Read *Dynamic Backdoor Attacks against Machine Learning Models* and See How the Authors Use the System of Identification to Help Readers Stay on Topic

**Abstract**

Readers stray easily off topic because they lack the authors firsthand experience of conducting the experiments, gathering the results, and thinking to interpret the significance. To help readers follow the points throughout a paper, authors are advised to learn how to focus their prose. Focus begins from the selection of research content to cover in one paper or in one portion of the paper. But focus is, as well, eminently achievable through the writing. In this parse, I demonstrate how the paper *Dynamic Backdoor Attacks against Machine Learning Models* achieves such focus through the skillful deployment of grammatical Identification.

1 Introduction

Complex research seems most suited to complex prose. However, the experience of reading such prose leaves most readers baffled. Alternatively, focus in scientific prose impacts readers. They know the topic of any sentence, and they know where any topic comes in the large scheme of the whole text. Basically, focus in the prose causes focus in the reading of that prose.

The exact source of prose focus is a chicken-and-egg kind of problem. Which is the decisive factor in achieving focused prose? Does the focus derive from selections made in the subject matter itself; in other words, is focus not about the writing but about the topic of the writing? Or instead, is the topic irrelevant because focus is achieved through a focussed way of presenting topics through deft composition? The short answer here is both.

The real question, though, is which exact proportions of research topic and prose composition will achieve focus. *That is* a question every researcher must be asking as a project progresses from inception, through intermediary phases, down to submission. It's the asking that matters, not definitive answers, because the
question to the balance of topic selection and prose composition provides itself answer. As the details of the background and the experiments and the results appear ever clearer to the authors' thinking, so too do the authors become, as well, abler to compose that thinking in focussed prose. Constant questioning to what to leave in and what to leave out improve in the abilities of authors to say it straight.

The authors of *Dynamic Backdoor Attacks against Machine Learning Models* demonstrate such abilities. Their paper says what it’s going to say and then sets straight about saying that thing. The paper reports on three innovative techniques for backdoor attacks, and so it keeps to just that one topic: the backdoor attack. The innovation is contrasted to current techniques, but only as much detail is provided as we readers will need if we are to understand precisely where the current techniques are innovated upon by the proposed techniques. In short, the paper reports primarily on triggers and labels.

We cannot know whether the author team began from much complexer thinking and subject matter and progressed to focussed prose. However, just such progression is the development in most projects, especially the ones which end in impactful papers. Every object of research interest demands hard thinking. Few problems in the research are solved nicely and finally. Therefore, author teams everywhere face the challenges of bringing complex science into plain and simple prose. They achieve that through focus.

Focus is eminently visible in the writing. It is possible to parse a paper to see just how the writing contributes to focus. Therefore, my parse of *Dynamic Backdoor Attacks against Machine Learning Models* reveals the system of Identification, that is, the discoursal resources deployed in order to introduce backdoor attacks and the operative controls in such attacks, viz., triggers and labels.

2 Methodology

Identification performs two closely related functions in text.

The first function is to present entities into the discourse. This presenting sort of Identification is marked red here in ¶1:
Machine learning (ML) has made tremendous progress during the past decade and is being adopted in various critical real-world applications. However, recent research has shown that ML models are vulnerable to multiple security and privacy attacks. In particular, backdoor attacks against ML models have recently raised a lot of awareness. A successful backdoor attack can cause severe consequences, such as allowing an adversary to bypass critical authentication systems.

The phrases in red make no presumptions. The reader is not expected to know in the opening paragraph of the Abstract that machine learning or tremendous progress will be entities covered by the ensuing discourse. Therefore, these entities get presented, and this presentation triggers features of indefiniteness in the grammar; that is, the noun phrases have no the or that or this or possessive like their.

The second function performed by Identification follows up on the first and so presumes entities which have already been presented into the discourse. Here the Identification tracks the presumed entities as these reappear throughout the discourse. This presuming sort of Identification is marked blue here in ¶1:

Machine learning (ML) has made tremendous progress during the past decade and is being adopted in various critical real-world applications. However, recent research has shown that ML models are vulnerable to multiple security and privacy attacks. In particular, backdoor attacks against ML models have recently raised a lot of awareness. A successful backdoor attack can cause severe consequences, such as allowing an adversary to bypass critical authentication systems.

