Thibaud Southiratn

2024, June 13th

Efficient Molecule Captioning

Machine Learning in Bioinformatics
Term Project

Background – Baselines

(Model – **TextChemT5**)
Multi-task & Multi-domain language model (LM)

text2text
text2mol
mol2mol
mol2text = Molecule Captioning

(Dataset)
160,492 'Molecule-Description' pairs

Unifying Molecular and Textual Representations via Multi-task Language Modelling

Dimitrios Christofidellis [*1] **Giorgio Giannone** [*1 2 3]
Jannis Born [1 4] **Ole Winther** [2 5] **Teodoro Laino** [1] **Matteo Manica** [1]

"TextChemT5"

+

***L+M-24*: Building a Dataset for Language+Molecules @ ACL 2024**

Carl Edwards[1], **Qingyun Wang**[1], **Lawrence Zhao**[2] and **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign [2]Yale University
{cne2, qingyun4, hengji}@illinois.edu, larry.zhao@yale.edu

Motivation – TextChemT5

(Christofidellis et al, 2023)

<u>Good performance</u>

Coherent Word Matching (BLEU & ROUGE)

&

Semantic Similarity (METEOR)

→ What about the **meaning** ?

Table 3: Results of the SMILES to Caption (mol2text) task. The baselines include Transformer (Edwards et al., 2022), T5 (fine-tuned), and MolT5 (Edwards et al., 2022). The metrics used in the table include BLEU-2, BLEU-4, Rouge-1, Rouge-2, Rouge-L, and Meteor, all of which are common metrics used to evaluate text generation models. The table shows that our proposed model, Text+Chem T5, outperforms the other baselines in all the metrics. Overall, *Text+Chem T5* is able to generate more accurate and informative captions for SMILES.

	Size	BLEU-2 ↑	BLEU-4 ↑	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑	Meteor ↑
Transformer (Edwards et al., 2022)	-	0.061	0.027	0.188	0.0597	0.165	0.126
T5 (fine-tuned) (Raffel et al., 2020)	small	0.501	0.415	0.602	0.446	0.545	0.532
MolT5 (Edwards et al., 2022)	small	0.519	0.436	0.620	0.469	0.563	0.551
***Text+Chem T5* (ours)**	small	0.553	0.462	0.633	0.481	0.574	0.583
***Text+Chem T5-augm* (ours)**	small	**0.560**	**0.470**	**0.638**	**0.488**	**0.580**	**0.588**
T5(fine-tuned) (Raffel et al., 2020)	base	0.511	0.424	0.607	0.451	0.550	0.539
MolT5 (Edwards et al., 2022)	base	0.540	0.457	0.634	0.485	0.578	0.569
***Text+Chem T5* (ours)**	base	0.580	0.490	0.647	0.498	0.586	0.604
***Text+Chem T5-augm* (ours)**	base	**0.625**	**0.542**	**0.682**	**0.543**	**0.622**	**0.648**

Input: Caption the following smile: COC1=C(C=C2C3CC4=CC(=C(C=C4C(N3)CC2=C1)OC)OC)OC

Expected output: The molecule is a racemate comprising equimolar amounts of <u>(R,R)- and (S,S)-pavine</u>. It has <u>a role as a plant metabolite</u>. It contains a <u>(R,R)-pavine and a (S,S)-pavine</u>. It is a conjugate base of a pavine(1+).

From descriptive → more **comprehensive** captions.

