

Final Project Introduction

Minlie Huang (黃民烈)

aihuang@tsinghua.edu.cn

Conversational Artificial Intelligence Group

<http://coai.cs.tsinghua.edu.cn/hml/>



清华大学
Tsinghua University



Overview



- We provide two options for the final project this year (choose one):
 - 1. Paper Reproduction (using **Jittor**)
 - 2. Research-Oriented Project

Computing Resources



- Jittor:

- ◆ Type: SSH Server. Titan RTX (24G)
- ◆ Advance Application: A100 (80G*N)
- ◆ Only for Reproduction



- 曙光算力中心:

- ◆ Type: 国产显卡 (16G * 4)

- Each team can apply one type of computing resource at most

- See later announcements on 网络学堂

- We encourage you:

- ◆ Try free resources like Google Colab
- ◆ Join lab projects

Paper Reproduction



- The paper reproduction requires you to reimplement the main results of the paper using **the Jittor framework**.
- Build the community of Jittor (Use, Feedback, Share)

- We provide a list of papers that already have PyTorch / Tensorflow codes.
- Choose one paper to reproduce.
- Each team consists up to 3 students.

Requirements



◎ Mandatory

- ◆ Reproduce the main result (**red box in the following pages**)
- ◆ Build an open-source repository: Readme, dependencies, command for train/evaluate ...

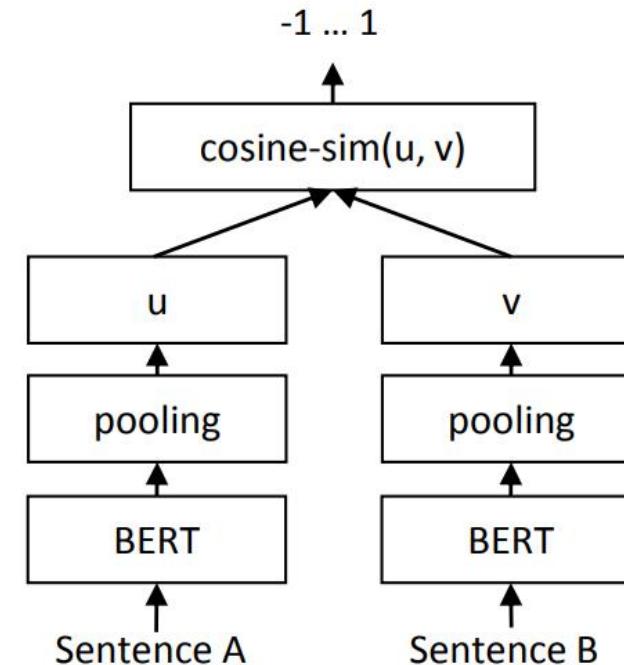
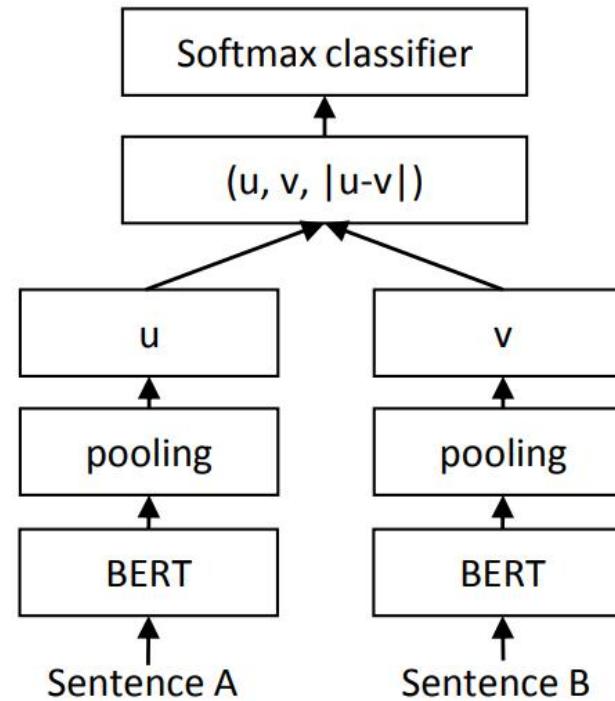
◎ Optional (At least choose one)

- ◆ Profiling: Analysis the bottleneck of running speed
 - Compare with PyTorch / Tensorflow
 - Improve the efficiency (training or test)
- ◆ Improve the model and enhance the performance
- ◆ Transfer to other tasks / datasets
- ◆ Build an interactive demo
- ◆ **Any contributions to Jittor community**

Paper Candidates



- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks



Paper url: <https://arxiv.org/pdf/1908.10084>

Official code: <https://github.com/UKPLab/sentence-transformers>

Paper Candidates



- Task: complete training to do Semantic Textual Similarity / Retrieve & Re-Rank ...

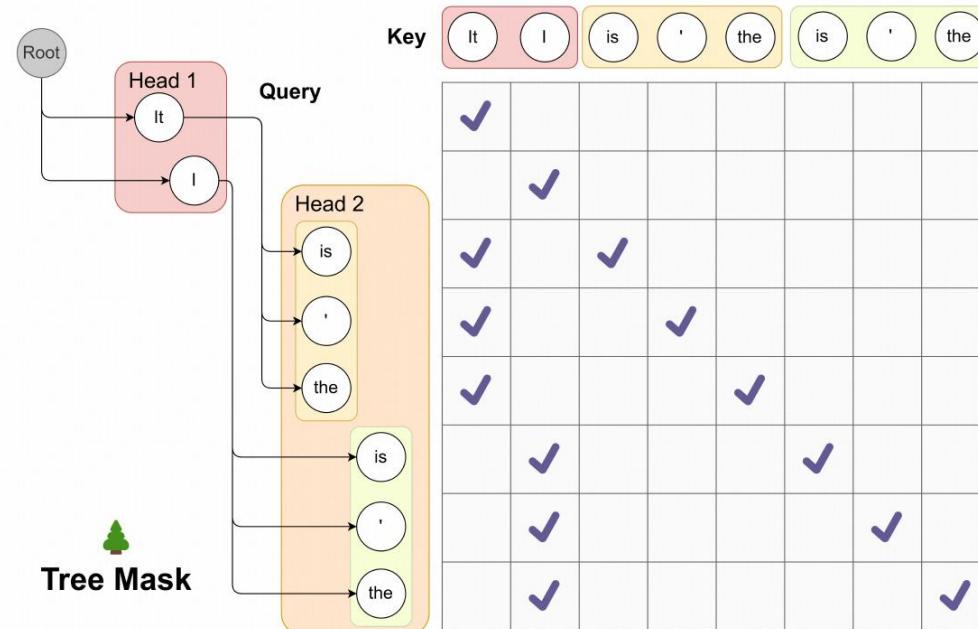
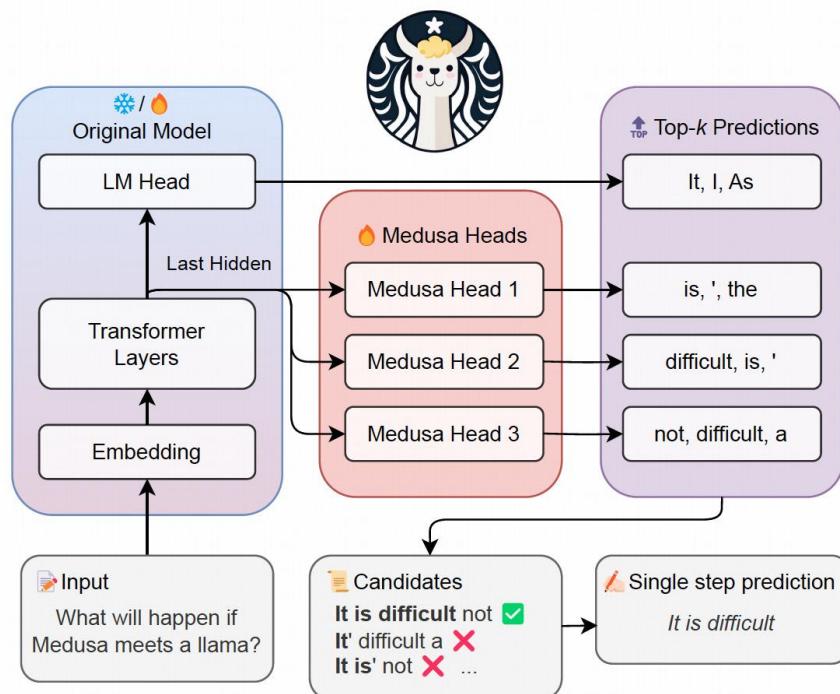
Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	78.46	74.90	80.99	76.25	79.23	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68

- Task: ablation study on network design and other details.

