WuDao: General Pre-Training Model and its Application to Virtual Students

Tang Jie
Tsinghua University
Beijing Academy of Artificial Intelligence (BAAI)

The slides is available at http://keg.cs.tsinghua.edu.cn/jietang (or Google “Jie Tang”)
GPT-3

- **GPT-3**: OpenAI released GPT-3 in 2020/5 with super-scale parameters: 175B

- **Google’s Switch Transformer**: 1.6T

---

**GPT-3 for QA**

---

**DALL·E**: (Generating Images by Text)

- a snail made of harp
Open-domain Chatbot

- Xiaolce (2018)
- Meena by Google (2020)
- Blender by Facebook (2020)
Chatbot w/ Grounded Knowledge

[Background Description]
2022年北京冬季奥运会
abstract: 第24届冬季奥林匹克运动会，即2022年北京冬季奥运会，是由中国举办的国际性奥林匹克赛事，于2022年2月4日开幕，......
The 24th Winter Olympic Games, namely the Beijing 2022 Winter Olympic Games, is an international Olympic event held by China and opened on February 4, 2022, ......
The Power of Generative Models

**Domain-specific**

- **unconditional generation**
  - StyleGAN2 on FFHQ

**Domain-specific**

- **text-to-image generation**
  - DF-GAN on CUB
  - This bird has a white belly and breast, with a blue crown and nape.

**Domain-general**

- **text-to-image generation**
  - DF-GAN on COCO
  - Close-up of a man eating a piece of pizza while holding a plate.

CogView2

Virtual Student

Psychology

Knowledge Graph

Pretraining Model
Virtual Front Desk
AI driven Sign Language
AI driven Sign Language
比如你可以问我以下问题或者其他问题。如果你没有疑问，你可以跟我说“再见”或者点击“结束面试”按钮。
1：公司目前在全国各个项目团队都是多少人？
2：咱们公司现在这个岗位的五险一金是怎么交的？
3：这个项目组目前的盈利业务有哪几条线？
WUDAO 2.0 is one of the largest models

1.75 Trillion

10X larger than GPT-3 parameters
What is WuDao 2.0

01. 1.75 Trillion Parameters
02. both text and images
03. train on a supercomputer
04. Bilingual (Cn and En) data: 4.9T text and images

Largest
Universal
Domestic
Knowledge
What is WuDao 2.0

FastMoE

- Support complex balance strategies such as Switch and GShard
- Support different experts and different models.
- Deploy in Alibaba PAI platform
- Explore the application to the Alipay intelligent service system
- Training on a supercomputer

FastMoE
- Redesign all operators
- Efficient communication strategy
- MoE training with tens of thousands of experts

Z. Ma et al. BAGUALU: Targeting Brain Scale Pretrained Models with over 37 Million Cores. PPoPP’22.
Question: Which moment did you want to live in forever?

User: High-school student

WUDAO: In an instant I recall the second year of high school. That afternoon, I took a sunbath on the playground for a long time. I then played basketball downstairs with my classmates. As the sun set slowly, the three of us stood together on the corridor, leaning on the railing to face the sunset. I’ve always wanted to live in that moment forever, but unfortunately I never encountered a similar moment again. I just want to live in that moment.
作诗图灵测试

查找：Bitcoin

Easy  Hard  Lunatic  Extra

在作诗图灵测试的Easy模式中，您将会被展现5组诗歌（包括标题、作者及内容），每组包括1首由诗人创作的诗歌和1首AI创作的诗歌，请您选择您认为由人创作的诗歌。所有组选择完成后，您将会得知有多少组结果正确。
外挖无穷洞，机神犹未休。
卡中窥币影，池里验沙流。
屡载吸金主，孤深渍盗求。
方知区块链，本是古来游。

Mining an infinite hole outside
The god of machine is still not satisfied
Illusion of bitcoins seen inside the GPU
Verified in pools to confirm the workflow
Several years it has attracted many investors
Deep and limited, even the robbers pursue
Eventually it turns out that Blockchain
May only be a financial trick from ancient