Motivation – L+M24

3 Data Sources
- PubChem
- Chemical Function (CheF)
- ChemFOnt (Chemical Functional Ontology resource)

Chemical Properties
- Molecule-specific descriptions
- Functional properties

GPT4 Templates → Captions
- **Compositional**
- **Functional**
- **Abstract**

Light/Electricity, Biomedical, Agriculture/Industry, Human Interactions & Organoleptics

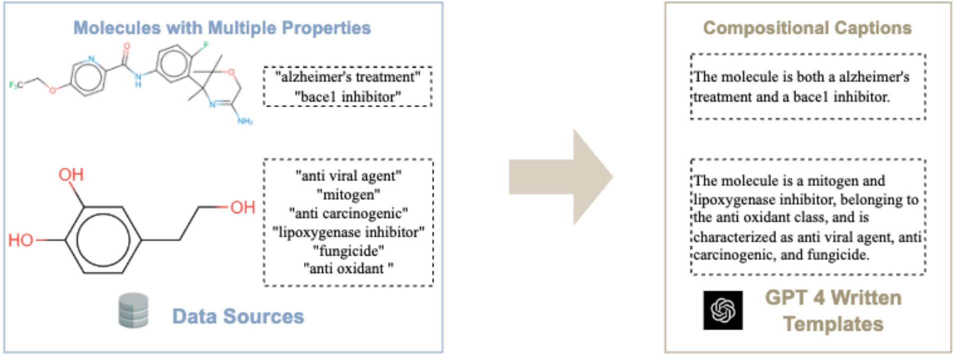
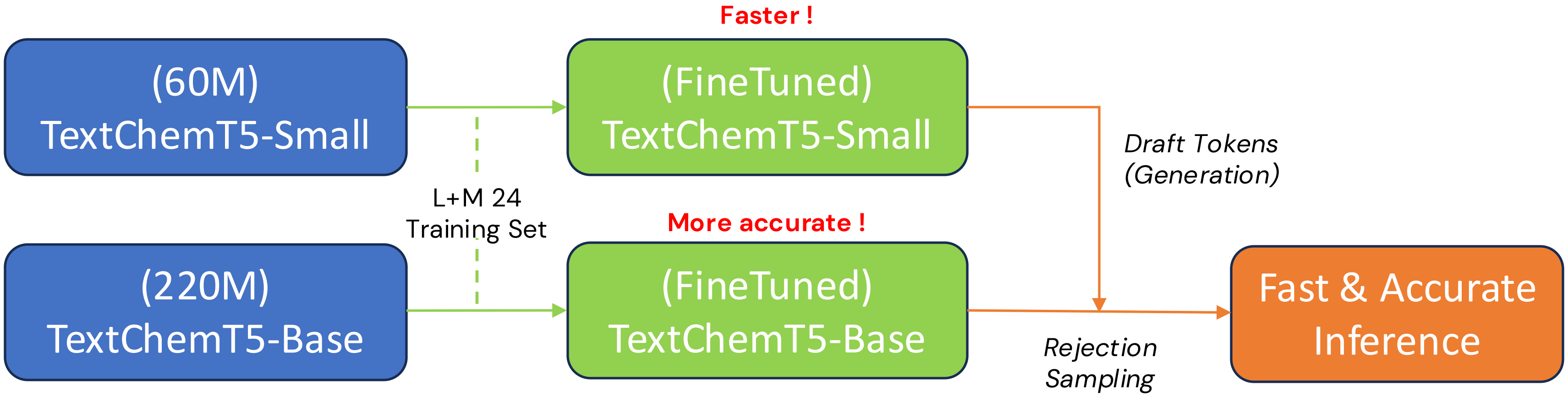


Figure 1: Example descriptions created for molecules from the training set.

Goal & Approach

Two Steps:

1. Improve performance → **Finetune** TextChemT5 on L+M 24 Dataset
2. Accelerate inference speed → **Speculative Decoding***



** LLM Technique, initially introduced by Google (Leviathan et al, 2023) & DeepMind (Chen et al, 2023)*

Improving Performance

Training Settings:

20 epochs
Optimizer: Adafactor*
Learning Rate: 2e-5
Batch Size: 128 (Small) & 96 (Base)

Input Max Length: 128
Output Max Length: 128

Prompt: "Caption the following SMILES: <input>"

