Commentary on the Abstract of Latent Backdoor Attacks on Deep Neural Networks *(Yao et al. CCS 2019).*

Functions are literally everywhere in language, and so too in this Abstract.

To begin with, consider the individual words used. The Abstract totals 215 word-tokens, but 22% of that total is comprised by just eight word-types: a, and, attack, backdoor, be, latent, model, the. *Here, have a look for yourself — the eight word-types have been boldfaced.*

> Recent work proposed **the** concept **of backdoor attacks** on deep neural networks (DNNs), where misclassification rules **are** hidden inside normal **models,** only to **be** triggered by very specific inputs. However, these "traditional" **backdoors** assume **a** context where users train their own **models** from scratch, which rarely occurs in practice. Instead, users typically customize "Teacher" **models** already pretrained by providers like Google, through **a** process called transfer learning. This customization process introduces significant changes to **models** **and** disrupts hidden **backdoors,** greatly reducing **the** actual impact **of backdoors** in practice.
>
> In this paper, we describe **latent backdoors, a** more powerful **and** stealthy variant **of backdoor attacks** that functions under transfer learning. **Latent backdoors** are incomplete **backdoors** embedded into **a** "Teacher" **model, and** automatically inherited by multiple "Student" **models** through transfer learning. If any Student **models** include **the** label targeted by **the backdoor,** then its customization process completes **the backdoor and** makes it active. We show that **latent backdoors** can **be** quite effective in **a** variety **of** application contexts, **and** validate its practicality through real-world **attacks** against traffic sign recognition, iris identification **of** volunteers, **and** facial recognition **of** public figures (politicians). Finally, we evaluate 4 potential defenses, **and** find that only one **is** effective in disrupting **latent backdoors,** but might incur **a** cost in classification accuracy as tradeoff.

That's a lot of repetition. However, the repetition is of two different sorts.

The first sort is high-frequency words in the language. Literally the four most common words in English (i.e., a, and, be, the) *make for 10% of the tokens total. But that means that the other four words outmatch these highest-frequency words for frequency. Those four words represent the second sort of repetition: the topic words of the research. Basically, the* *topic words are those which repeat and repeat, throughout the paper, because the reader just keeps learning more and more about the things the words refer to.*

So, what's the takeaway?

The takeaway is this: Learn how to repeat well. In fact, you might profitably view the writing of scientific prose as an exercise in knowing when to say once more again yet another piece of information about the same basic topic. This means overcoming your inbred fear of USING THE SAME WORD TWICE ! To write well, you'll very often use the same word twice. Consider again the second paragraph in the Abstract, reproduced here without repetition.

> In this paper, we describe **latent backdoors,** a more powerful and stealthy variant of **similar such attacks** that functions under transfer learning. **They** are incomplete **ones** embedded into a "Teacher" **model,** and automatically inherited by multiple "Student" **ones** through transfer learning. If any Student **ones** include the label targeted by **one,** then its customization process completes **it** and makes it active. We show that **they** can be quite effective in a variety of application contexts, and validate its practicality through real-world **ones** against traffic sign recognition, iris identification of volunteers, and facial recognition of public figures (politicians). Finally, we evaluate 4 potential defenses, and find that only one is effective in disrupting **them,** but might incur a cost in classification accuracy as tradeoff.

Admittedly, this is an ultra-extreme reduction of the repetition, but it's meant to demonstrate, by way of removal, the functions which repetition brings to a text. Repetition, as you have seen, is topic creation. The topic of this Abstract, for instance, is latency *and* backdooring. *The transmission of this topic is right there in the word choice:* backdoor, backdoor, backdoor, latent backdoor.

Now, functions occur in the writing elsewhere than just in the words themselves. Remember, functions are everywhere.

Functions are also served, for example, by the order which the words are put in. Take, for instance, the opening sentence of the Abstract — the words highlighted yellow function as appraisal of the research topic, *while the words highlighted blue function as* background on the research topic.

Recent work proposed the concept of backdoor attacks on deep neural networks (DNNs), where misclassification rules are hidden inside normal models, only to be triggered by very specific inputs.

Again, much as with the repetition above, let's edit this sentence to disrupt the word order.

The concept of backdoor attacks on deep neural networks (DNNs), where misclassification rules are hidden inside normal models, only to be triggered by very specific inputs, has been propose in recent work.

Once more, I will admit, the edit is extreme. I mean, I've converted a solid sentence into a floppy-loppy sentence.

Still, the edit serves my illustrative purposes, because what I want to show is the overriding function of the opening sentence in this as in most any Abstract. That function is create value. *The opening sentence in nearly every Abstract is to establish value for the topic the authors have conducted research on. Readers want to see, and see right away, the value of the* topic, *or at least the value of the* area *which the topic belongs in. Readers are not going to pay attention to background information if they can't see why they're being asked to pay attention to the background information. It is the writer's job to say why before the writer explains how!*

Your takeaway: First establish value, then provide any background relevant to that value. So, that means, it is value, value, value, *and then* a little background.

Again, just take a look at how well this Abstract achieves value. Once more, for illustrative purposes, here's the original, with the function of appraisal *and then the function of* background*:*

Recent work proposed the concept of backdoor attacks on deep neural networks (DNNs), where misclassification rules are hidden inside normal models, only to be triggered by very specific inputs.

Notice how the research topic actually belongs inside of the appraisal *function. In other words, the readers are hearing about backdoors in connection with valuable work. Because in the edited version here, things change drastically:*

`The concept of backdoor attacks on deep neural networks (DNNs), where misclassification rules are hidden inside normal models, only to be triggered by very specific inputs, has been propose in recent work.`

In this edited version, backdoors are just information. The reader is getting lectured to. And the thin band of yellow highlight which functions as appraisal *really actually sounds more like this: "Oh, right, and all this backdoor stuff matters because of blah, blah, blah."*

Your big takeaway: Choose words according to the function of your research topic — and order those words according to the function of your appraisal of that research.