https://wudao.aminer.cn/turing-test/v2/
<table>
<thead>
<tr>
<th>文源</th>
<th>文澜</th>
<th>文溯</th>
<th>文汇</th>
</tr>
</thead>
<tbody>
<tr>
<td>磴溪连灞水，商岭接秦山。青汉不回驾，白云长掩关。雀喧知鹤静，凫戏识鸥闲。却笑南昌尉，悠悠城市间。</td>
<td>晨趋禁掖暮郊园，松桂苍苍烟露繁。明月上时群动息，雪峰高处正当轩。</td>
<td>初辞云浦寂，再别竺岩难。各分当时意，空驰海远寒。依稀真弟子，守至四明涯。临别殷勤语，重违誓去盘。</td>
<td>近年容尚白，卿士累资名。投契顾怜旧，照乘拟次行。常嫌官爵小，更睹道心清。多谢白公意，作为肺腑情。</td>
</tr>
<tr>
<td>文源</td>
<td>文澜</td>
<td>文溯</td>
<td>文汇</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>老翁岩畔卧，数载避喧惊。棋局尘生涧，茶烟景近屏。刀枪犹野外，吟啸亦琴清。更有深交意，相期岁晚行。</td>
<td>罗隐虽来频道寺，篷门未肯便归休。殷勤与见我无倦，竟日相随不暂留。</td>
<td>上国随缘住，来途若梦行。浮天沧海远，去世法舟轻。水月通禅观，鱼龙听梵声。惟怜一灯影，万里眼中明。</td>
<td>太空秋色凉，独鸟下微阳。三径池塘静，六街车马忙。渐能高酒户，始是入诗狂。官冷且无事，追陪慎莫忘。</td>
</tr>
</tbody>
</table>

https://wudao.aminer.cn/turing-test/v2/
**Draw Pictures**

https://wudao.aminer.cn/CogView/index.html

- A woman in a black and purple dress poses in front of some tall grass.
- A woman is on a bench overlooking the city.
- A couple of young boys playing a game of soccer.
- A man that is on a surfboard in some water.
- A woman in a white blouse is holding a remote in her hands.
- A bird perched on top of a leafless tree under a blue sky.
- A clock hanging outside of a house in a nice neighborhood.
- A red bus is driving on the road.
- A beautiful young blond woman talking on a phone.
- A red bowl filled with food and leafy greens.
• Draw Pictures-Image completion

• A girl with a surgical mask
• A girl wear a tie
• **Draw Pictures-Image completion**

  • A man with a red ball
  • 一个胖子在吃一碗面
Turing test

https://wudao.aminer.cn/turing-test/v2/
Core Algorithms

Controllable Generation via Inverse Prompting

CogView: Text-to-Image Generation

CogVideo: Text-to-Video Generation

GLM: General Language Model Pretraining

How to build a trillion-scale model
GLM: General Language Model Pretraining with Autoregressive Blank Infilling

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL'22.
Pretrained LMs and NLP Tasks

None of the pretraining frameworks performs the best for all tasks.

- Autoregressive model: GPT, GPT-2, GPT-3
- Autoencoding model: BERT, RoBERTa, ALBERT
- Encoder-Decoder model: MASS, BART, PALM

**Framework** | **NLU** | **Cond. Gen.** | **Uncond. Gen.**
--- | --- | --- | ---
Autoregressive | ✗ | ✗ | ✓
Autoencoding | ✓ | ✗ | ✗
Encoder-Decoder | ✗ | ✓ | ✗
GLM | ✓ | ✓ | ✓

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL'22.
All NLP Tasks are Generations

Figure 3. GLM finetune framework. (a) Formulation of the sentiment classification task as blank infilling with GLM. (b) GLM for text generation given the context. This can be the language modeling in the zero-shot setting, or seq2seq with fine-tuning.

NLU, Cond. Gen, Uncond. Gen can be unified into the GLM generation framework

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL'22.
A New Pretraining Framework

Figure 1. Illustration of GLM. We blank out text spans (green part) and GLM is trained to generate them in an autoregressive fashion.

Multi-task pretraining
1. Sample 15% in the middle as the generation objective
2. Sample 50-100% in the end as the generation objective
GLM: Autoregressive Blank Filling

(a) Sample spans from the input text

Part A: $x_1$ $x_2$ [M] $x_4$ [M]

Part B: $x_5$ $x_6$ $x_3$

(b) Divide the input into Part A and Part B

(c) Generate the Part B spans autoregressively

GLM (Transformer w/ masked self-attention)

Position 1: $1$ $2$ $3$ $4$ $5$ $5$ $5$ $3$ $3$
Position 2: $0$ $0$ $0$ $0$ $1$ $2$ $3$ $1$ $2$