Paper Candidates



- MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads



Paper URL: <https://arxiv.org/abs/2401.10774>

Official Code: <https://github.com/FasterDecoding/Medusa>

Paper Candidates



- MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads
 - ◆ **MEDUSA-1:** Finetuning a MEDUSA head on top of a frozen backbone LLM Vicuna-7B
 - ◆ Compare the speed and quality with the backbone LLM Vicuna-7B

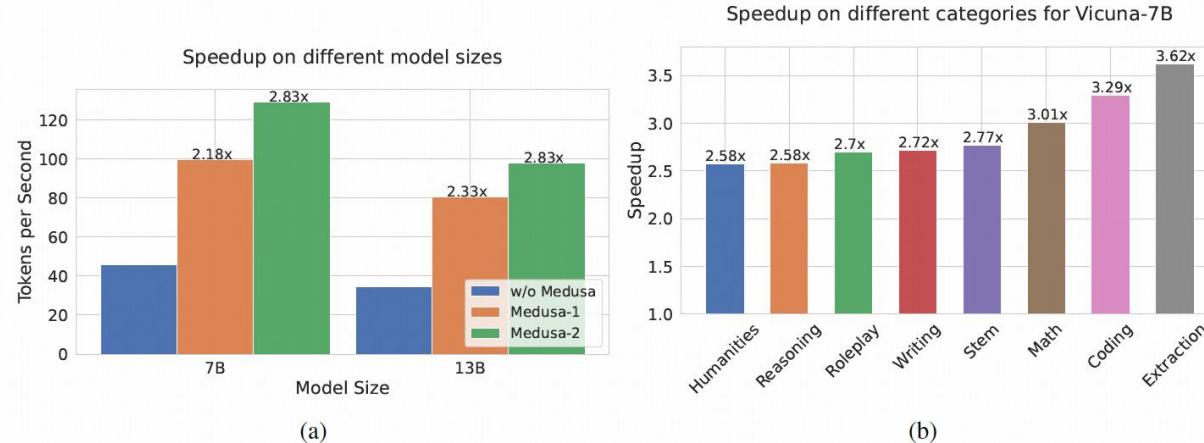


Figure 3. Left: Speed comparison of baseline, MEDUSA-1 and MEDUSA-2 on Vicuna-7B/13B. MEDUSA-1 achieves more than 2× wall-time speedup compared to the baseline implementation while MEDUSA-2 further improves the speedup by a significant margin. Right: Detailed speedup performance of Vicuna-7B with MEDUSA-2 on 8 categories from MT-Bench.

Model Name	Vicuna-7B	Zephyr-7B	Vicuna-13B	Vicuna-33B
Acc. rate	3.47	3.14	3.51	3.01
Overhead	1.22	1.18	1.23	1.27
Quality	6.18 (+0.01)	7.25 (-0.07)	6.43 (-0.14)	7.18 (+0.05)
$S_{\text{SpecDecoding}}$	1.47	-	1.56	1.60
S_{MEDUSA}	2.83	2.66	2.83	2.35

Table 1. Comparison of various MEDUSA-2 models. The first section reports the details of MEDUSA-2, including accelerate rate, overhead, and quality that denoted the average scores on the MT-Bench compared to the original models. The second section lists the speedup (S) of SpecDecoding and MEDUSA, respectively.

Paper Candidates



- ⦿ MEDUSA: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads
 - ◆ (Optional) **MEDUSA-2**: Finetuning the MEDUSA head together with the backbone LLM Vicuna-7B
 - ◆ (Optional) Compare the speed and quality with the backbone LLM and MEDUSA-1
 - ◆ (Optional) **Ablation Study**: Configuration of tree attention and thresholds of typical acceptance

Baseline	Direct Fine-tuning	MEDUSA-1	MEDUSA-2
Quality	6.17	5.925	6.23
Speedup	N/A	N/A	2.18

Table 2. Comparison of Different Settings of Vicuna-7B. Quality is obtained by evaluating models on MT-Bench using GPT-4 as the judge (higher the better).

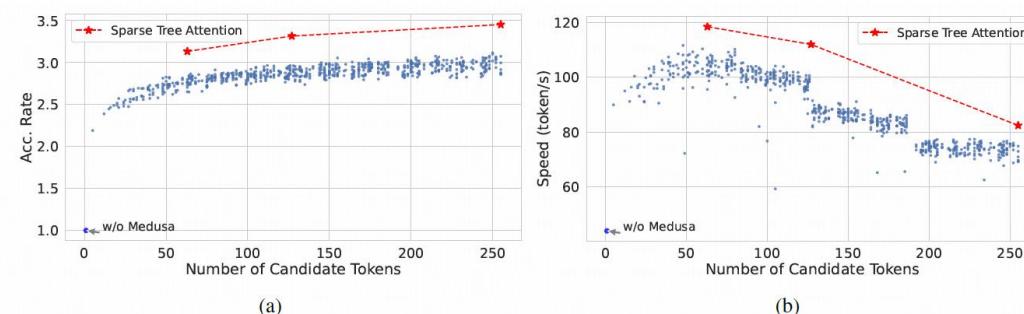


Figure 4. Effectiveness of numbers of candidate tokens for decoding introduced by trees (default number of candidate token for decoding is 1 when using KV cache). Left: The acceleration rate for randomly sampled dense tree settings (blue dots) and optimized sparse tree settings (red stars). Right: The speed (tokens/s) for both settings. The trend lines indicate that while the acceleration rate remains relatively stable for sparse trees, there is a notable decrease in speed as the candidate tokens increases.

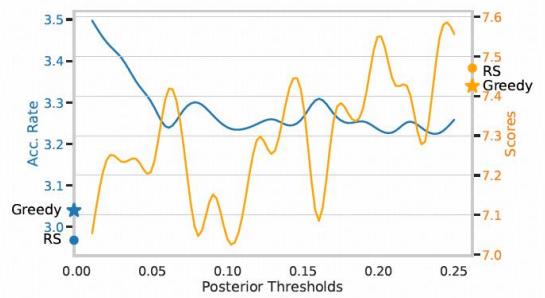


Figure 5. Performance comparison of MEDUSA using proposed typical sampling. The model is fully fine-tuned from Vicuna-7B. The plot illustrates the acceleration rate and average scores on the writing and roleplay (MT-Bench) with a fixed temperature of 0.7 for 3 different settings: greedy sampling and random sampling (RS) plotted as the star and the dot, and typical sampling curves under different thresholds.

Paper Candidates



Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection

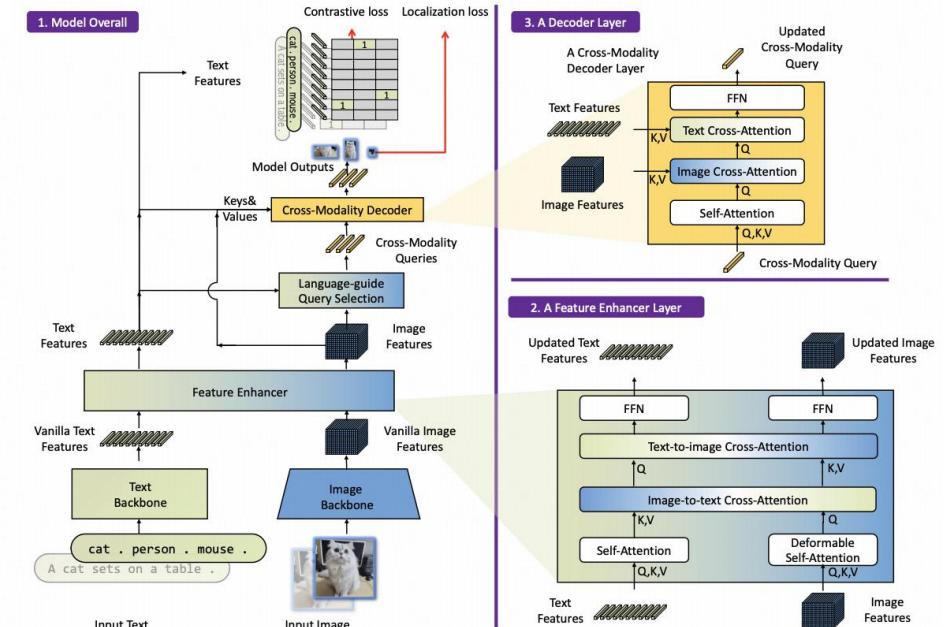
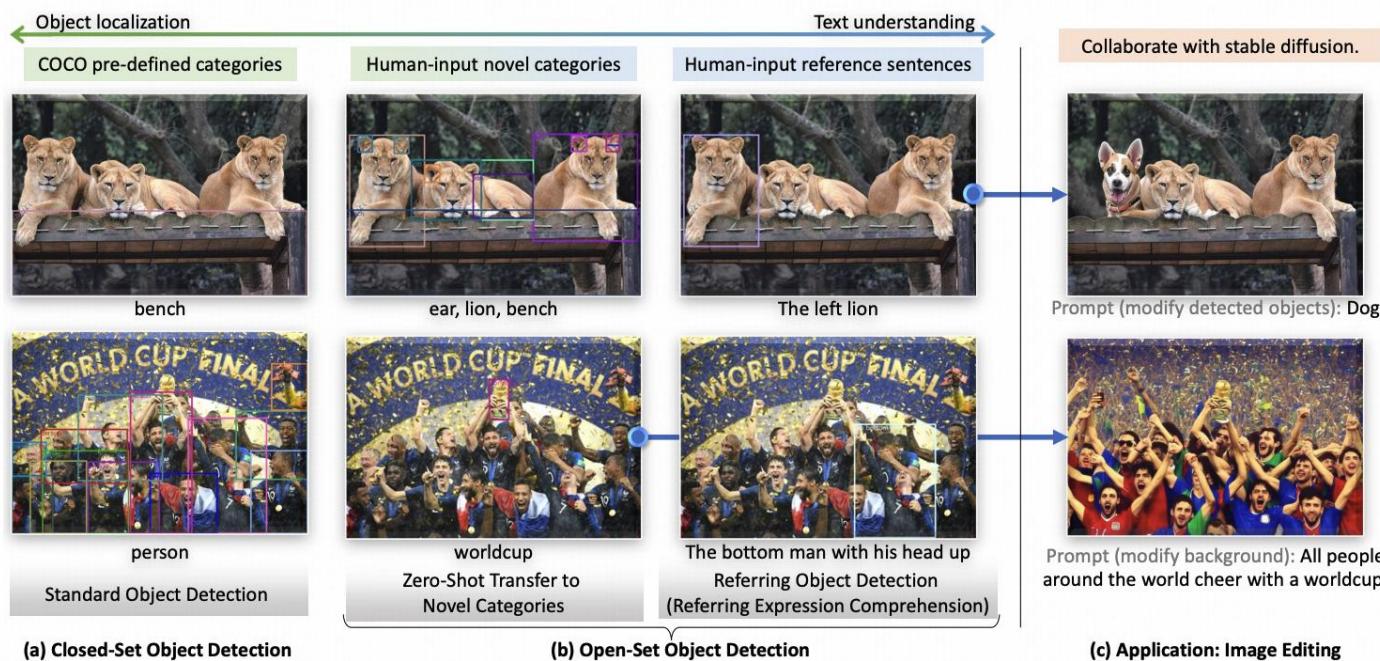


Fig. 3: The framework of Grounding DINO. We present the overall framework, a feature enhancer layer, and a decoder layer in block 1, block 2, and block 3, respectively.