Multi-GPU: From ~16 to <1 GPU hour per epoch

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small †	70.9	51.2	74.5	55.8	54.4	70.1
MolT5-Base †	73.8	53.5	75.0	55.9	53.9	71.8
MolT5-Large †	76.9	55.6	77.7	58.0	55.7	74.3
Meditron-7B †	79.2	57.6	79.7	60.2	57.5	75.7
(Baseline - Not FineTuned)						
TextChemT5-Small	7.1	3.1	17.1	8.2	15.4	14.9
TextChemT5-Base	7.0	3.2	13.7	6.3	12.3	10.9
(Fine-Tuned)						
TextChemT5-Small	**76.3**	**55.4**	**76.2**	**57.1**	**54.9**	**72.9**
TextChemT5-Base	**77.1**	**55.5**	**77.1**	**57.6**	**55.2**	**73.8**

Molecule captioning results on the validation split of L+M-24.
(Marked models' results are reported from the original L+M-24 Manuscript)

(Edwards et al, 2023)

Results:

- Small (60M) > Base (220M)
- Base (220M) performs similarly to Large (738M)

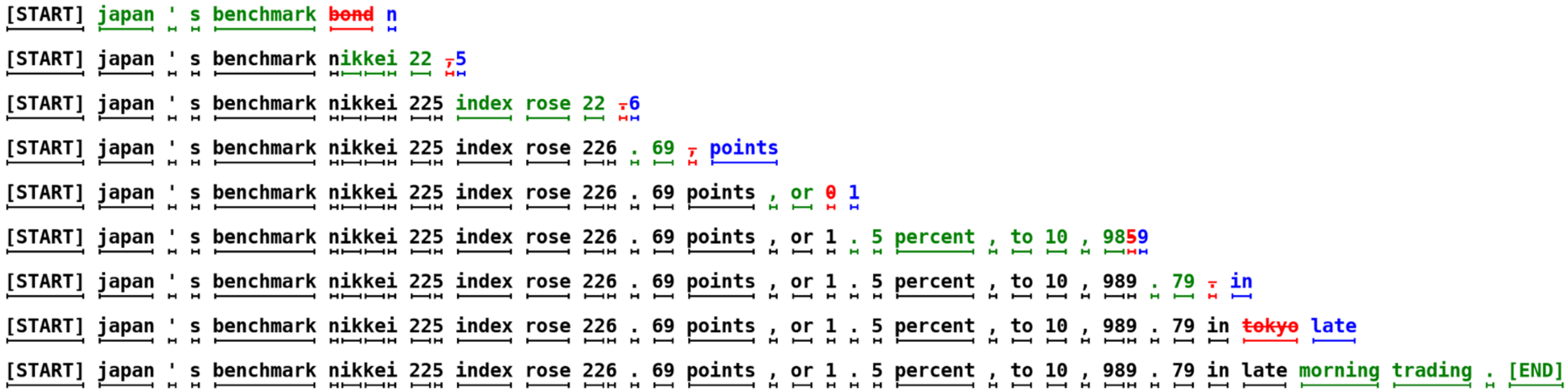
* *Adafactor from Shazeer et al, 2018*

Accelerating Inference (1)

<u>Speculative Decoding – Keypoints :</u>

- Idea: In a **single forward pass,**
    - Use a <u>smaller</u> model to decode (generate) MULTIPLE tokens faster
    - Iteratively Accept/Reject the tokens through the <u>larger</u> model.

- <u>Best case</u>: all "draft tokens" are accepted → **We saved time !**
- <u>Average case</u>: first "draft tokens" are accepted, then the next ones are rejected → **We saved time !**
- <u>Worst case</u>: the very first "draft token" is rejected → We don't save time… but **we don't lose time either !**



(Leviathan et al, 2023)