(d) Self-attention mask
Table 2. Results on the SuperGLUE dev set. Models with * are pre-trained for two times the number of steps of other methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>ReCoRD F1/Acc.</th>
<th>COPA Acc.</th>
<th>WSC Acc.</th>
<th>RTE Acc.</th>
<th>BoolQ Acc.</th>
<th>WiC Acc.</th>
<th>CB F1/Acc.</th>
<th>MultiRC F1/a/EM</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT_Base</strong></td>
<td>65.4/64.9</td>
<td>66.0</td>
<td>65.4</td>
<td>70.0</td>
<td>74.9</td>
<td><strong>68.8</strong></td>
<td>70.9/76.8</td>
<td>68.4/21.5</td>
<td>66.1</td>
</tr>
<tr>
<td><strong>GLM_Base</strong></td>
<td><strong>73.5/72.8</strong></td>
<td><strong>71.0</strong></td>
<td><strong>72.1</strong></td>
<td><strong>71.2</strong></td>
<td><strong>77.0</strong></td>
<td>64.7</td>
<td><strong>89.5/85.7</strong></td>
<td><strong>72.1/26.1</strong></td>
<td><strong>70.7</strong></td>
</tr>
<tr>
<td><strong>BERT_Large</strong></td>
<td>76.3/75.6</td>
<td>69.0</td>
<td>64.4</td>
<td>73.6</td>
<td>80.1</td>
<td><strong>71.0</strong></td>
<td>94.8/92.9</td>
<td>71.9/24.1</td>
<td>72.0</td>
</tr>
<tr>
<td>UniLM_Large</td>
<td>80.0/79.1</td>
<td>72.0</td>
<td>65.4</td>
<td>76.5</td>
<td>80.5</td>
<td>69.7</td>
<td>91.0/91.1</td>
<td>77.2/38.2</td>
<td>74.1</td>
</tr>
<tr>
<td>GLM_Large</td>
<td>81.7/81.1</td>
<td>76.0</td>
<td><strong>81.7</strong></td>
<td>74.0</td>
<td><strong>82.1</strong></td>
<td>68.5</td>
<td>96.1/94.6</td>
<td><strong>77.1/36.3</strong></td>
<td>77.0</td>
</tr>
<tr>
<td>GLM_Large (multi-task)</td>
<td>80.2/79.6</td>
<td>77.0</td>
<td>78.8</td>
<td>76.2</td>
<td>79.8</td>
<td>63.6</td>
<td><strong>97.3/96.4</strong></td>
<td>74.6/32.1</td>
<td>75.7</td>
</tr>
<tr>
<td>GLM_410M (multi-task)</td>
<td>81.5/80.9</td>
<td>80.0</td>
<td><strong>81.7</strong></td>
<td><strong>79.4</strong></td>
<td>81.9</td>
<td>69.0</td>
<td>93.2/96.4</td>
<td>76.2/35.5</td>
<td>78.0</td>
</tr>
<tr>
<td>GLM_515M (multi-task)</td>
<td><strong>82.3/81.7</strong></td>
<td><strong>85.0</strong></td>
<td><strong>81.7</strong></td>
<td>79.1</td>
<td>81.3</td>
<td>69.4</td>
<td>95.0/96.4</td>
<td>77.2/35.0</td>
<td><strong>78.8</strong></td>
</tr>
<tr>
<td><strong>T5_Base</strong></td>
<td>76.2/75.4</td>
<td>73.0</td>
<td>79.8</td>
<td>78.3</td>
<td>80.8</td>
<td>67.9</td>
<td>94.8/92.9</td>
<td>76.4/40.0</td>
<td>76.0</td>
</tr>
<tr>
<td><strong>T5_Large</strong></td>
<td>85.7/85.0</td>
<td>78.0</td>
<td><strong>84.6</strong></td>
<td>84.8</td>
<td>84.3</td>
<td>71.6</td>
<td>96.4/98.2</td>
<td>80.9/46.6</td>
<td>81.2</td>
</tr>
<tr>
<td><strong>BART_Large</strong></td>
<td>88.3/87.8</td>
<td>60.0</td>
<td>65.4</td>
<td>84.5</td>
<td>84.3</td>
<td>69.0</td>
<td>90.5/92.9</td>
<td>81.8/48.0</td>
<td>76.0</td>
</tr>
<tr>
<td>RoBERTa_Large*</td>
<td>89.0/88.4</td>
<td><strong>90.0</strong></td>
<td>63.5</td>
<td>87.0</td>
<td><strong>86.1</strong></td>
<td><strong>72.6</strong></td>
<td>96.1/94.6</td>
<td><strong>84.4/52.9</strong></td>
<td>81.5</td>
</tr>
<tr>
<td><strong>GLM_RoBERTa</strong></td>
<td><strong>89.6/89.0</strong></td>
<td>82.0</td>
<td>83.7</td>
<td><strong>87.7</strong></td>
<td>84.7</td>
<td>71.2</td>
<td><strong>98.7/98.2</strong></td>
<td>82.4/50.1</td>
<td><strong>82.9</strong></td>
</tr>
</tbody>
</table>

- Better than BERT，T5，RoBERTa
Before
Train three different models

After
1.25 X Larger GLM can do all the three tasks with one model!