Paper url: <https://arxiv.org/pdf/2303.05499>

Official Code: <https://github.com/IDEA-Research/GroundingDINO>

Training Code: <https://github.com/longzw1997/Open-GroundingDino>

Paper Candidates



◎ Grounding DINO

◆ Task: Lvis [1]

Model	Backbone	Pre-Training Data	MiniVal [18]		
			AP	APr/APc/APf	
<i>Zero-Shot Setting</i>					
GLIP-T (C)	Swin-T	O365,GoldG	24.9	17.7/19.5/31.0	
GLIP-T	Swin-T	O365,GoldG,Cap4M	26.0	20.8/21.4/31.0	
DetCLIPv2	Swin-T	O365,GoldG,CC15M	40.4	36.0/41.7/40.0	
Grounding DINO T	Swin-T	O365,GoldG	25.6	14.4/19.6/32.2	
Grounding DINO T	Swin-T	O365,GoldG,Cap4M	27.4	18.1/23.3/32.7	
Grounding DINO L	Swin-L	O365,OI,GoldG,Cap4M, COCO,RefC	33.9	22.2/30.7/38.8	
<i>Fine-Tune Setting</i>					
MDETR	RN101	GoldG,RefC	24.2	20.9/24.9/24.3	
Mask R-CNN	RN101	-	33.3	26.3/34.0/33.9	
DetCLIPv2 [51]	Swin-T	O365,GoldG,CC15M	50.7	44.3/52.4/50.3	
Grounding DINO T	Swin-T	O365,GoldG	52.1	35.4/51.3/55.7	

Table 3: Model results on LVIS.

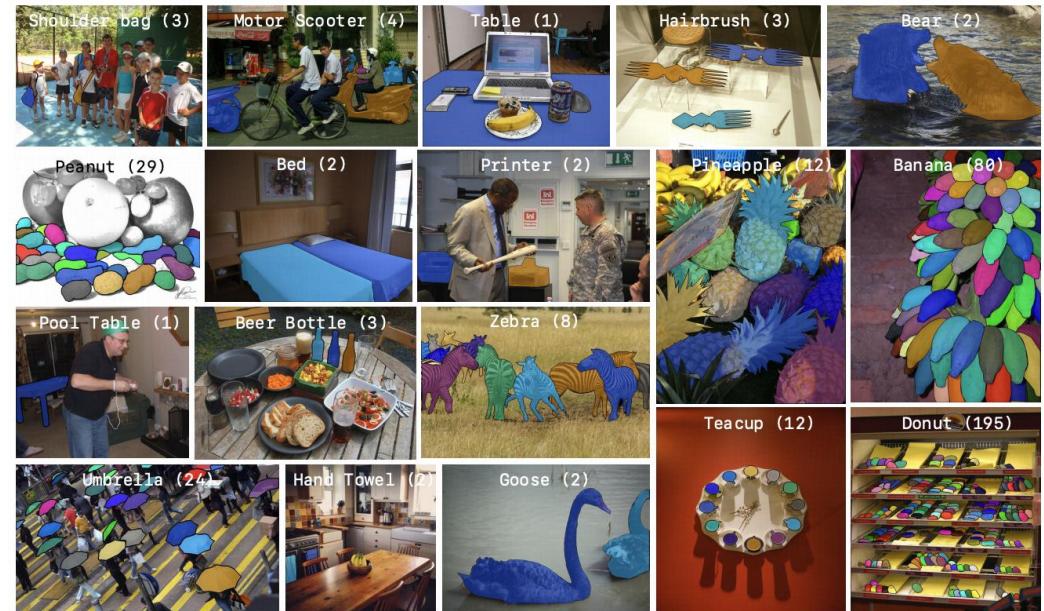


Figure 3. Example LVIS annotations (one category per image for clarity). See <http://www.lvisdataset.org/explore>.

[1] Gupta, Agrim, Piotr Dollar, and Ross Girshick. "Lvis: A dataset for large vocabulary instance segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Paper Candidates



● Grounding DINO

- ◆ Optional (recommend): Compare the performance of VLMs and DINO on the grounding task

Original Image (left) and Visualized Grounding (right)



Prompt: Tell me the position of the couple in the picture. The short-haired guy is wearing a pink top and blue shorts, and the girl is in a cyan dress. Answer in $[x1,y1,x2,y2]$ format.

Model Response:

<think> The image contains multiple people. We need to find the specific couple where the short - haired guy is wearing a pink top and blue shorts, and the girl is in a cyan dress. First, we scan through all the individuals in the image to identify the person matching the description of the short-haired guy with a pink top and blue shorts. Once we locate him, we then look for the girl in a cyan dress near him. After identifying both of them precisely within the image, we determine their bounding box coordinates in the $[x1,y1,x2,y2]$ format by measuring the left - most ($x1$), top - most ($y1$), right - most ($x2$), and bottom - most ($y2$) points of their combined area or individual areas if they are separate but clearly associated as a couple.</think><answer> The position of the couple in the picture, where the short-haired guy is wearing a pink top and blue shorts, and the girl is in a cyan dress, is [835,626,931,883]. </answer>

Paper Candidates



- Versatile Diffusion: Generate text, images and variations with one diffusion model.

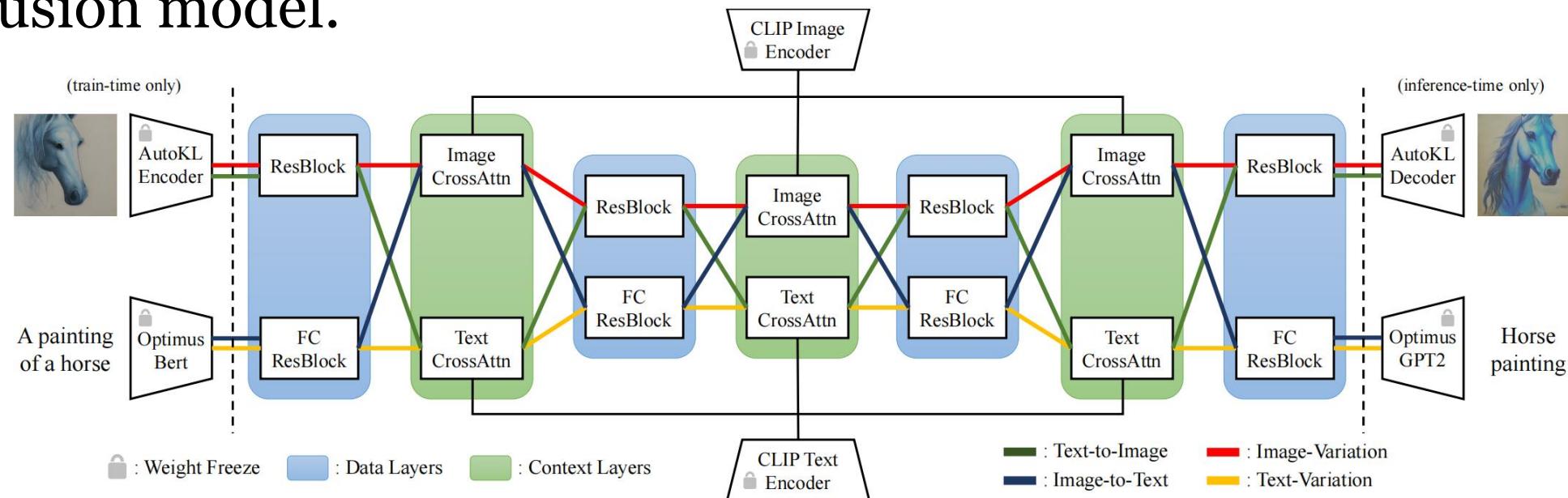


Figure 3: The overall structure of four-flow Versatile Diffusion (VD). Each color line depicts a single flow of VD that represents one supported task (*i.e.* green line for text-to-image). The VAE encoders at the far left are only used in training and are replaced with Gaussian noise inputs during inference. Oppositely, the VAE decoders at the far right are only used in inference for output generation, not train-time loss computation. For simplicity, we hide global layers in this figure. Better viewed in color.

Paper url: <https://arxiv.org/pdf/2211.08332.pdf>

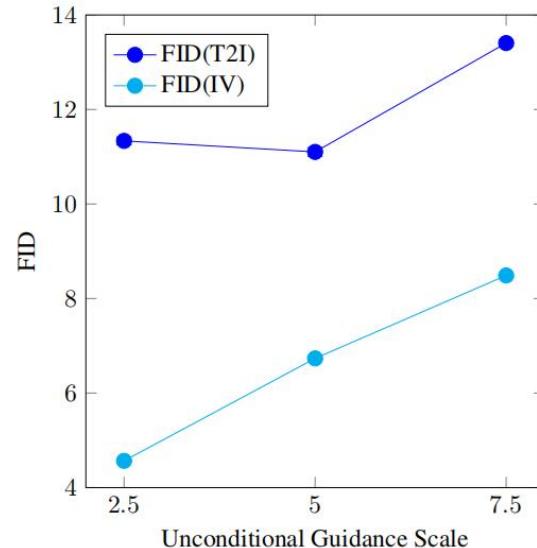
Official code: <https://github.com/shi-labs/versatile-diffusion>

Paper Candidates



- Task: Text-to-Image, Image-Variation (Quantitative, COCO-caption) and Image-to-Text (Qualitative)

Method	FID ↓
(A) Text-to-Image Synthesis	
CogView [18]	27.10
LAFITE [111]	26.94
GLIDE [62]	12.24
Make-a-Scene [23]	11.84
LDM [73]	12.63
SD (baseline)	11.21 ± 0.03
VD (four-flow)	11.10 ± 0.09
(B) Image-Variation Synthesis	
SD (baseline)	18.81 ± 0.06
VD (four-flow)	4.57 ± 0.02