The phrases in blue do make presumptions. For instance, the phrase the past decade does presume that a reader will know the date of publication and as well, the ten-year period in which the date belongs. The phrase the past decade is definite, that is, just the opposite sort of grammar to that of presenting Identification.

Identificatory tracking also uses pronouns to presume identity, as in this example from the Introduction (¶6 S28):

As we can see, Badnets in this case uses a white square as a trigger and always places it in the top-left corner of all inputs.

Here the pronoun it tracks the white square through the developing topic, and so the reader learns its location in the images. Pronoun use is not frequent in security papers; therefore, identificatory tracking relies far more heavily on the grammatical
resource of definiteness. However, another resource of the grammar used even more frequently is repetition.

In the discourse of security research, very many entities are generic in reference. Generic reference simply means that the entity being picked out and referred to in the real world is not a specific entity, but a representative one. For example, take again S3 and S4, and note how the underlined phrases have generic reference:

In particular, backdoor attacks against ML models have recently raised a lot of awareness. A successful backdoor attack can cause severe consequences, such as allowing an adversary to bypass critical authentication systems.

The backdoor attacks referred to are not any specific attacks carried out by any individual adversaries on any identified models. No, the attacks are just any backdoor attacks; or to put that another way, the attacks meant are any and all attacks which can be defined as backdoor attacks, to the exclusion of other classes of attack.

The other underlined phrases above prove similarly to be generic, and thus you see how generic reference appears in the grammar as either a plural noun (e.g., backdoor attacks) or as a singular noun without antecedent in the foregoing discourse (e.g., an adversary).

In sum, the system of Identification presents entities into the discourse and continues tracking the entities so that readers can know what is said about which entities. The skillful deployment of Identification resources can, as in Dynamic Backdoor Attacks against Machine Learning Models, create an on-point and focussed handling of complex research.

3 Evaluation

My parse of the system of Identification demonstrates that the authors have selected a finely circumscribed set of details in their work and foreground these and just these details by deploying Identification resources to best effect.
To begin with, here is the organization of the Title and Abstract and Introduction and Preliminaries, where indentation and color represent position in the textual hierarchy, and where the pilcrow indicates *paragraph* and the capital S indicates *sentence*:

**Title**

**Abstract**

1 S1 - S4  
2 S5 - S10  
3 S11 - S13

1. Introduction

4 S14 - S17  
5 S18 - S25  
6 S26 - S32  

1.1. Our Contributions

7 S33 - S40  
8 S41 - S44

Random Backdoor

9 S45 - S47  

Backdoor Generating Network (BaN)

10 S48 - S52  

conditional Backdoor Generating Network (c-BaN)

11 S53 - S56  
12 S59 - S62  

13 S63 - S69  
14 S70  
• S71  
• S72  
• S73

2. Preliminaries

15 S74 - S75  

2.1. Machine Learning Classification

16 S76 - S80  

2.2. Backdoor in Machine Learning Models

17 S81 - S82  
18 S83 - S84  
19 S85 - S87  
20 S88 - S90  
21 S91 - S94  

2.3. Threat Model

22 S95 - S97  
23 S98 - S101
Immediately apparent is the decision by the authors to distinguish their contributions above and beyond the conventional bulleted list in §14. No, the authors create instead a subsection, they call it *Our contributions*, and they mark off the three innovative techniques even further by paragraph headings. Essentially, we readers enter the authors’ actual study at the very moment we cross that threshold, 1.1. *Our Contributions*.

Next, open **BgrTxt_9_paper_parsed** — it’s the other download on this post. Through the first four sections, I have highlighted the Identification of these two content points:

- **backdoor attacks**: So, that is for example, the new techniques Random Backdoor, Backdoor Generating Network (BaN), and conditional Backdoor Generating Network (c-BaN), but also all other backdoor attacks as well as their counterpart, the defense against backdoor attacks. Basically, the yellow highlights pick up noun phrases that identify the macro-level technicality of the research focus.

- **triggers and labels**: Here we’re at the next step down in generality, that is, we are reading about the meso-level technical details crucial to the operation of backdoor attacks. It is here that the innovation of the proposed techniques can be demonstrated and understood.