Table 3. Results on Gigaword abstractive summarization

<table>
<thead>
<tr>
<th>Model</th>
<th>RG-1</th>
<th>RG-2</th>
<th>RG-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASS</td>
<td>37.7</td>
<td>18.5</td>
<td>34.9</td>
</tr>
<tr>
<td>UniLM_{Large}</td>
<td>38.5</td>
<td>19.5</td>
<td>35.8</td>
</tr>
<tr>
<td>GLM_{Large}</td>
<td>38.6</td>
<td>19.7</td>
<td>36.0</td>
</tr>
<tr>
<td>GLM_{Large} (multi-task)</td>
<td>38.5</td>
<td>19.4</td>
<td>35.8</td>
</tr>
<tr>
<td>GLM_{410M} (multi-task)</td>
<td><strong>38.9</strong></td>
<td><strong>20.0</strong></td>
<td><strong>36.2</strong></td>
</tr>
</tbody>
</table>

Table 4. Zero-shot language modeling results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lambada (Accuracy)</th>
<th>BookWiki (Perplexity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM_{Large} (uni)</td>
<td>0.0</td>
<td>&gt; 100</td>
</tr>
<tr>
<td>GLM_{Large} (multi-task,uni)</td>
<td>47.4</td>
<td>15.1</td>
</tr>
<tr>
<td>− 2d positional encoding</td>
<td>45.8</td>
<td>15.1</td>
</tr>
<tr>
<td>GLM_{410M} (multi-task,uni)</td>
<td>49.5</td>
<td>14.5</td>
</tr>
<tr>
<td>GLM_{515M} (multi-task,uni)</td>
<td><strong>50.4</strong></td>
<td><strong>13.9</strong></td>
</tr>
<tr>
<td>GLM_{Large} (bi)</td>
<td>10.6</td>
<td>&gt; 100</td>
</tr>
<tr>
<td>GLM_{Large} (multi-task,bi)</td>
<td>48.5</td>
<td>14.9</td>
</tr>
<tr>
<td>− 2d positional encoding</td>
<td>47.3</td>
<td>15.0</td>
</tr>
<tr>
<td>GLM_{410M} (multi-task,bi)</td>
<td><strong>53.5</strong></td>
<td><strong>14.3</strong></td>
</tr>
<tr>
<td>GLM_{515M} (multi-task,bi)</td>
<td><strong>54.9</strong></td>
<td><strong>13.7</strong></td>
</tr>
<tr>
<td>GPT_{Large} (uni)</td>
<td>50.1</td>
<td>14.4</td>
</tr>
</tbody>
</table>
CogView: Mastering Text-to-Image Generation via Transformers.

CogView: Text-to-Image Generation

- CogView: 4B、6B
- Training with 30M image-text pairs on 512 V100 for 400 hours

![Diagram of Text-to-Image Generation]

Input Text:
(The head of a lovely cat.)
一只可爱的小猫的头像。

Text Tokenizer (sentence pieces)
Text tokens, ranging from 8192 to 58192.

Input Image:

Transformer (GPT)

Image Tokenizer (VQAE)
1024 Image tokens, ranging from 0 to 8191.
Instability in Pretraining: Solutions

Sandwich LayerNorm
Block up the layer-wise magnification.

- Precision-Bottleneck relaxation
- Make the computation of attention and LayerNorm precision-friendly.

DALL-E uses another solution, perResblock loss scaling and throw back some parameters back to fp32.
Towards higher resolution

The Big Ben

- Since CogView is trained on the most complex distribution of domain-general images, details of the objects have already been covered.

- Finetuning it into an super-resolution model should not be hard. (16*16 tokens => 32*32 tokens, 1 DGX*day)

(a) Center-continuous sliding window

(b) Different super-resolution results for “a tiger is playing football”.
Results

37% of the users favor the generated images by CogView better than DALL-E.

Codes and Models: https://github.com/THUDM/CogView
Demo website: http://wudao.aminer.cn/CogView/index.html

Table 1: Metrics for machine evaluation. Statistics about DALL-E are extracted from their figures. FID-k means that all the images are blurred by a Gaussian Filter with radius k.

