Sapienza NLP Reading Group

09/03/2022 - Editing Factual Knowledge in Language Models

Nicola De Cao, Wilker Aziz, Ivan Titov

SAPIENZA UNIVERSITÀ DI ROMA

Babelscape

SAPIENZA NLP

Summarizer

Summarizer (Bacciu)

- This paper presents an innovative technique for updating language model information.

- This update is achieved by avoiding fine tuning of the model or training from scratch.

- This allows old information such as "Donald Trump is the president of the United States" to be updated with "Biden is the president of the United States".

Summarizer (Bacciu)

- They propose KNOWLEDGE EDITOR which is a hyper-network (Ha et al., 2017) a neural network that predicts the parameters of another network.
    - That hyper-network learns to modify implicit knowledge stored within LM parameters efficiently and reliably.

- In deep learning area this technique is called *Learning to Update*.

Summarizer (Bacciu)

- They use closed-book for testing Fact-Checking and Question Answering

- To evaluate the performance of this innovative method the authors also propose a set of new metrics: success rate, *retain accuracy*, *equivalence accuracy*, *performance deterioration*.

- This approach obtains successful results, exceeding the fine-tuning baseline.

Reviewer 1

Reviewer 1 (Carlos)

Main advantages of the paper:

1. Present a new task of **Knowledge Editing**:
   a. A method to **modify** the implicit knowledge of the **LM parameters**.
   b. Defining a set of **metrics** to measure the efficacy of the task.
2. The method can easily and **efficiently** modify the knowledge acquire by the LM
3. Show **first insights** that prove the **effectiveness** of the method comparing with other baselines

Reviewer 1 (Bejgu)

Propose a novel universal approach to correct factual knowledge in any pre-trained LM:

1. Strong performances on **success rate**, **reliability**, and **consistency** without significant **performance deterioration**
2. Strong **consistency** performances using paraphrases but were automatically generated
3. Minor performance improvement compared to **simple** fine-tuning, but authors claim their approach is **more efficient** (not demonstrated in the paper).

SAPIENZA NLP

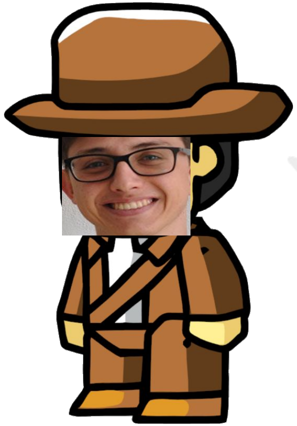
Reviewer 2

Reviewer 2 (Pere-Lluis)

- Handcrafted metrics.
- Overselling what is in fact, slightly modified backpropagation.
- Improvement comes from using extra data (paraphrases).
- Efficiency of the method not covered.
- No significance testing.

Archeologist

Niccolò Campolungo



Refresher

How to constrain?

$$\min_{\phi} \sum_{\hat{x} \in \mathcal{P}^x} \mathcal{L}(\theta'; \hat{x}, a)$$

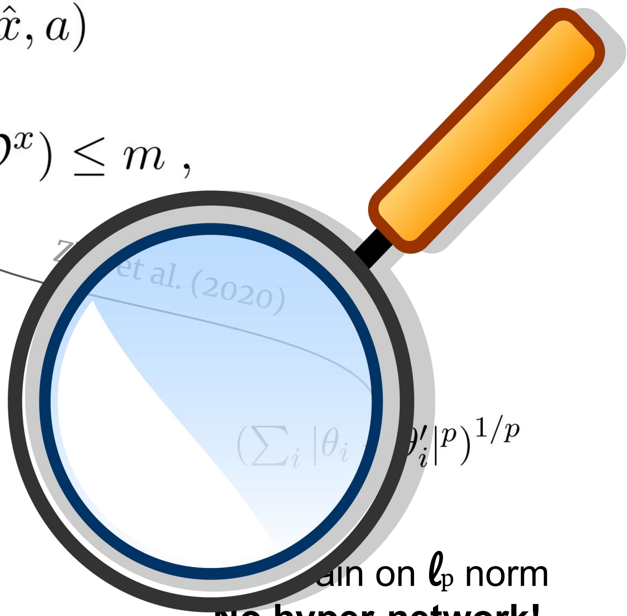
$$\text{s.t.} \quad \mathcal{C}(\theta, \theta', f; \mathcal{O}^x) \leq m \ ,$$