A glance at the highlighted text makes this much immediately apparent: Something like 40% is identifiably backdoor attacks or triggers and labels. The readers of *Dynamic Backdoor Attacks against Machine Learning Models* stay on topic because the paper gives them few other topics to be straying off to. In fact, most stretches of text that are not highlighted are precisely those points of detail the authors want to connect to the two main topics of backdoor attacks and triggers and labels. Take, for example, §7:

In this work, we propose the first class of backdooring techniques against deep neural networks (DNN) models that generate dynamic triggers, in terms of trigger pattern and location. We refer to our techniques as dynamic backdoor attacks. Dynamic backdoor attacks offer the adversary more flexibility, as they allow triggers to have different patterns and locations. Moreover, our techniques largely reduce the efficacy of the current defense mechanisms demonstrated by our empirical evaluation. Figure 1b shows an example of our dynamic backdoor attacks implemented in a model trained on the CelebA dataset [28]. In addition, we extend our techniques to work for all labels of the backdoored ML model, while the current backdoor attacks only focus on a single or a few target labels. This further increases the difficulty of our backdoors being mitigated.
To be begin with, there are at total 130 words, and 70 of these are highlighted. That means just over half the paragraph is topic subject-matter. Moreover, note how the topic subject-matter is extended only by the most integral of details. Thus, backdooring techniques generate triggers, and dynamic backdoor attacks offer flexible adversarial modes of operation, and also, triggers acquire different patterns and locations. None of these details are unrelated to the topics nor irrelevant to this stage in the discourse. Essentially, you read in these details the micro-level technicality that completes the explanation of the meso- and macro-level technicalities.

Certain stretches of text are conspicuously uncolored, meaning the system of Identification there is non-operative. For instance, ¶1 is almost totally devoid of highlighting, while ¶4 is totally devoid. However, it's not hard to surmise why this should be. At both of these stages in the discourse (the opening paragraphs of Abstract and Introduction), the topic subject-matter has yet to be properly established or even introduced, and so the grammar cannot yet present or presume those topics which will become the mainstay of the remainder of the paper.

Particularly effective use of the Identification is made in ¶9 through ¶12, the stage in the discourse where the three new techniques are first detailed. I have mentioned the focus achieved through the paragraph headings, but note, as well, how it is only ¶10 which presents large stretches of text un by Identification. There is good reason for this.

Look again at ¶10 here:

Backdoor Generating Network (BaN): In our second technique, we propose a generative ML model, i.e., BaN, to generate triggers. To the best of our knowledge, this is the first backdoor attack which uses a generative network to automatically construct triggers, which increases the flexibility of the adversary to perform backdoor attacks. BaN is trained jointly with the backdoor model, it takes a latent code sampled from a uniform distribution to generate a trigger, then place it at a random location on the input, thus making the trigger dynamic in terms of pattern and location. Moreover, BaN is essentially a general framework under which the adversary can change and adapt its loss function to their requirements. For instance, if there is a specific backdoor defense in place, the adversary can evade the defense by adding a tailored discriminative loss in BaN.
First off, the opening to S49 draws explicit attention to one research contribution, though the remainder of the sentence does, expectedly, restate the advantage of flexibility for adversarial modes of operation. S50 provides the essential details about model construction, which really are the same as for Random Backdoor, and therefore are again to be expected. But S51 and S52 do develop new and crucial points of detail about tailored discriminative loss. Here a lot of points are expanded beyond the established core topics of backdoor attacks more generally and the triggers and labels in specific. However, this expansion of the detail is required if the point of tailored discriminative loss is to be made understandable.

The authors of *Dynamic Backdoor Attacks against Machine Learning Models* remain in control of their topic subject-matter, and by doing so, they enable their readers to remain in control as well.

**Conclusion**

This parse has demonstrated by the paper *Dynamic Backdoor Attacks against Machine Learning Models* just how Identification can be deployed (A) to arrange topic subject-matter in the discourse and (B) to assist readers in their progression through the discourse and into the research contributions of the author team. Focus is brought to prose by pre-selection of the topics to cover, but prose is brought to a focus by the skillful deployment of Identification resources in the grammar.