<table>
<thead>
<tr>
<th>Model</th>
<th>FID-0</th>
<th>FID-1</th>
<th>FID-2</th>
<th>FID-4</th>
<th>FID-8</th>
<th>IS</th>
<th>CapS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttnGAN</td>
<td>35.2</td>
<td>44.0</td>
<td>72.0</td>
<td>108.0</td>
<td>100.0</td>
<td>23.3</td>
<td>0.02763</td>
</tr>
<tr>
<td>DM-GAN</td>
<td>26.0</td>
<td>39.0</td>
<td>73.0</td>
<td>119.0</td>
<td>112.3</td>
<td>32.2</td>
<td>0.02801</td>
</tr>
<tr>
<td>DF-GAN</td>
<td>26.0</td>
<td>33.8</td>
<td>55.9</td>
<td>91.0</td>
<td>97.0</td>
<td>18.7</td>
<td>0.02802</td>
</tr>
<tr>
<td>DALL-E</td>
<td>27.5</td>
<td>28.0</td>
<td>45.5</td>
<td>83.5</td>
<td>85.0</td>
<td>17.9</td>
<td>—</td>
</tr>
<tr>
<td>CogView</td>
<td>27.1</td>
<td>19.4</td>
<td>13.9</td>
<td>19.4</td>
<td>23.6</td>
<td>18.2</td>
<td>0.17403</td>
</tr>
</tbody>
</table>
CogView2

**Input Text:**
(The head of a lovely cat.)

**Input Image:**

![Image Tokenizer (sentence pieces)](image)

**Transformer**

Hidden representation

Mix Tokens

Text

Text Tokens
(sentencepiece, Cn-En)
(0-127)

Image

Image Token (VQ-VAE)
(200128-208319)

All NLP tasks are generation tasks

All [MASK] are generation tasks [START] NLP tasks
CogView2: Cross-Modality General Language Model

- A unification of “GPT(CogView) + MAE + Image captioning”
- Simplicity. No sentinel tokens, no order changing or token replacement.
- Only 20 * 20 tokens... 6 billion parameters!
- As a basic for future uses (super-resolution / video generation)!
CogView2: Faster *(LoPAR)* and better *(Bidirectional/higher-resolution)* text-to-image generation
Results

Table 1: Machine Evaluation Results on MS-COCO. (Downsampling CogView2 images to 256×256.) “*” means fine-tuning on MS-COCO. “—” technique is the ablation study without this technique. CogView2 achieves the best blurred FIDs over all comparable methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>FID-0</th>
<th>FID-1</th>
<th>FID-2</th>
<th>FID-4</th>
<th>FID-8</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttnGAN* [35]</td>
<td>35.2</td>
<td>44.0</td>
<td>72.0</td>
<td>108.0</td>
<td>100.0</td>
<td>23.3</td>
</tr>
<tr>
<td>DM-GAN* [40]</td>
<td>26.0</td>
<td>39.0</td>
<td>73.0</td>
<td>119.0</td>
<td>112.3</td>
<td>32.2</td>
</tr>
<tr>
<td>DF-GAN* [28]</td>
<td>26.0</td>
<td>33.8</td>
<td>55.9</td>
<td>91.0</td>
<td>97.0</td>
<td>18.7</td>
</tr>
<tr>
<td>DALL-E [26]</td>
<td>27.5</td>
<td>28.0</td>
<td>45.5</td>
<td>83.5</td>
<td>85.0</td>
<td>17.9</td>
</tr>
<tr>
<td>CogView [3]</td>
<td>27.1</td>
<td>19.4</td>
<td>13.9</td>
<td>19.4</td>
<td>23.6</td>
<td>18.2</td>
</tr>
<tr>
<td>XMC-GAN* [36]</td>
<td>9.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30.5</td>
</tr>
<tr>
<td>NUWA* [33]</td>
<td>12.9</td>
<td>13.8</td>
<td>15.7</td>
<td>19.3</td>
<td>24</td>
<td>27.2</td>
</tr>
<tr>
<td>LAFTE [39]</td>
<td>26.9</td>
<td>23.0</td>
<td>18.7</td>
<td>15.7</td>
<td>14.8</td>
<td>26.0</td>
</tr>
<tr>
<td>Make-A-Scene* [8]</td>
<td>7.55</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DALL-E-2 [27]</td>
<td>10.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CogView2</td>
<td>24.0</td>
<td>19.7</td>
<td>16.8</td>
<td>17.2</td>
<td>17.2</td>
<td>22.4</td>
</tr>
<tr>
<td>– clustering sampling</td>
<td>36.4</td>
<td>32.4</td>
<td>28.9</td>
<td>28.5</td>
<td>30.4</td>
<td>18.8</td>
</tr>
<tr>
<td>– attention upweighting</td>
<td>24.6</td>
<td>20.4</td>
<td>17.5</td>
<td>17.9</td>
<td>18.9</td>
<td>21.1</td>
</tr>
<tr>
<td>CogView2*</td>
<td>17.5</td>
<td>13.4</td>
<td>10.9</td>
<td>10.6</td>
<td>10.4</td>
<td>25.2</td>
</tr>
</tbody>
</table>

Figure 7: The results of human evaluation. CogView2 performs the best in all the aspects.
ArXiv arxiv.org/abs/2204.14217: Text-to-image system (CogView2) competitive with DALL-E-2. 6B-parameter hierarchical transformer for cross-modal language model. Fine-tuning by super-resolution modules. Tokens are re-masked and re-generated in a local parallel auto-regressive way.