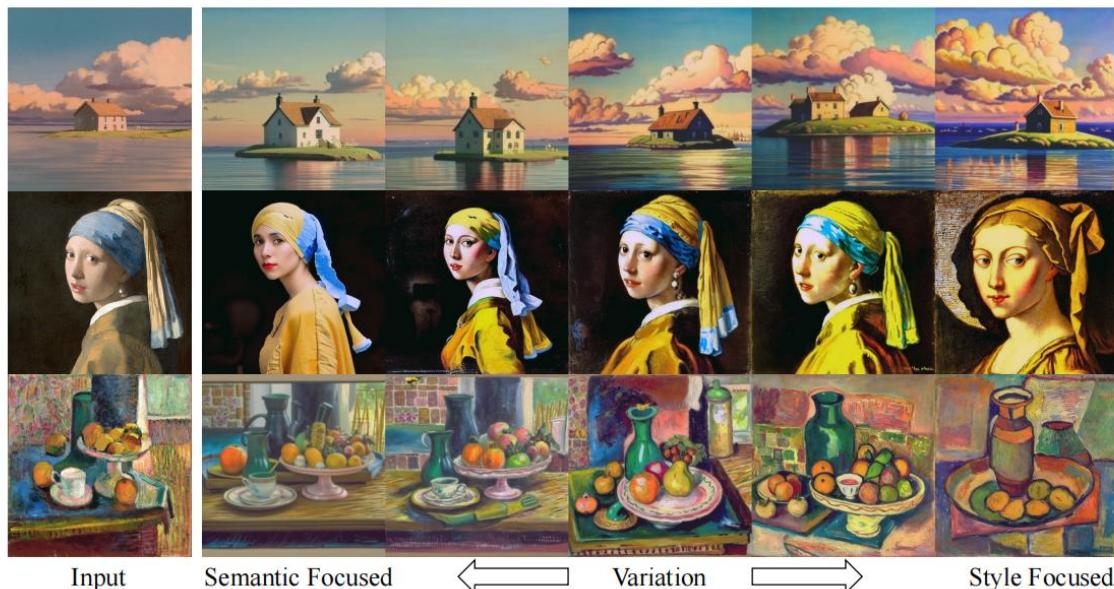
- Two young kids pointing at the sky.
- A girl and boy with a kite flying in the sky.
- Two children sitting near each other kids, pointing their fingers at the sun.
- Two young kids looking up at the sky.
- There are stars that a child is watching about.
- Two young girls and a boy standing near a star.
- Two young girls are watching a star.
- Kids standing for their starts.
- An old house on the water.
- An island with houses in it.
- A house in the water.
- An island in the sky.
- A pictures of a rocky houses on the shore of a lake with a Mountains silhouette on the horizon.
- Skyport, a houseboat on a rocky cliff, very near the shore, pictured and the shore.
- House on a lake, sky overlooking, and on the shore of the mountain.
- House on a lake by the sky.
- Fireworks and some people watching.
- Fireworks with lots of colored bursts.
- Fireworks and some clouds.
- Fireworks over the skyline and mountains.
- Fireworks at night.
- Fireworks night.
- A fireworks after huge fireworks.
- Fireworks bright lights at a party in downtown.

Figure 6: FID scores of VD comparing with baseline and prior approaches, and under various unconditional (classifier-free) guidance scales.

Paper Candidates



- Optional: Compare the performance of various versions of VD on Image-to-Text (Quantitative is better).
- Optional: Experiments for disentanglement of style and semantic, and/or dual- and multi-context blender.



Paper Candidate Summary



◎ NLG

- ◆ Sentence-BERT
- ◆ MEDUSA

◎ CV

- ◆ Grounding DINO
- ◆ Versatile Diffusion

Important Dates



- **Project Proposal:** 第8周，11月9日，星期日（网络学堂提交）
 - ◆ 做好预先调研，确定项目，确定额外任务的内容
- **Project Presentation :** Poster, 第16周（线下）
- **Final Report :** 第16周，1月4日，星期日（网络学堂提交）
- 请用中文写所有的报告，推荐使用Latex；助教会提供模板

Grading



- 40 points in total
- Project presentation (10 points)
 - ◆ Poster
- Open-source Repository (15 points)
 - ◆ Model Performance
 - Comparisons with the original paper
 - ◆ Codes & Instructions
 - Quick Start; Usage; Pretrained Model;
- Final Report (15 points)
 - ◆ Reproduction Details
 - ◆ The optional tasks (contribution & innovation)
 - ◆ Your workload should not be less than other projects

Research-Oriented Project



- **Apply what you have learned to a real problem of your interest**
- **A chance to show your research potential**
- **Each team consists of up to 3 students**
- Two types of projects are welcome:
 - ◆ **Applications:** Apply existing models to an existing or new problem
 - ◆ **Innovations:** Build a new model (algorithm) with NN, or a new variant of existing models
- It can be your research project undergoing (**before half complete**), but must relate to
 - ◆ **Neural Network, Deep Learning**

How to Choose a Good Project?



- **Significance and importance:** why do I choose this topic?
- **Novelty:** anything new?
- **Substances:** how much work did I do?
- **Results:** How well did I do?
 - ◆ Lessons, experiences learned?
 - ◆ New, interesting discoveries?
 - ◆ Inspiration to others? Any vision?

Overall Guideline – Proposal



- **Proposal** should be **at most 2** pages using the provided template. The following is a suggested structure for your proposal:
- **Task and Problem Definition:** Define what you will do precisely. Is it a well-defined problem? Try to use math symbols to formulate the task.
- **Dataset:** Existing dataset or create a new one? (**Collecting data should not be the main part of your workload, though it is very important in research**)
- **Challenges and Baselines:** What is difficult for you? State-of-the-art?
- **Proposal:** State your idea very clearly. Connect the idea with the math and even the codes
- **Feasibility:** State why this is possible for your team
 - ◆ Computation Cost? Did you run a baseline, implement a similar model?

Some Notes



- ◎ Necessary literature review
- ◎ It is hard to outperform baselines in some tasks
 - ◆ But the results of baselines are always needed
 - ◆ Novel ideas are more important
- ◎ Don't waste too much time on setting up environment
 - ◆ Use the existing tools
 - ◆ Only data labeling can't be your main part of work
- ◎ If someone in your group didn't do anything, you can report to TAs privately.

Important Dates



- **Project Proposal**: 第8周, 11月9日, 星期日 (网络学堂提交)
- **Project Presentation** : Poster, 第16周 (线下)
- **Final Report** : 第16周, 1月4日, 星期日 (网络学堂提交)

- 请用中文写所有的报告, 推荐使用Latex; 助教会提供模板

Grading



- 40 points in total
- Project presentation (10 points)
 - ◆ Poster
- Final submission (30 points)
 - ◆ Codes
 - ◆ Report

Reminders



- About scoring: The more you pay, the more you gain.
- Three points
 - ◆ Novelty
 - ◆ Substance
 - ◆ Results

Recommended Project Topics



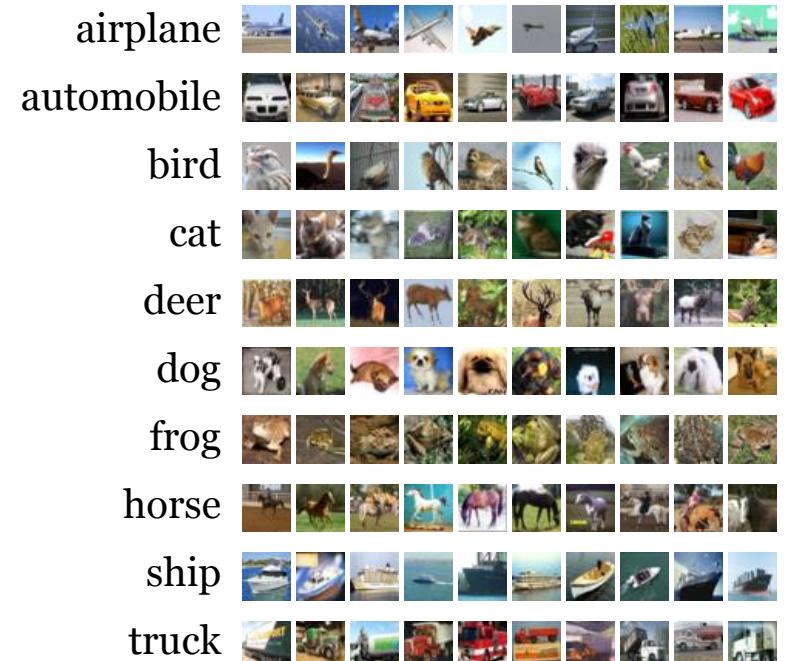
- CV
 - ◆ Image Recognition / Segmentation
 - ◆ Image Generation (Diffusion, ViT)
 - NLP
 - ◆ Text Classification
 - ◆ Reading Comprehension
 - ◆ Text Generation
 - ML
 - ◆ Zero / Few-shot Learning
 - ◆ Meta / Transfer Learning
 - ◆ Adversarial Attack / Defense
 - ◆ Pruning / Quantization
- Other Options
- RL theory is not the main concern of this course!
- ◆ Inform TA in advance if you want to conduct RL projects

You are not limited to these topics

Image Recognition



- Single label/Multi-label classification
- Well-studied, hard to improve performance
- But new ideas, new tasks are worth of study



