A Case of Plagarism in Machine Learning Research

by **Nicholas Carlini** 2022-04-08

I recently came to be aware of a case of plagiarism in the machine learning research space. The paper A Roadmap for Big Model plagiarized several paragraphs from one of my recent papers Deduplicating Training Data Makes Language Models Better. (There is some irony in the fact that the Big Models paper copies from a paper about data copying. This irony was not lost on us.) This is unfortunate, but to my dismay, our paper was not the only paper copied from: the Big Models paper copied from at least a dozen other papers.

In the grand scheme of things, this particular form of copying isn't the worst thing ever. It's not like a paper has directly copied the method of a prior result and claimed it as its own. But even putting aside the fact that claiming someone else's writing as one's own is wrong, the value in survey papers is in how they re-frame the field. A survey paper that just copies directly from the prior paper hasn't contributed anything new to the field that couldn't be obtained from a list of references.

*[**Update 4/12**: This article has received a lot more attention than I expected. (Context: every hour more people visit this page than viewed my entire website last week.) So a plea: let's not turn this into a witch hunt. I've seen some people say things like this should result in immediate dismissal of all those involved / people should be banned from arXiv / etc. I don't pretend to know the situation that resulted in this paper having copied from so many sources. Without knowing what happened behind the scenes, I'd like to refrain passing judgement. Maybe some junior authors meant well and thought that a citation was enough to then copy text. Maybe there was pressure from above that made some students feel like their only choice to deliver on time was to cut corners. For the part of the senior authors, they may have read over the text and thought that it looked perfectly reasonable and only made a few tweaks to the*

text here and there without being aware of where it came from. The point is we don't know. With 100 authors on this paper anything could have happened.

My hope with this post was just to draw some attention to something that I've seen happen not infrequently. For example, roughly 1% of published-and-accepted papers have a higher data-copying-fraction than this paper. I should have given this context when I wrote this post initially. So, again, please let's not come down to harshly on this paper in particular. This is a problem I've noticed with the field in general, this case was just the tipping point for me because it was a paper of mine where this happened. Hopefully we can treat this as a learning experience to improve the field as a whole. With that out of the way, back to your regularly scheduled programming...]

See below for a few of the more egregious examples of this, with text from the Big Models paper on the left and the corresponding text from the original paper on the right. Copied text is highlighted in green.

Text from the "Big Models" Paper	Text from the Original Paper
The risks of data memorization, for example, the ability to extract sensitive data such as valid phone numbers and IRC usernames, are highlighted by Carlini et al. [41]. While their paper identifies 604 samples that GPT-2 emitted from its training set, we show that over 1 of the data most models emit is memorized training data. In computer vision, memorization of training data has been studied from various angles for both discriminative and generative models Deduplicating training data does not hurt perplexity: models trained on deduplicated datasets	[Original]: The privacy risks of data memorization, for example the ability to extract sensitive data such as valid phone numbers and IRC usernames, are highlighted by Carlini et al. (2020). While their paper finds 604 samples that GPT-2 emitted from its training set, we show that over 1% of the data most models emit is memorized training data. In computer vision, memorization of training data has been studied from various angles for both discriminative and generative models Deduplicating training data does not hurt perplexity: models trained on

have no worse perplexity compared to baseline models trained on the original datasets. In some cases, deduplication reduces perplexity by up to 10%. Further, because recent LMs are typically limited to training for just a few epochs [47, 47], the models can reach higher accuracy faster by training on higher-quality data. The simplest technique to find duplicate examples would be to perform exact string matching between all example pairs, but as we will show, this is insufficient. We introduce two complementary methods for performing deduplication. First, using a suffix array [49], we remove duplicate substrings from the datasets if they occur verbatim in more than one example. Second, we use MinHash [48] , an efficient algorithm for estimating the n-gram similarity between all pairs of examples in a corpus, to remove entire examples from the dataset if they have high n-gram overlap with any other example. causes researchers to over-estimate model accuracy and biases model selection towards models and hyperparameters that intentionally overfit	deduplicated datasets have no worse perplexity compared to baseline models trained on the original datasets. In some cases deduplication reduces perplexity by up to 10%. Further, because recent LMs are typically limited to training for just a few epochs (Radford et al., 2019; Raffel et al., 2020), by training on higher quality data the models can reach higher accuracy faster. The simplest technique to find duplicate examples would be to perform exact string matching between all example pairs, but as we will show, this is insufficient. We introduce two complementary methods for performing deduplication. First, using a suffix array (Manber and Myers, 1993), we remove duplicate substrings from the dataset if they occur verbatim in more than one example. Second, we use MinHash (Broder, 1997), an efficient algorithm for estimating the n-gram similarity between all pairs of examples in a corpus, to remove entire examples from the dataset if they have high n-gram overlap with any other example. causes researchers to over-estimate model accuracy, but also biases model selection towards models and hyperparameters that intentionally overfit