De Cao et al. (2021)

$$\sum_{x' \in \mathcal{O}^x} \sum_{c \in \mathcal{Y}} p_{Y|X}(c|x', \theta) \log \frac{p_{Y|X}(c|x', \theta)}{p_{Y|X}(c|x', \theta')}$$

Constrain on KL Divergence

Z... et al. (2020)



$(\sum_i |\theta_i - \theta'_i|^p)^{1/p}$

...ain on ℓ_p norm

No hyper-network!

Modifying Memories in Transformer Models

Zhu et al. (2020)

- **Core idea**: constrain on ℓ_∞ norm

- Experimental settings
  - Fine-tune on:
    - Full **unmodified** facts dataset (FT)
    - **Modified** facts only (FTM)
    - **Mixture** of modified and unmodified facts (FTA)
  - Layers:
    - Specific layer
    - All together

SAPIENZA NLP

Main findings

Zhu et al. (2020)

- **Tuning specific layers** is (generally) better
- FT before FTM/FTA helps
  - Surprisingly, FTM > FTA
- Updating **too many facts** leads to performance degradation
- You **forget** *some* **previously-known facts** when updating the model's knowledge...

SAPIENZA NLP

Futurist

Futurist (Riccardo)

Futurist (Riccardo)

- Neural models are (usually) black boxes
  - No control over their (*implicit*) knowledge
- Successful knowledge editing can **facilitate maintaining models** in production environments
  - Correct biases/outdated information induced by training corpora
  - No need to re-train to add/correct facts over time
  - More robust **continual learning** setup
- Can these techniques be **applied in MTL** to cope with catastrophic forgetting?
- Or to **adapt** models to **new tasks/domains** on demand?

SOTA

Related Work

Sinitsin et al. (2020) propose a **meta-learning approach** for model modification that learns parameters that are easily editable at test time. To have a reliable method, they employ a regularized objective **forcing the updated model not to deviate from the original one**.

Zhu et al. (2020) use **constrained optimization.** They re-finetune on a specific downstream task (with altered data). Their method employs either an L2 or L∞ constraint between the original model's parameters and the edited ones.

$$\mathcal{C}_{L_p}(\theta, \theta', f; \mathcal{O}^x) = (\sum_i |\theta_i - \theta'_i|^p)^{1/p}$$

SAPIENZA NLP

SoTA

Method	Fact-Checking Success rate ↑	Fact-Checking Retain acc ↑	Fact-Checking Equiv. acc ↑	Fact-Checking Perform. det ↓	Question Answering Success rate ↑	Question Answering Retain acc ↑	Question Answering Equiv. acc ↑*	Question Answering Perform. det ↓
Fine-tune (1st layer)	100.0	99.44	42.24	0.00	98.68	91.43	89.86 / 93.59	0.41
Fine-tune (all layers)	100.0	86.95	95.58	2.25	100.0	67.55	97.77 / 98.84	4.50
Zhu et al. (1st layer)	100.0	99.44	40.30	0.00	81.44	92.86	72.63 / 78.21	0.32
Zhu et al. (all layers)	100.0	94.07	83.30	0.10	80.65	95.56	76.41 / 79.38	0.35
KnowledgeEditor	98.80	98.14	82.69	0.10	94.65	98.73	86.50 / 92.06	0.11
+ loop†	100.0	97.78	81.57	0.59	99.23	97.79	89.51 / 96.81	0.50
+ $\mathcal{P}^x$ ‡	98.50	98.55	95.25	0.24	94.12	98.56	91.20 / 94.53	0.17
+ $\mathcal{P}^x$ + loop‡	100.0	98.46	94.65	0.47	99.55	97.68	93.46 / 97.10	0.95

Table 1: Accuracy scores on fact-checking and question answering for the metrics presented in Section 2.2. *We report both the accuracy on the set of generated paraphrases (left) and human-annotated (right). †Apply updates in a loop, stopping when the update is a success or when reaching a maximum number of iterations (only at test time). ‡Using paraphrases (semantically equivalent inputs) as additional supervision (only at training time).