Cifar-10 Dataset

Image Segmentation

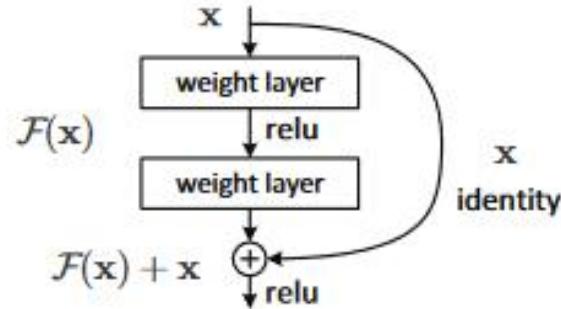


- Cluster parts of an image together which belong to the same object class

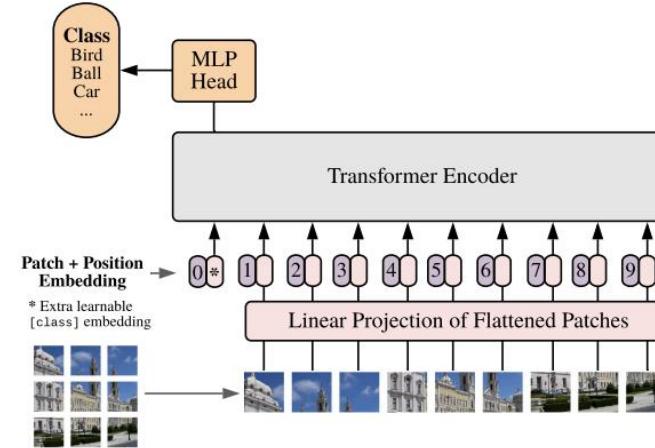


<https://paperswithcode.com/task/semantic-segmentation>

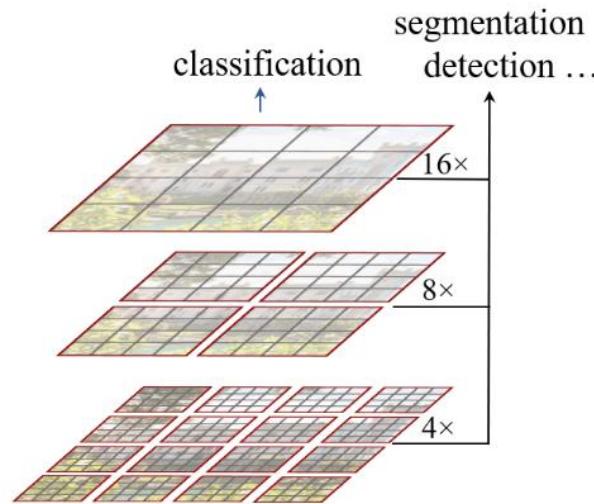
Image Recognition / Segmentation



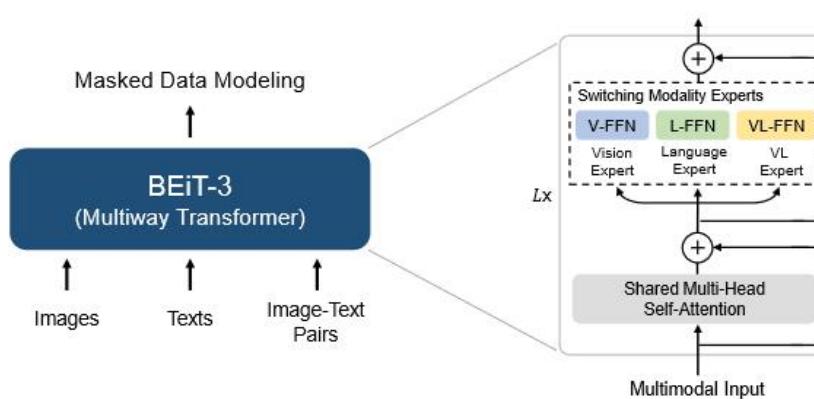
ResNet



Vision Transformer (ViT)



Swin-
Transformer



BEiT

Image Generation



- Generate a natural image
- Fill blanks in images
- Super resolution
- Text-to-image generation



Figure 1: Class-conditional samples generated by our model.

TEXT PROMPT an armchair in the shape of an avocado....

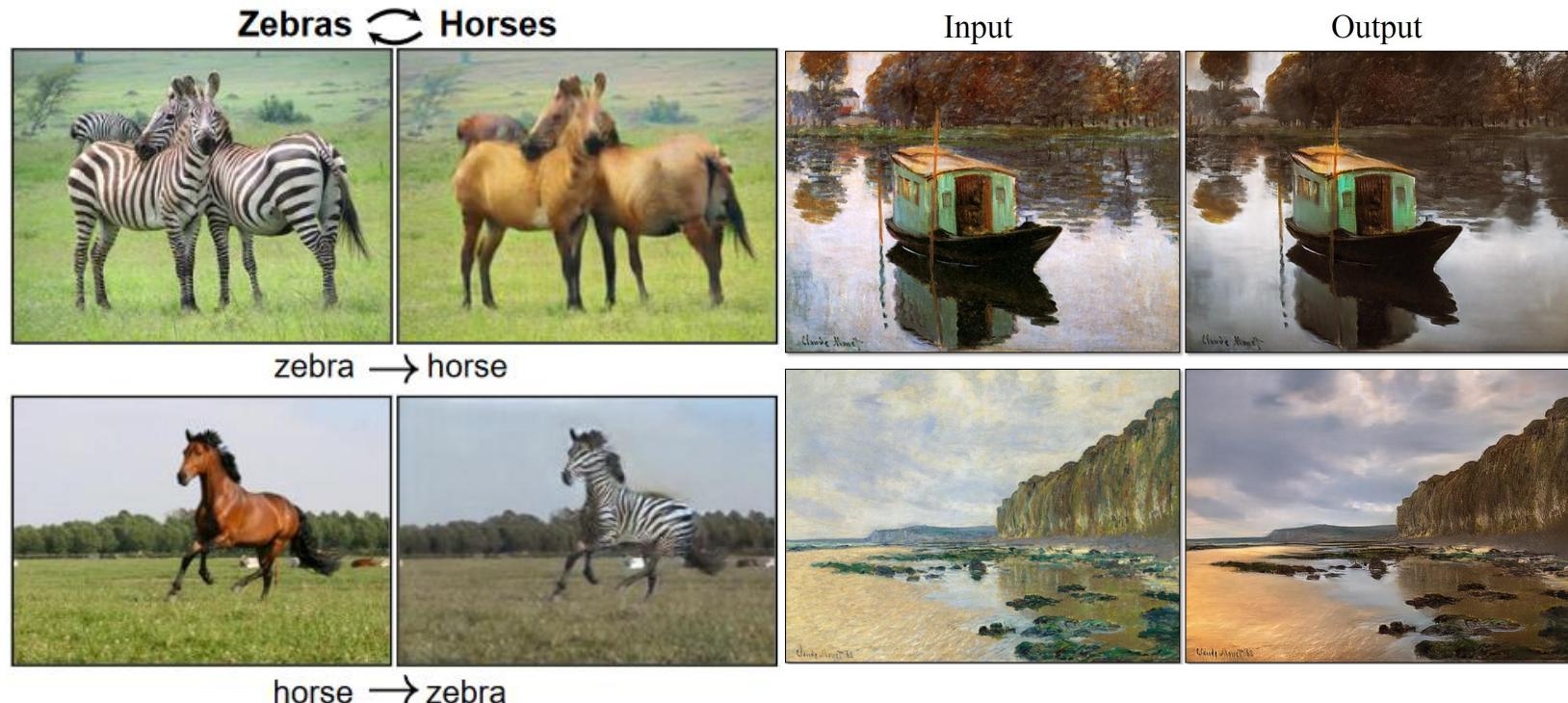
AI-GENERATED
IMAGES



Image Generation

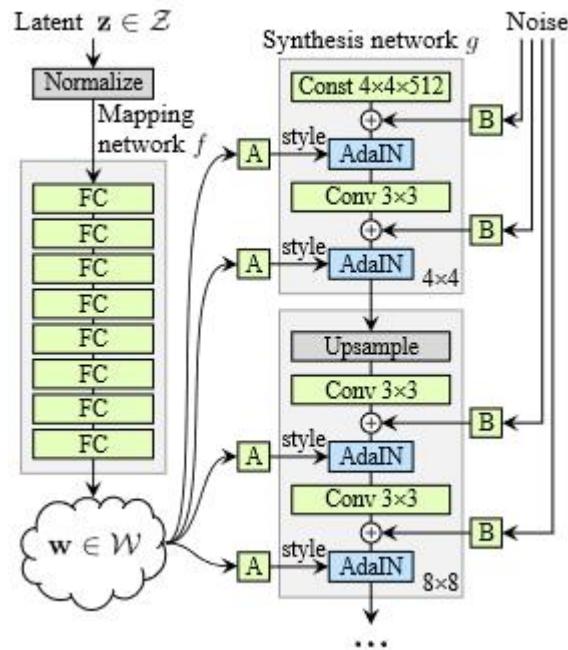


◎ Image Style Transfer

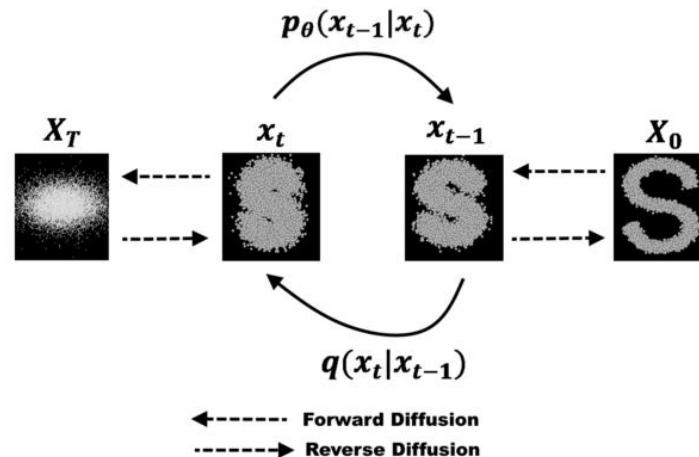


Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
Style Transfer for Anime Sketches with Enhanced Residual U-net and Auxiliary Classifier GAN