propose the notion of a World Scope (WS) as a lens through which to audit progress in NLP. They define five WSs, and they note that the most popular pre-training in NLP operates in the WS2 (Internet)	[Original]: propose the notion of a World Scope (WS) as a lens through which to audit progress in NLP. We describe five WSs, and note that most trending work in NLP operates in the second (Internet-scale data).
In addition to BERT, where masked words are predicted from the non-masked words in the language modality, LXMERT proposes cross-modality model architecture that could predict masked words from the visual modality as well so as to resolve ambiguity. For example, it is hard to determine the masked word carrot from its language Who is eating the carrot?, but the word choice is clear if the visual information is available	[Original]: In addition to BERT where masked words are predicted from the non-masked words in the language modality, LXMERT, with its cross-modality model architecture, could predict masked words from the vision modality as well, so as to resolve ambiguity. For example, as shown in Fig. 2, it is hard to determine the masked word 'carrot' from its language context but the word choice is clear if the visual information is considered
a number of information-seeking questions such as what is the definition of ... as the prompts, [...] show that this self-talk method substantially improves the performance of zero-shot big model baselines on four out of six commonsense benchmarks, and competes with models that obtain knowledge from external knowledge bases.	[Original]: a number of information seeking questions such as "what is the definition of ..." to discover [...] that the self-talk procedure substantially improves the performance of zero-shot language model baselines on four out of six commonsense benchmarks, and competes with models that obtain knowledge from external KBs.
even if the social bias is eliminated at the word level, the sentence-level bias can still exist due to the	[Original]: even if the social bias is eliminated at the word level,the sentence-level bias can still be

imbalanced combination of words [..] replacing sensitive words in the original sentence with words in a similar semantic but different bias directions.	caused by the unbalanced combination of words [..] by replacing sensitive words in the original sentence with words in a similar semantic but different bias directions.
It proposes two methods to learn cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. [...] Both the CLM and MLM objectives are unsupervised and only require monolingual data. For improving cross-lingual pre-training, they introduce a new translation language modeling (TLM) objective. They consider cross-lingual language model pre-training with either CLM, MLM, or MLM is used in combination with TLM.	[Original]: We propose two methods to learn cross-lingual lan- guage models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. [...] Both the CLM and MLM objectives are unsupervised and only require monolingual data. We introduce a new translation language modeling (TLM) objective for improving cross-lingual pretraining. In this work, we consider cross-lingual language model pretraining with either CLM, MLM, or MLM used in combination with TLM.
to large-scale mono-lingual corpora across many languages. The input texts are noised by masking phrases and permuting sentences, and a single Transformer model is learned to recover the texts.	[Original]: to large-scale monolingual corpora across many languages. The input texts are noised by masking phrases and permuting sentences, and a single Transformer model is learned to recover the texts.
Such models not only have lower inference latency, but they also do	[Original]: Such models not only have lower inference latency, but

not suffer from the problem of errors that propagate from one component to the next	they also do not suffer from the problem of errors that propagate from one component to the next
has presented a study of adapters for multilingual ST and shown that language-specific adapters can enable a fully trained multilingual ST model to be further specialized in each language pair.	[Original]: have presented a study of adapters for multilingual ST and shown that language-specific adapters can enable a fully trained multilingual ST model to be further specialized in each language pair
147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017. DialoGPT [...] The GPT-2 transformer model adopts the generic transformer language model [25] and leverages a stack of masked multi-head self-attention layers to train on massive web-text data. [...] DialoGPT inhered a 12-to-48 layer transformer with layer normalization, a initialization scheme that accounts for model depth that we modified, and byte pair encodings [1236] for the tokenizer MMI employs a pre-trained backward model to predict source sentences from given responses [...] a strong preference can be observed for DialoGPT over PersonalityChat [1432]	[Original]: 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017, DialoGPT [...] The GPT-2 transformer model adopts the generic transformer language model (Vaswani et al., 2017) and leverages a stack of masked multi-head self-attention layers to train on massive web-text data. [...] Our model inherits from GPT-2 (Radford et al., 2018), a 12-to-48 layer transformer with layer normalization, a initialization scheme that accounts for model depth that we modified, and byte pair encodings (Sennrich et al., 2016) for the tokenizer. MMI employs a pre-trained backward model to pre- dict source sentences from given responses. [...] strong preference can be observed for DialoGPT over PersonalityChat.

How did we find these examples?

One of my coauthors was reading the Big Models paper and noticed that some of the text seemed oddly familiar, and after quickly looking things over we found that in fact a bunch of the text was directly copied from our paper.

Given that this happened to us, we then set out to see if there were other examples too. As part of a prior project, I had collected a dataset of PDFs for (almost) every accepted paper at top machine learning venues (ICML/ICLR/NeurIPS/AAAI/ACL/etc). So all I did to find the above copied text was to take these PDFs, extract out all of the text and dump it into a single .txt file, and then run our dataset deduplication tools (that we developed for the paper that was copied from!) to find all repeated sequences that were contained both in the Big Models paper along with some other prior publication. To rule out false positives, I only considered sequences of

1. at least 10 words (after whitespace normalization),
2. that are contained sequentially in the Big Models paper,
3. and also present in a prior paper,
4. but are not present in more than one prior paper.

This ensures that I won't flag any common phrases as copied (e.g., copyright blocks, citations to prior paper titles or author names, etc).

And then from there, it was just a matter of quickly manually reviewing a few of the most egregious cases (shown above). There were other examples of self-plagiarism where the paper that was copied from shared an author with the new paper that I have omitted–while this isn't an ideal practice, it's less concerning.

Because of this filtering process, and because my dataset of papers is not exhaustive over all prior publications (notably, it only contains accepted papers, not arXiv preprints), it is possible there is more copying going on here than I have identified. However even what we have found so far is already more than should happen, and I am saddened that this is happening at all.

If you want to be notified the next time I write something (maybe like this, maybe not, who knows) enter your email address here. you@example.com Submit

There's also an RSS Feed if that's your thing.