Fine-tune =using standard gradient descent, minimizing the loss for the fact/prediction that needs revision.

For this, authors followed **Sinitsin et al. (2020)** and employed **RMSProp (Tieleman and Hinton, 2012).**

PI: Nicola de Cao (Martelli)

- PhD candidate at the Institute for Logic, Language and Computation (University of Amsterdam) and permanent visiting of the School of Informatics (University of Edinburgh)
- Personal website: https://nicola-decao.github.io/
- His research interests are:
  - Machine reading comprehension and question answering
  - Supervised and unsupervised deep neural network applications
  - Reasoning and reinforcement methods
- Recommended readings:
  - GenIE: Generative Information Extraction (https://arxiv.org/pdf/2112.08340.pdf)

SAPIENZA NLP

PI: Wilker Aziz (Martelli)

- Assistant professor at the Institute for Logic, Language and Computation (University of Amsterdam)
- Personal website: https://wilkeraziz.github.io/
- His research interests are:
  - Natural Language Understanding tasks like information extraction, machine translation, language modeling
  - Interpretability of deep learning models
- Recommended readings:
  - How do decisions emerge across layers in neural models? Interpretation with differentiable masking (https://arxiv.org/pdf/2004.14992.pdf)

SAPIENZA NLP

PI (Tedeschi): Ivan Titov

- Associate Professor at University of Amsterdam and University of Edinburgh
- His research interests are:
  - Natural Language Understanding tasks like information extraction, machine translation, question answering and semantic parsing
  - Meta-learning
  - Interpretability and controllability of deep learning models
- His research have been supported by several grants (e.g. ERC)
- Related research:
  - Sparse Interventions in Language Models with Differentiable Masking (https://arxiv.org/abs/2112.06837)

SAPIENZA NLP

Social Impact

Social Impact

Cesare Campagnano and **Luigi Procopio**

SAPIENZA NLP

What is Social Impact?

- Quite the odd role

  Identify how this paper self-assesses its impact on the world. Have any positive social impacts left out? What are possible negative ones that were overlooked?

- Aka, **Ethicist from the Future**
- *It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process*

SAPIENZA NLP

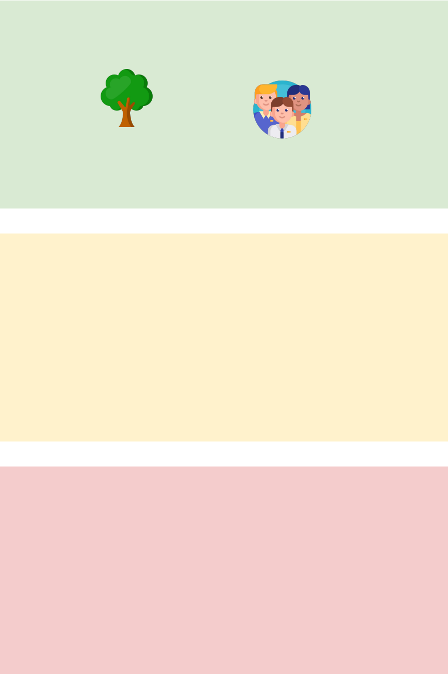
Avoid...