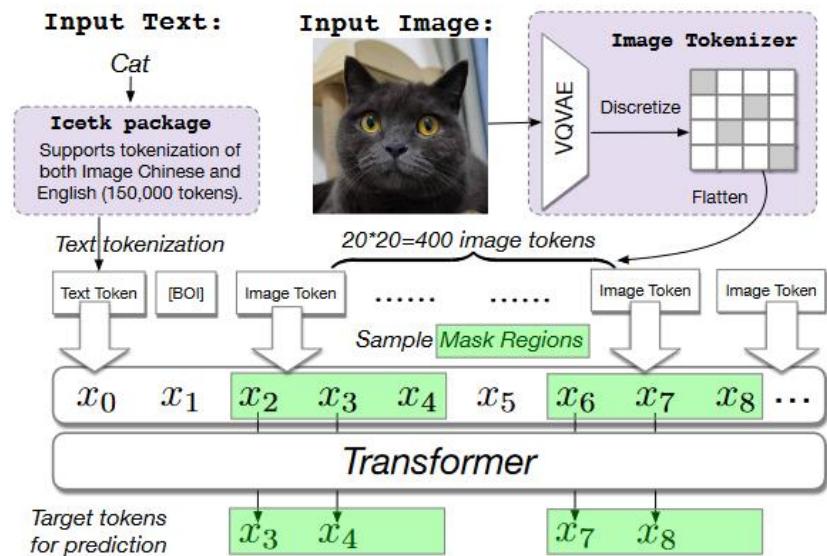
Image Generation Models



StyleGAN



Diffusion Model



Autoregressive Generation
Cogview / DALL-E

Text Classification



- Document-level classification
- Sentence-level classification

- Topic classification: sports, economics
- Sentiment classification:
 - ◆ Positive vs. negative
 - ◆ Subjective vs. objective
 - ◆ Emotion
- Relation classification

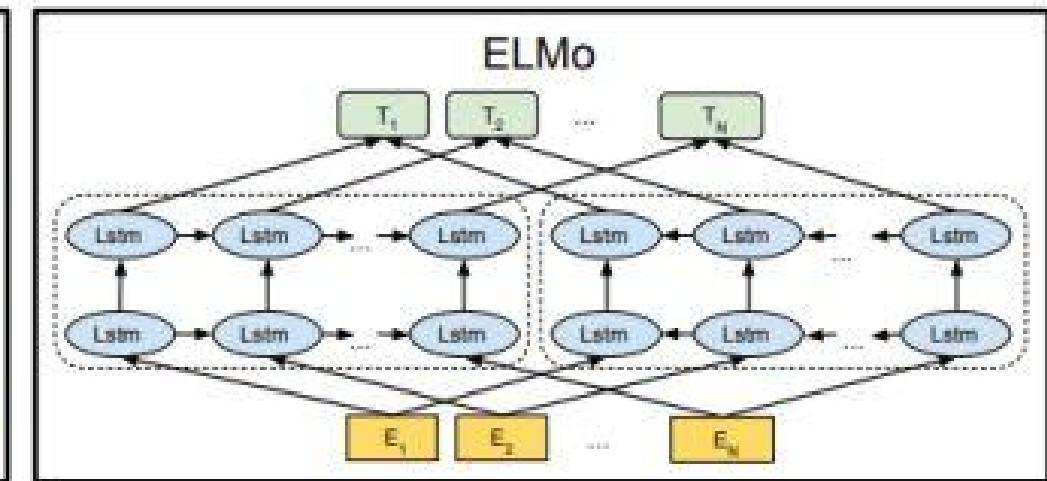
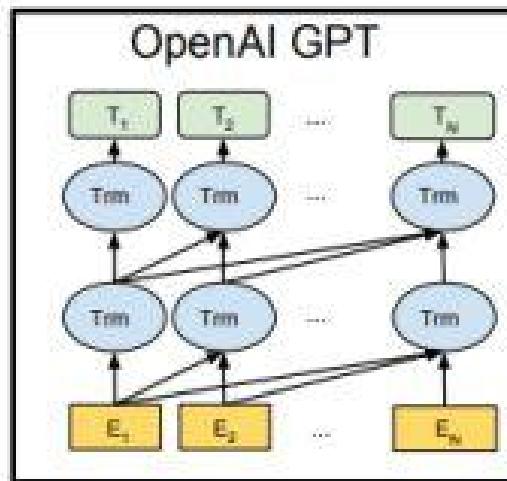
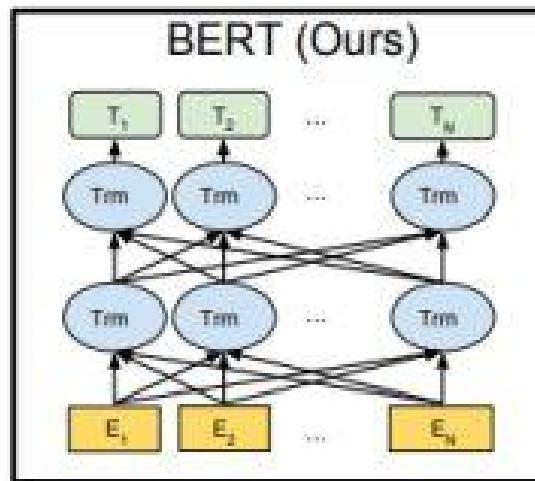
- Multi-class / Fine-grained classification
- Few-shot classification

Text Classification Models



○ Representation learning:

- ◆ BERT / GPT / Elmo



Datasets for Text Classification



- [YELP](#) / [IMDB](#) Large-scale datasets
- [GLEU](#) dataset Multi-task language dataset
- [Stanford Sentiment Treebank](#) (SST) dataset contains 11,855 sentences, and has split into the training/validation/test parts, respectively containing 8,544/1,101/2,210 sentences.
 - ◆ Specially, the dataset has provided phrase-level annotation on all inner nodes, so that it is more suitable for Tree-structured Model.

Reading Comprehension



- Single-document extractive RC

- ◆ natural language form questions
- ◆ answer is a word/span extracted from passages

- Cloze-style RC

- ◆ question is marked by a blank space;
- ◆ single-word answer from candidates is selected;

- Choice-style RC

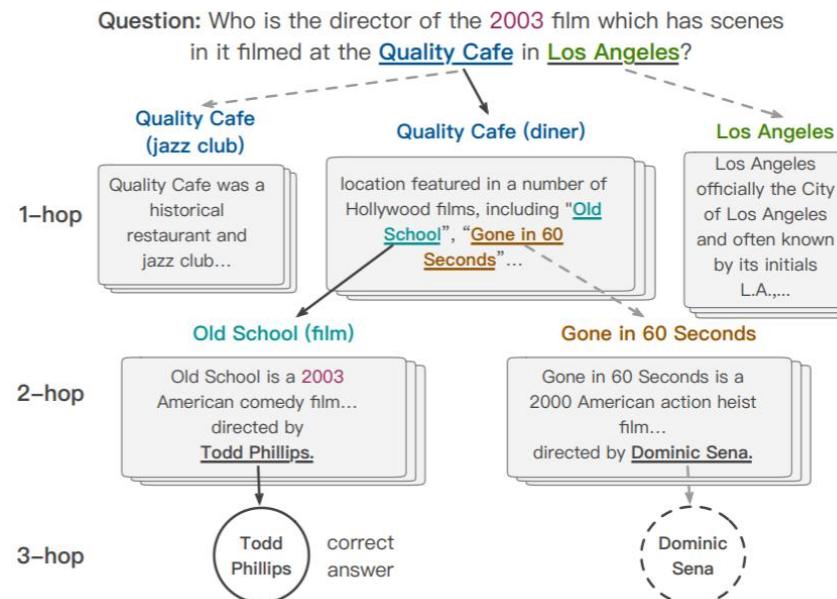
- ◆ natural language form questions
- ◆ Few answer sentences

Reading Comprehension



Multi-hop extractive RC

- ◆ natural language form questions
- ◆ answer is extracted (and then abstracted) based on passages
- ◆ need multi-hop reasoning on the reference passages



Reading Comprehension



◎ Commonsense reasoning RC

- ◆ natural language form questions
- ◆ some choices, one of which is plausible
- ◆ need external commonsense knowledge to help reasoning

Where on a river can you hold a cup upright to catch water on a sunny day?

✓ waterfall, X bridge, X valley, X pebble, X mountain

Where can I stand on a river to see water falling without getting wet?

X waterfall, ✓ bridge, X valley, X stream, X bottom

I'm crossing the river, my feet are wet but my body is dry, where am I?

X waterfall, X bridge, ✓ valley, X bank, X island

CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge (NAACL 2019)



○ Machine Translation

- ◆ Translating a sequence of words from Language A to Language B.

○ Summarization

- ◆ Summarize the key points of a document into one or few sentences.

○ Dialog Generation

- ◆ Chit-chat (with emotion, personality)
- ◆ Task-oriented AI (help you book tickets, etc.)

