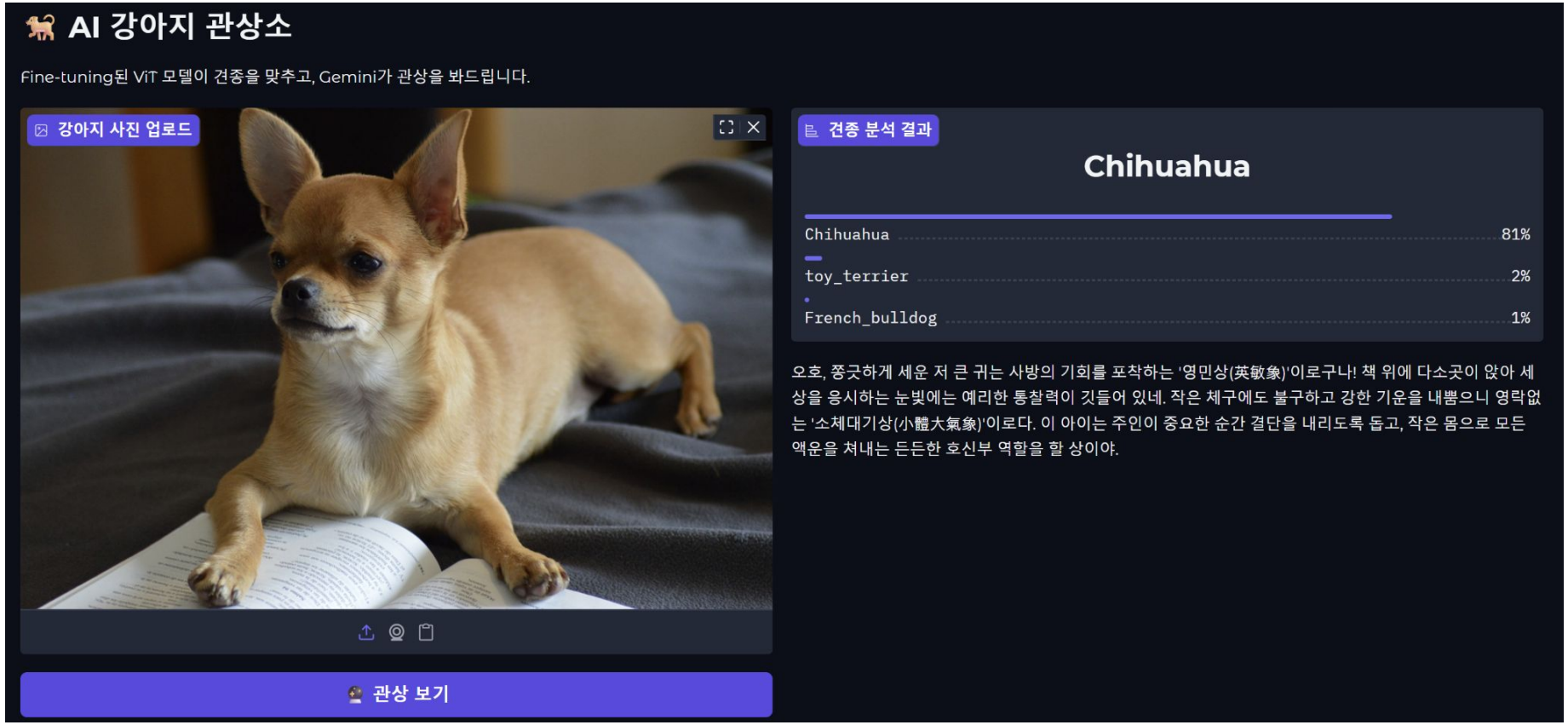
- few shot: 3개 제공

Input: 말티즈

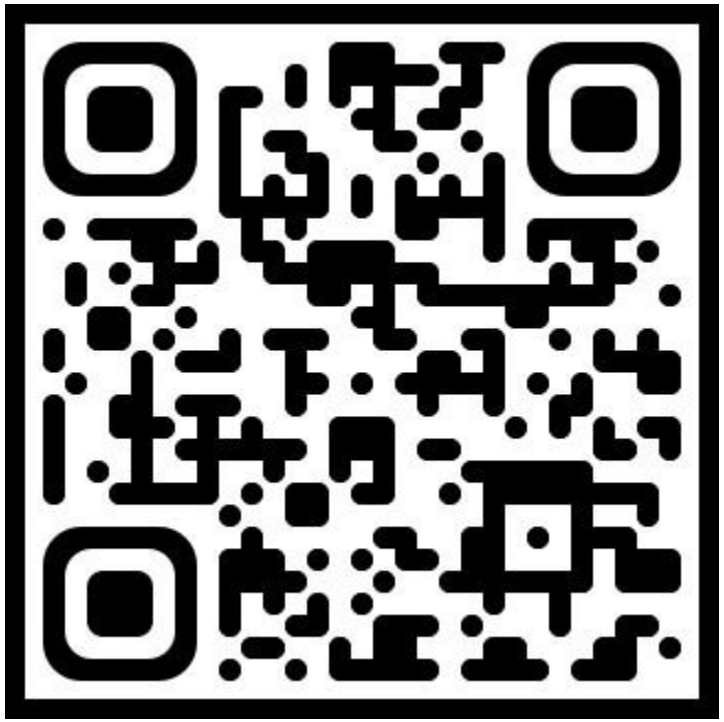
Output: 오호, 눈망울이 밤다래처럼 맑고 초롱초롱한 것을 보니 영락없는 '청명상(**淸明象**)'이로구나! 하얀 털은 집안의 나쁜 기운을 정화하는 기운을 가졌으니, 이 아이가 머무는 곳마다 웃음꽃이 피어날 것이야. 주인에게는 맑은 정신과 평안을 가져다줄 복덩이이니 애지중지 아끼도록 하거라.

3) User Experience & Value

- Interactive UI : Gradio 기반 직관적 웹 서비스
- Emotional Engagement : AI와의 유쾌한 대화 및 재미 선사



III Dev & Ops: 3. ViT Fine-Tuning: Inference & Deployment



III Dev & Ops: 4. Conclusion

ViT 기술 공부
모델 최적화 + LLM 연동
웹 서비스 배포

GPU
Transformer

https://github.com/torilove11/ViT-fine-tuning

Thank You!

Q&A