Interesting research, though probably best way to show this is to let people compare prompts across both models, so should get the (mentioned in arxiv) demo website up for sure, this is a must read.
CogView2: Cross-Modality General Language Model

- A unification of “GPT(CogView) + MAE + Image captioning”
- Simplicity. No sentinel tokens, no order changing or token replacement.
- Only 20 * 20 tokens... 6 billion parameters!
- As a basic for future uses (super-resolution / video generation)!
CogVideo: Towards text-to-video generation

https://github.com/THUDM/CogVideo

Examples generated by CogVide

A woman is drinking waters.

A woman is running on beach at dusk.

A lion man is drinking a glass of water.

https://github.com/THUDM/CogVideo
A video in the training set

Split to fixed frame-rate(5) clips as real samples.

But... Each sample are companied with text “drinking”. Could the model learn the whole action?
Challenges: text-video alignment

Solution:

- Sample frames to cover the long video.
- Lead to different frame-rate?
- Treat the frame-rate as a condition!
- Recursive Interpolation
- Another model
Results

(a) Human preference. The percentage of being chosen as the best.
Controllable Generation from Pretrained Language Models via Inverse Prompting


Question: Which moment did you want to live in forever?

I have been in a relationship with my boyfriend for two years. I studied abroad and he studied domestically. Alone in a foreign country, I felt really lonely and sad. When I was in school, I felt that it was ok to be together, working hard together and quarrelling together. I kept thinking in this way. In the last few months of our time at the high school, I felt that he was my walking little brother. In fact, the happiest moment was when we were together in the “orphanage”.

The relativity is weak!

Question: Which moment did you want to live in forever?

Inverse Prompting:

In an instant I recall the second year of high school. That afternoon, I took a sunbath on the playground for a long time. I then played basketball downstairs with my classmates. As the sun set slowly, the three of us stood together on the corridor, leaning on the railing to face the sunset. I’ve always wanted to live in that moment forever, but unfortunately I never encountered a similar moment again. I just want to live in that moment.
Figure 1: The generation process of open-domain traditional Chinese poems under inverse prompting. Using title *New York* (《咏纽约》) as an example.
Inverse Prompting scoring method: Inversely prompt the title to improve the relativity.

Input:
Title: New York
Genre: Poem
Text: Clouds touch the rain in Manhattan.

Output:
曼哈顿中云拂雨

"曼哈顿中云拂雨"出自
诗歌《咏纽约》

目标输出：
诗歌《咏纽约》

Compute the log likelihood(-26.52), score=50-26.52=23.48
## Evaluation: QA

<table>
<thead>
<tr>
<th>Method</th>
<th>Fluency (1-5)</th>
<th>Inform.(^1) (1-5)</th>
<th>Relevance (1-5)</th>
<th>Overall (1-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM [27]</td>
<td>2.66</td>
<td>2.47</td>
<td>2.36</td>
<td>4.32</td>
</tr>
<tr>
<td>Prompting Baseline</td>
<td>3.44</td>
<td>3.25</td>
<td>3.21</td>
<td>5.97</td>
</tr>
<tr>
<td>Inverse Prompting</td>
<td><strong>3.61</strong></td>
<td><strong>3.43</strong></td>
<td><strong>3.59</strong></td>
<td><strong>6.51</strong></td>
</tr>
<tr>
<td>Human Answers</td>
<td>3.80</td>
<td>3.61</td>
<td>3.67</td>
<td>6.85</td>
</tr>
</tbody>
</table>

\(^1\) Informativeness
## Evaluation: Poem

<table>
<thead>
<tr>
<th>Method</th>
<th>Format (1-5)</th>
<th>Innov.(^1) (1-5)</th>
<th>Relevance (1-5)</th>
<th>Aes.(^2) (1-5)</th>
<th>Overall (1-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiuge [28] Search Baseline</td>
<td>3.60</td>
<td>2.47</td>
<td>1.99</td>
<td>3.12</td>
<td>3.57</td>
</tr>
<tr>
<td>Inverse Prompting</td>
<td>2.56</td>
<td>2.71</td>
<td>2.92</td>
<td>2.33</td>
<td>4.00</td>
</tr>
<tr>
<td>Inverse Prompting +ST</td>
<td>2.42</td>
<td>2.92</td>
<td>3.65</td>
<td>2.18</td>
<td>4.40</td>
</tr>
</tbody>
</table>