Text Generation



○ Visual Storytelling



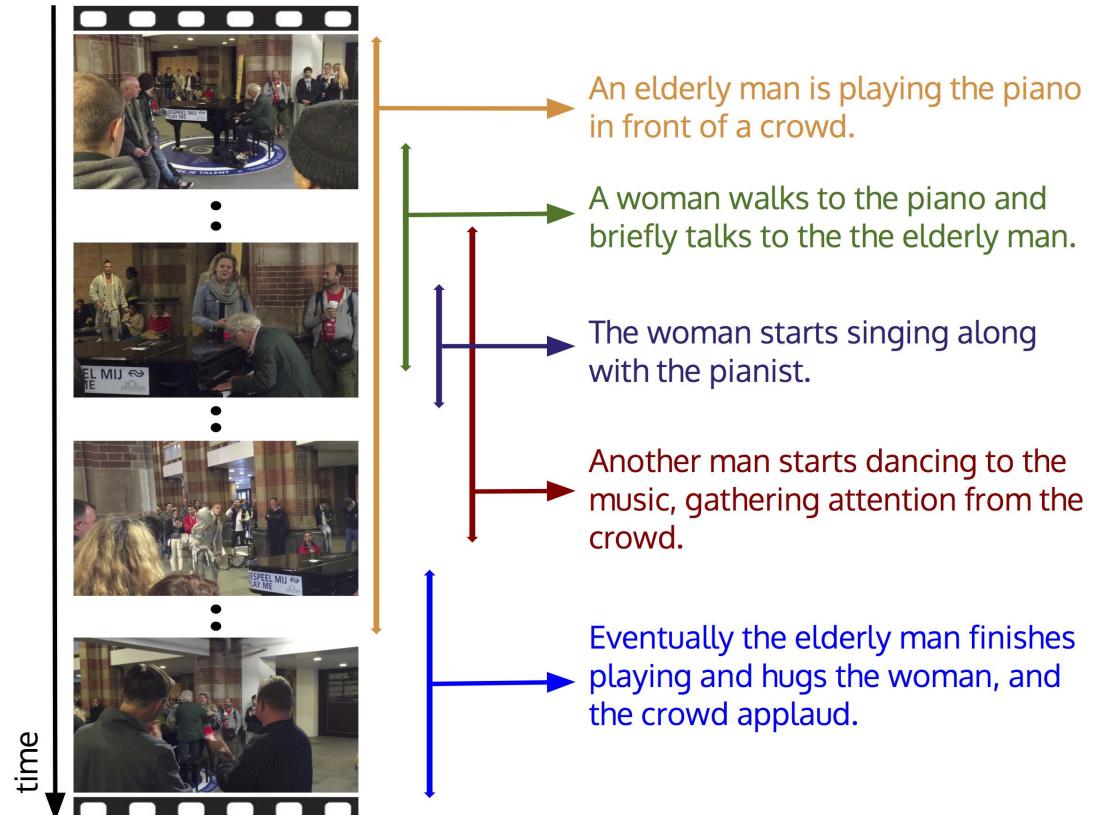
(a) (b) (c) (d) (e)

Captions:

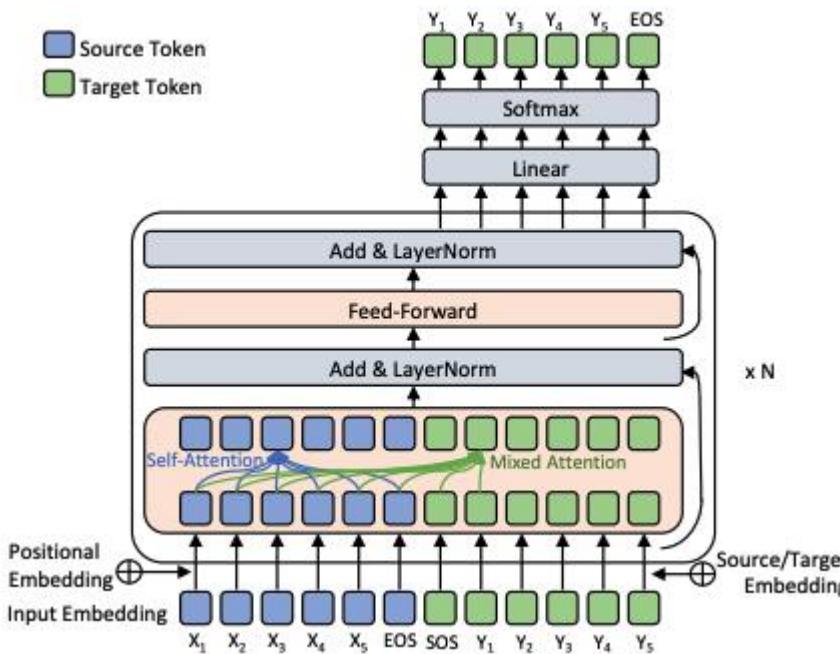
- (a) A small boy and a girl are sitting together.
- (b) Two kids sitting on a porch with their backpacks on.
- (c) Two young kids with backpacks sitting on the porch.
- (d) Two young children that are very close to one another.
- (e) A boy and a girl smiling at the camera together.

Story #1: The brother and sister were ready for the first day of school. They were excited to go to their first day and meet new friends. They told their mom how happy they were. They said they were going to make a lot of new friends . Then they got up and got ready to get in the car .

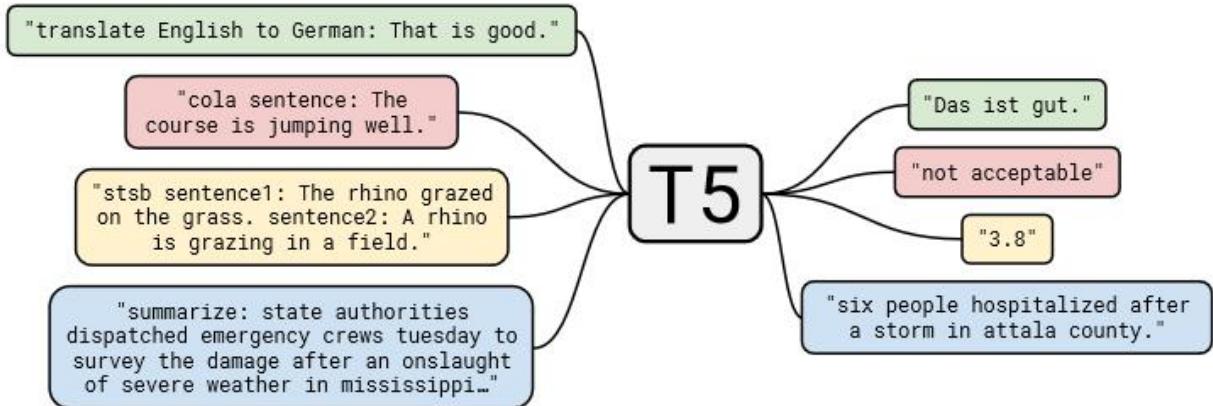
○ Video Captioning



Text Generation Models



GPT-2



T5

Datasets for Text Generation



- [Google 1 Billion Word Corpus](#) dataset makes available a standard corpus of reasonable size (0.8 billion words) to train and evaluate language models.
- [Workshop on Statistical Machine Translation](#) (WMT) dataset contains training data for five language pairs, and a common framework (including a baseline system). The task is to improve methods current methods.
- [Extreme Summarization \(XSum\)](#) contains 200 K document-summarization pairs collected from BBC and is designated for abstractive summarization.
- [Ubuntu Dialogue Corpus](#), a dataset containing almost 1 million multi-turn dialogues, with a total of over 7 million utterances and 100 million words.
- [WritingPrompts](#) contains 300K human-written stories with prompts, which is used for training models to generate coherent and fluent passages of text about a topic.

Datasets for Dialogue Generation (by CoAI)



- ⦿ **Commonsense Conversation Dataset** contains one-turn post-response pairs with the corresponding commonsense knowledge graphs. Each pair is associated with some knowledge graphs retrieved from ConceptNet.
- ⦿ **Knowledge-Driven Conversation Dataset** contains multi-turn dialogues with annotated knowledge triples extracted from knowledge graphs of various domains.
- ⦿ **Large-scale Chinese Dialogue Dataset** contains 12 million turns of clean dialogue data collected from Weibo and open-source data. We together release a large-scale pre-training dialogue model CDial-GPT.

<http://coai.cs.tsinghua.edu.cn/tools/>

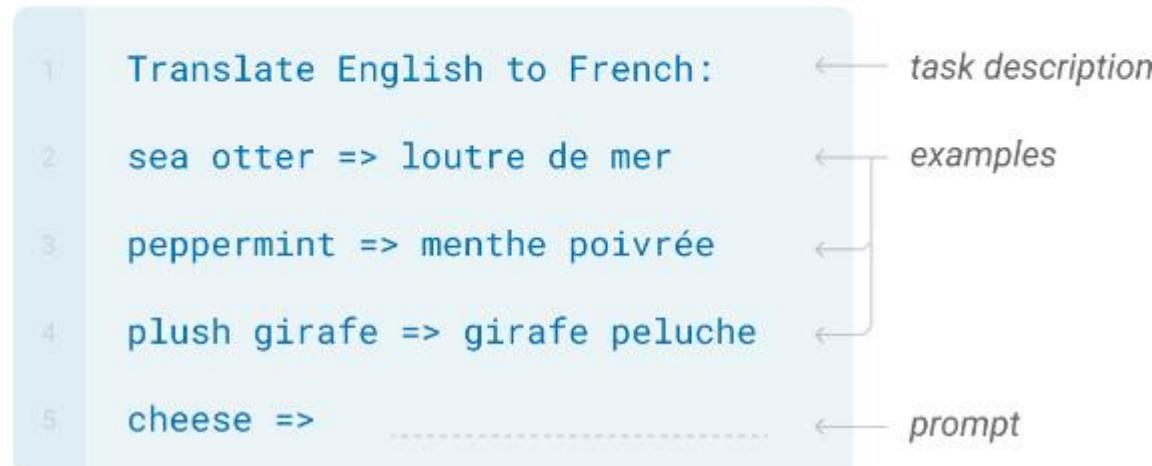
Zero / Few-shot Learning



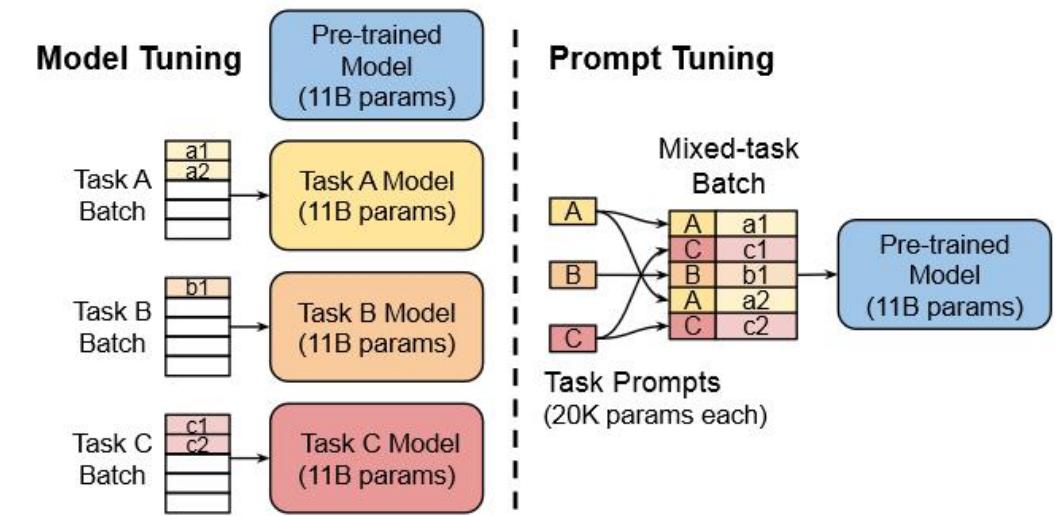
- Accomplish tasks with several or no training samples
 - ◆ Large-scale pretrained model + Prompt