Accelerating Inference (2)

```
> /gamma 4
Gamma: 4
> Caption the following SMILES: CC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CCCCCC(=O)OC[C@H](COP(=O)(O)OC[C@H](O)COP(=O)(O)OC[C@@H](COC(=O)CCC
CCCC/C=C\CCCCCC)OC(=O)CCCCCCCCCCCCC)OC(=O)CCCCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CC
========== Speculative ==========
Out: The molecule is a stabilizing mitochondrial structure, proton trap for oxidative phosphorylation, stabilizing cytochrome oxid
ase that impacts aging and tangier disease. The molecule is a cholesterol translocation and a apoptosis that impacts non-alcoholic
 fatty liver disease, barth syndrome, and diabetic heart disease.
Acceptance rate: 1.000
Throughput: 211.7 tokens/s
========== Speculative ==========
=========== Target AR ===========
Out: The molecule is a stabilizing mitochondrial structure, cholesterol translocation, stabilizing cytochrome oxidase that impacts
 tangier disease and barth syndrome. The molecule is a proton trap for oxidative phosphorylation and a apoptosis that impacts agin
g, non-alcoholic fatty liver disease, and diabetic heart disease.
Throughput: 149.0 tokens/s
=========== Target AR ===========
Throughput increase: 142.1%
```

```
> /gamma 25
Gamma: 25
> Caption the following SMILES: CC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CCCCCC(=O)OC[C@H](COP(=O)(O)OC[C@H](O)COP(=O)(O)OC[C@@H](COC(=O)CCC
CCCC/C=C\CCCCCC)OC(=O)CCCCCCCCCCCCC)OC(=O)CCCCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CC
========== Speculative ==========
Out: The molecule is a stabilizing mitochondrial structure, cholesterol translocation, proton trap for oxidative phosphorylation t
hat impacts aging and non-alcoholic fatty liver disease. The molecule is a stabilizing cytochrome oxidase and a apoptosis that imp
acts tangier disease, barth syndrome, and diabetic heart disease.
Acceptance rate: 0.860
Throughput: 193.4 tokens/s
========== Speculative ==========
=========== Target AR ===========
Out: The molecule is a stabilizing mitochondrial structure, cholesterol translocation, stabilizing cytochrome oxidase that impacts
 tangier disease and barth syndrome. The molecule is a proton trap for oxidative phosphorylation and a apoptosis that impacts agin
g, non-alcoholic fatty liver disease, and diabetic heart disease.
Throughput: 148.9 tokens/s
=========== Target AR ===========
Throughput increase: 129.9%
```

(Code is adapted from [R Storai, Speculative Decoding, Github Repository 2024, https://github.com/romsto/Speculative-Decoding](https://github.com/romsto/Speculative-Decoding))

Accelerating Inference (3)

Out of 20 random samples (gamma 25):

- Average acceptance rate: 89.3%
- Average throughput increase: 136.5%

With the same accuracy !

Recap

Two Steps:

1. Improve performance → **Finetune** TextChemT5 on L+M 24 Dataset
2. Accelerate inference speed → **Speculative Decoding***

What's Next ?

Performance & Speed:

- Train Larger Models (T5 Large/3B/11B)

Interpretability:

- Finetune on approved drugs dataset using more specific knowledge (pharmacogenomics, metabolic reactions…)

e.g. DrugBank

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small †	70.9	51.2	74.5	55.8	54.4	70.1
MolT5-Base †	73.8	53.5	75.0	55.9	53.9	71.8
MolT5-Large †	76.9	55.6	77.7	58.0	55.7	74.3
Meditron-7B †	79.2	57.6	79.7	60.2	57.5	75.7
(Baseline - Not FineTuned)						
TextChemT5-Small	7.1	3.1	17.1	8.2	15.4	14.9
TextChemT5-Base	7.0	3.2	13.7	6.3	12.3	10.9
(Fine-Tuned)						
TextChemT5-Small	**76.3**	**55.4**	**76.2**	**57.1**	**54.9**	**72.9**
TextChemT5-Base	**77.1**	**55.5**	**77.1**	**57.6**	**55.2**	**73.8**

Molecule captioning results on the validation split of L+M-24.

(Marked models' results are reported from the original L+M-24 Manuscript)

(Edwards et al, 2023)

Thank you