\(^1\) Innovation  
\(^2\) Aesthetics
## Turing Test

<table>
<thead>
<tr>
<th>Method</th>
<th>Total</th>
<th>Selected</th>
<th>Selection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse Prompting +ST</td>
<td>1,656</td>
<td>748</td>
<td>45.2%</td>
</tr>
<tr>
<td>Ancient Human Poems</td>
<td>1,656</td>
<td>908</td>
<td>54.8%</td>
</tr>
</tbody>
</table>

- 45.2%
Online Generation & Offline Knowledge Curation

**XDAI:** eXploiting large-scale pre-trained language models in building grounded Dialogue AI systems

Online Generation & Offline Knowledge Curation

Dialogue History Selection
- Most relevant historical utterances
  \[ \alpha^{-1} \cdot \text{cosine}(\langle U_i; S_i \rangle, U_t) \]

Dialogue Knowledge Injection
- Most similar QA-formatted knowledge to \([U_t; S_t]\)
- Injected before the corresponding dialogue utterance.

Background Knowledge Addition
- Description-formatted as the background
- Injected as the prefix of the entire prompt text.

---

**Background Knowledge**
- Description-formatted as the background
- Injected as the prefix of the entire prompt text.

**Knowledge QA-pairs**
- Injected before the corresponding dialogue utterance.

**Chat History**
- Generated Text
  - Me too. Let's play badminton together!
  - I am good. How about you?
  - Recently I found badminton very interesting.
  - Yes, badminton is a sport that can improve the flexibility of your arms and shoulders. At the same time, it can train the shoulder, elbow, wrist, knee and other parts. It is an activity that can strengthen your body and cultivate your mind. I would play badminton when I'm free as well.

**Prompt (w/o Knowledge injected)**
- Truncated Response
  - 我也是诶，一起打羽毛球哦！
  - Me too. Let's play badminton together!

**Prompt (w/ Knowledge injected)**
- Truncated Response
  - 我也是诶，一起打羽毛球哦！
Grounded on personalized condition (emotion)

Prompt: <dialogue context> [MASK] <grounded post-response>
Mechanism: Auto-regressive Blank Infilling in GLM
Prompt: <dialogue context> [MASK] <grounded post-response>
Mechanism: Auto-regressive Blank Infilling in GLM
Open!

WuDaoCorpora: the world's largest publicly available dataset!

We released almost all codes in WuDao!

You can download >20 well-trained models!

Directly call an API to enjoy the power of WuDao!

Click here to find more: https://wudaoai.cn/
WuDao is very general and can be applied to different tasks.
WuDao can do QA, write poem, knowledge extraction, draw pictures, write articles, recognize pictures, etc.
WuDao is open to everyone.
WuDao is very expensive... and needs to reduce cost...
WuDao’s Tomorrow

Teach Machine to Think Like Humans: Beyond the Turing Test!

1. Adaptation and Learning
2. Definition and Contextualization
3. Character Setting
4. Priority and Access Control
5. Call Together and Control
6. Decision Making and Execution
7. Probing and Editing
8. Reflection and Self-Monitoring
9. Logic and Flexibility

The slides will be available soon at http://keg.cs.tsinghua.edu.cn/jietang (or Google “Jie Tang”)
Thanks to everyone!

**WUDAO ·Wenyuan**  
A large scale pre-training language model with Chinese as its core

Liu Zhiyuan, Huang Minlie, Han Wentao, Liu Yang, Zhu Xiaoyan, Sun Maosong  
Zhang Zhengyari, Gu Yuxian, Han Xu, Chen Shengqi, Xiao Chaojun, Yao Yuan, Qi Fanchao, Guan Jian, Ke Pei, Zhou Hao, Sun Zhenbo, Cai Yanzheng, Zeng Guoyang, Tan Zhixing, Qin Yujia, Su Yusheng Si Chenglei,  
Hu Xueyu, Li Wenhao, Wang Fengyu, Yi Jing, Wang Xiaozhi, Chen Weize, Ding Ning, Zhang Jiajie

**WUDAO ·Wenlan**  
Super large multi-modal pre-training model

Wen Jirong, Song Ruihua, Lu Zhiwu, Jin Qin, Zhao Xin, Pang Liang, Lan Yanyan, Dou Zhicheng  
Gao Yizhao, Huo Yuqi, Lu Haoyu, Wen Jingyuan, Yang Guoxing, Song Haoyang, Zhang Manli, Zhang Liang, Hu Anwen, Li Ruichen, Song Yuqing, Zhao Jinning, Zhao Yida, Fei Nanyi  
Sun Yuchong, Jin Chuhao, Hong Xin, Cui Wanqing, Hou Danyang, Li Yingyan, Xi Zongzheng, Liu Guangzheng, Liu Peiyu,  
Gong Zheng, Li Junyi