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



In-context Learning



Prompt Tuning

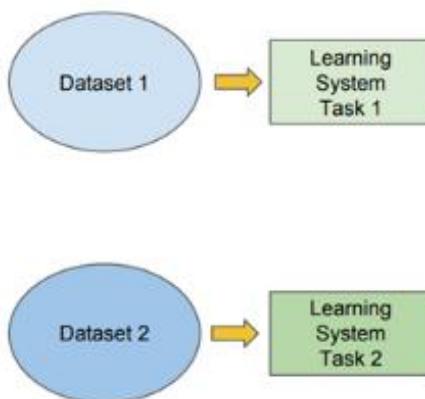
Meta / Transfer Learning



- Fast adaption to new datasets

Traditional ML

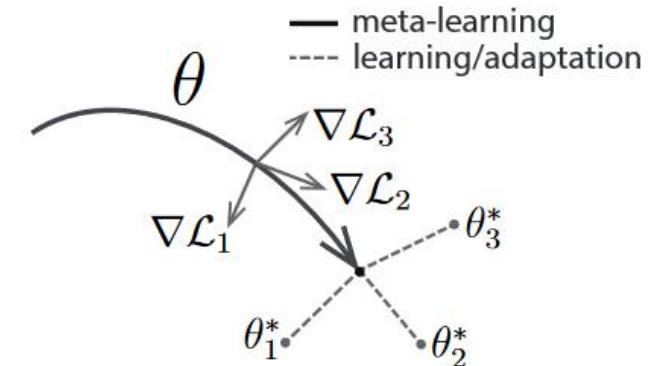
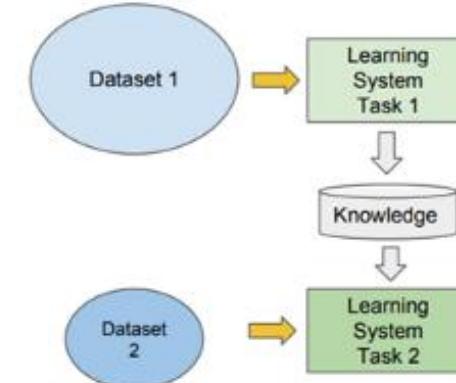
- Isolated, single task learning:
 - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks



vs

Transfer Learning

- Learning of a new tasks relies on the previous learned tasks:
 - Learning process can be faster, more accurate and/or need less training data

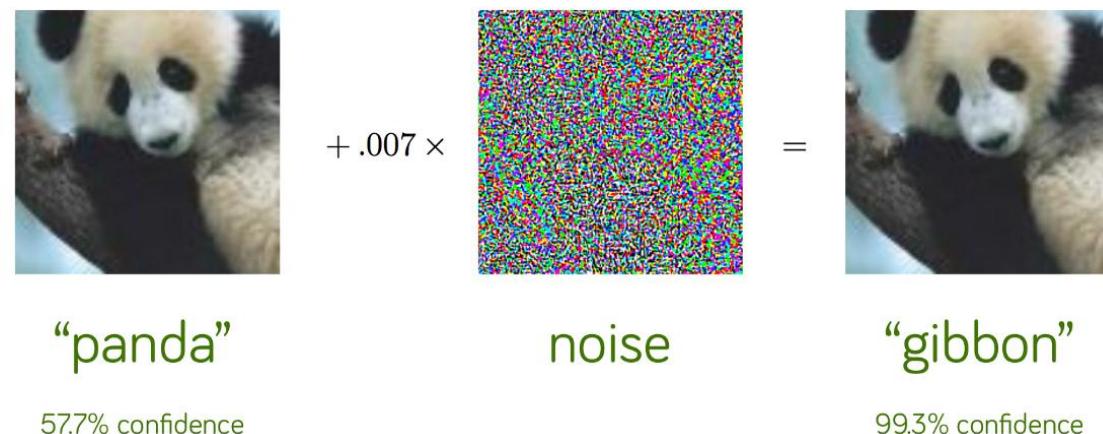


..model-agnostic meta-learning algorithm (MAML)

Adversarial Attack / Defense



- Disturb a sample with some noise and make model produce errors

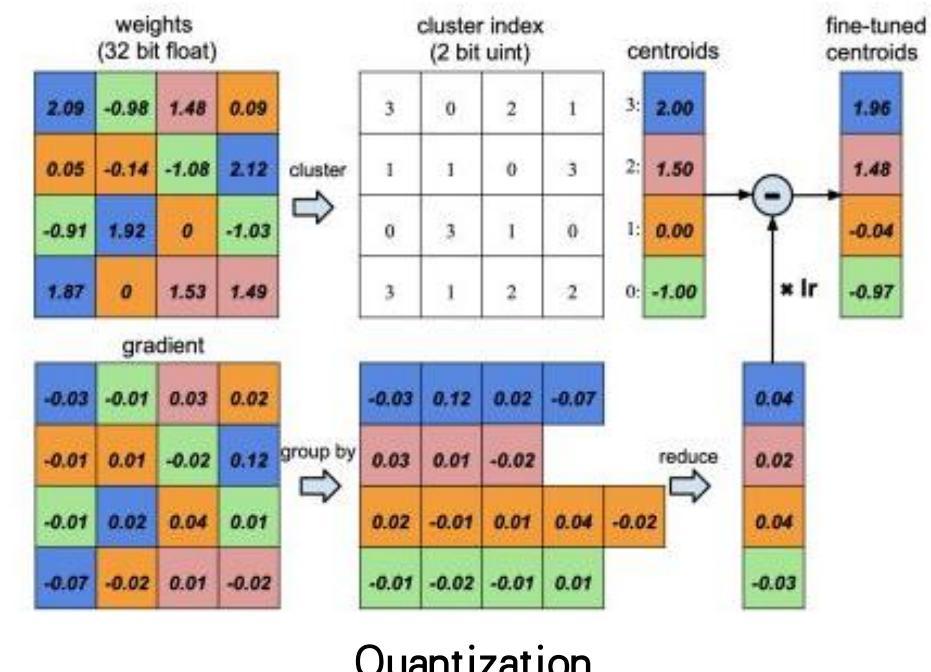
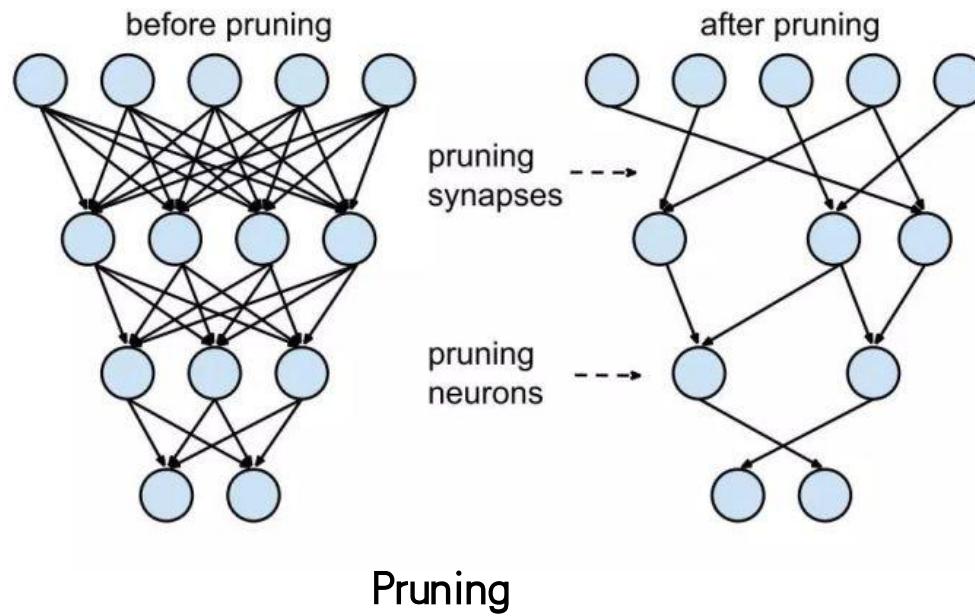


Original Prediction	Adversarial Prediction	Perturbed Texts
Positive Confidence = 96.72%	Negative Confidence = 74.78%	Ah man this movie was <i>funny (laughable)</i> as hell, yet strange. I like how they kept the shakespearean language in this movie, it just felt ironic because of how idiotic the movie really was. this movie has got to be one of troma's best movies. highly recommended for some senseless fun!

Pruning / Quantization



- Accelerate model by removing useless weights / reducing precision
 - Popular in industrial products ...



Other Options



- Actually neural network can do more than above, for example
- Deep reinforcement learning: Deepmind's Apollo project for general AI
-

Thanks for Your Attention



清华大学
Tsinghua University