- **Avoid** re-training large LMs...
- **Avoid** fine-tuning large LMs...
- To fix *a few problematic* predictions

Green AI

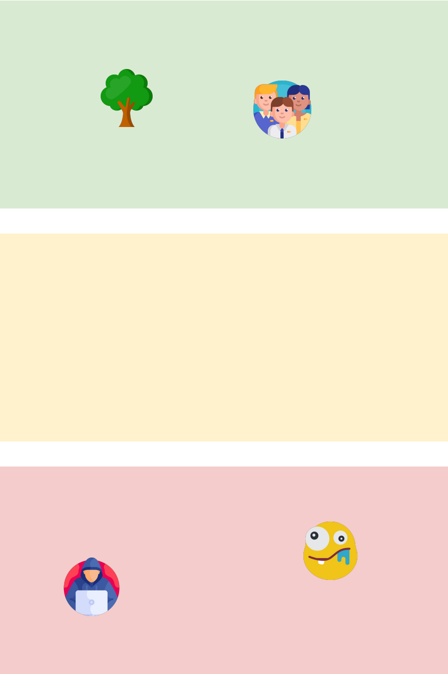
AI Democratization

So, we use a set D to fix the model...



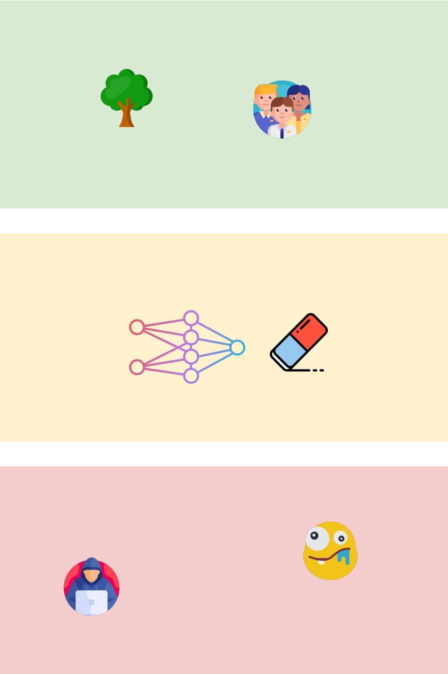
- What if (*Trudy*, aka the **malicious data scientist**) or (*Bob*, aka the **superficial data scientist**) write $\mathcal{D}$?
- might edit a CV screening network so that all women are auto-discarded (<u>sexist</u>)
- might accidentally produce an unbalanced D which still auto-discards all women
- This also applies for $\mathcal{O}^x$

So, we use a set D to fix the model...



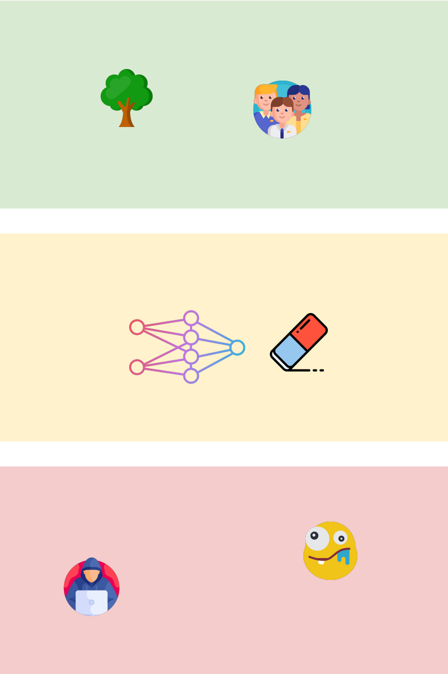
- *KnowledgeEditor* gives us a way to change (both positively and negatively) a network behavior
- Adversarial attacks need not craft special datasets whose long training/fine-tuning **might eventually** result in their goal
- They can **explicitly inject it**

Mind the Data



- So, as usual, **mind the data**
- We must be careful when crafting $\mathcal{D}$ and $\mathcal{O}^x$
- They must be:
    - Properly **balanced**
    - Unbiased (and maybe even **de-biasing**)
    - **Thoroughly examined**

What if...



Open Question

What if Trudy is the owner of the edited model? That is, what if the <u>biasing is intended</u>?

SAPIENZA NLP

Thank you!

Eventuale link al progetto o sottotitolo

Visit our website http://nlp.uniroma1.it and follow us on:

@SapienzaNLP

SAPIENZA NLP