**WUDAO ·Wenhui**  
A new super-large cognitive-oriented pre-training model

Tang Jie, Yang Zhilin, Yang Hongxia  
Du Zhengxiao, Ding Ming, Zou Xu, Qiu Jiezhang, Qian Yujie, Yinda, Zhong Qingyang, Yu Jifan, Liu Xiao, Zheng Yanan,  
He Jiaao, Zeng Aohan, Hong Wenyi, Yang Zhuoyi, Zheng Wendi, Zhou Jing, Du Jizhong  
Guo Zitong, Liu Jing, Zhou Chang, Lin Junyang

**WUDAO ·Wensu**  
Super large protein sequence prediction pre-training model

FastMoE and trillion large model

Tang Jie, Lu Bai  
Qiu Jiezhang, Xie Changyu, Xiao Yijia, Zeng Aohan, Li Zhaoh

Tang Jie, Zhai Jidong, Yang Hongxia, Chen Wenguang, Zheng Weimin  
Ma Zixuan, He Jiaao, Qiu Jiezhang, Cao Huanqi, Wang Yuanwei, Sun Zhenbo, Zheng Liyan, Wang Haojie, Tang Shizhi,  
Feng Guanyu, Zeng Aohan, Zhong Runxin, Shi Tianhui, Du Zhengxiao  
Ding Ming, Tiago Antunes, Peng Jinjun, Lin Junyang Zhang Jianwei

**WUDAO ·Wenyuan**  
A large scale pre-training language model with Chinese as its core

Liu Zhiyuan, Huang Minlie, Han Wentao, Liu Yang, Zhu Xiaoyan, Sun Maosong  
Zhang Zhengyari, Gu Yuxian, Han Xu, Chen Shengqi, Xiao Chaojun, Yao Yuan, Qi Fanchao, Guan Jian, Ke Pei, Zhou Hao, Sun Zhenbo, Cai Yanzheng, Zeng Guoyang, Tan Zhixing, Qin Yujia, Su Yusheng Si Chenglei,  
Hu Xueyu, Li Wenhao, Wang Fengyu, Yi Jing, Wang Xiaozhi, Chen Weize, Ding Ning, Zhang Jiajie

**WUDAO ·Wenlan**  
Super large multi-modal pre-training model

Wen Jirong, Song Ruihua, Lu Zhiwu, Jin Qin, Zhao Xin, Pang Liang, Lan Yanyan, Dou Zhicheng  
Gao Yizhao, Huo Yuqi, Lu Haoyu, Wen Jingyuan, Yang Guoxing, Song Haoyang, Zhang Manli, Zhang Liang, Hu Anwen, Li Ruichen, Song Yuqing, Zhao Jinning, Zhao Yida, Fei Nanyi  
Sun Yuchong, Jin Chuhao, Hong Xin, Cui Wanqing, Hou Danyang, Li Yingyan, Xi Zongzheng, Liu Guangzheng, Liu Peiyu,  
Gong Zheng, Li Junyi

**WUDAO ·Wenhui**  
A new super-large cognitive-oriented pre-training model

Tang Jie, Yang Zhilin, Yang Hongxia  
Du Zhengxiao, Ding Ming, Zou Xu, Qiu Jiezhang, Qian Yujie, Yinda, Zhong Qingyang, Yu Jifan, Liu Xiao, Zheng Yanan,  
He Jiaao, Zeng Aohan, Hong Wenyi, Yang Zhuoyi, Zheng Wendi, Zhou Jing, Du Jizhong  
Guo Zitong, Liu Jing, Zhou Chang, Lin Junyang

**WUDAO ·Wensu**  
Super large protein sequence prediction pre-training model

FastMoE and trillion large model

Tang Jie, Lu Bai  
Qiu Jiezhang, Xie Changyu, Xiao Yijia, Zeng Aohan, Li Zhaoh

Tang Jie, Zhai Jidong, Yang Hongxia, Chen Wenguang, Zheng Weimin  
Ma Zixuan, He Jiaao, Qiu Jiezhang, Cao Huanqi, Wang Yuanwei, Sun Zhenbo, Zheng Liyan, Wang Haojie, Tang Shizhi,  
Feng Guanyu, Zeng Aohan, Zhong Runxin, Shi Tianhui, Du Zhengxiao  
Ding Ming, Tiago Antunes, Peng Jinjun, Lin Junyang Zhang Jianwei
Wudao—Pretrain the world

Thanks!