Adversarial Machine Learning: A Literature Review

Sam Thomas(✉) and Nasseh Tabrizi

East Carolina University, Greenville, NC 27858, USA
thomass08@students.ecu.edu, tabrizim@ecu.edu

Abstract. Machine learning is becoming more and more utilized as a tool for businesses and governments to aid in decision making and automation processes. These systems are also susceptible to attacks by an adversary, who may try evading or corrupting the system. In this paper, we survey the current landscape of research in this field, and provide analysis of the overall results and of the trends in research. We also identify several topics which can better define the categorization.

Keywords: Adversarial machine learning · Literature survey

1 Introduction

Machine learning is the process by which a machine can learn to make decisions without being explicitly told what to do. This has been a boon to automation and data science as a whole. However, machine learning systems are not inherently robust. An entity that wishes to harm or evade an unguarded system's decision-making capabilities can do so with relative ease. This is the conceptual foundation of adversarial machine learning [1].

Adversarial machine learning (AdvML), broadly speaking, is where a machine learning system (i.e. the classifier) is in an adversarial environment – one in which it is challenged by some adversarial opponent. These opponent input samples, which have been designed to disrupt the performance of the ML system, alter the legitimate input samples by tricking the classifier into misclassifying the input. In fact, according to [1] machine learning systems can, and often are, trained to generate adversarial samples to use against the classifier Although the measures can be taken to protect a machine learning system, the protection is not total and not ensured to last. Thus, as reported in [2] this still remains an open problem.

Generally, AdvML is applicable wherever there is a machine learning system that is accessible to would-be attackers. Notably, it has applications in biometric verification, spam detection, malware detection, and the detection of network intrusions. In addition, generative adversarial networks (GAN) [1, 4, 5] have shown that they can be applied to a wide variety of tasks, such as medical image processing, image censoring, language generation, and learning representations of emotional speech, to name a few. These applications go beyond the scope of simply protecting a classifier from adversarial examples.

P. Perner (Ed.): MLDM 2018, LNAI 10934, pp. 324–334, 2018.
https://doi.org/10.1007/978-3-319-96136-1_26

As stated earlier, if a classifier is not protected from adversaries, then it is very susceptible to their attacks, resulting in misclassification rate of over 96% [3, 4, 8]. On-line machine learning systems are susceptible to being corrupted over time, and have been shown to become more inaccurate as the number of adversarial samples increases [6], and as reported in [7] there is also the ability to fool autonomous robotic patrolling by using game-theoretic approaches.

Our research offers the following contributions:

- A survey of research papers on the subject of "adversarial machine learning". The papers are counted and sorted by categories, and that data is compared and trends are analyzed.
- A refinement of the categorization of topics within the field of adversarial machine learning.

2 Related Works

This paper adapts the categorization used in a survey done by Kumar and Mehta of IBM Research, India [9]. The categories and descriptions can be seen in Fig. 1. We have outline the categories below, which include the adaptation we made for the purposes of this literature review. This categorization was chosen because it provides the basis for a deeper taxonomy than has been proposed by any other paper we found.

The categories and subcategories that we used are as follows:

Early Work

Applications: Spam filters, anti-virus/ malware, network intrusion detection, biometric verification and authentication

Approaches:

Game-theory based approaches – game between classifier and adversary. optimal classifier to automatically adjusts to adversary's evolving inputs; find the equilibrial prediction models

Signature-based intrusion detection systems - polymorphic techniques to generate attack instances that do not share fixed signatures - attackers can use polymorphic techniques to make attack instances look different from each other.

Polymorphic blending attacks - can evade byte-frequency based network anomaly intrusion detection systems by matching the statistics of mutated attack samples with normal samples.

Making classifiers robust - not assign too much weight to a single feature; game theoretic formalization to avoid over weighting single feature.

Multimodal biometric systems - likelihood ratio based fusion scheme and fuzzy logic based fusion scheme

Attacks: *exploratory, evasion* and *poisoning*

Exploratory attacks:

Model Inversion: black box access to model and some demographic information about a person, an attacker can predict private information such as genetic markers from a healthcare system.

Inferring information: an adversary can infer statistical properties from the relationship among dataset entries

Membership inference attack: given the black box access to model and a data sample, it can be inferred whether that data record was part of the training set or not.

Model Extraction using Online APIs: local models can be built that function very similar to proprietary models for which only API access is available. Confidence score and partial values for API are used to find key coefficients of the model.

Evasion attacks:

Adversarial Examples: systematic adversarial perturbation; DeepFool algorithm to efficiently compute perturbations that fool deep networks

Generative Adversarial Networks (GANs): simultaneously training two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G.

Adversarial classification: Learning to distinguish good inputs from malicious ones is known as adversarial classification. Useful in adversarial training.

Text-based systems: text classification systems can be fooled by carefully inserting, modifying or removing some text such that the meaning of text does not change for a human user.

Poisoning attacks:

Network Intrusion Detection: input samples to disturb the balance between false positives and false negatives therefore reducing effectiveness.

Support Vector Machine Poisoning: label flips attack; adversary can predict the change of the SVM's decision function to some extent by using malicious input. This can be used to craft malicious samples.

Defensive Distillation: Distillation is a training procedure that was designed to train a DNN using knowledge transferred from a different DNN. This technique is used for defense training.

Fig. 1. Category summary, adapted from [9].

- Applications
  - Spam Filters
  - Anti-virus/Malware Detection
  - Network Intrusion Detection
  - Biometric Verification and Authentication
  - Ill-Fit
- Approaches
  - Game-theoretic Approaches
  - Signature-based Intrusion Detection Systems
  - Polymorphic Blending Attacks
  - Making Classifiers Robust
  - Multimodal Biometric Systems
  - Ill-Fit
- Attacks
  - Exploratory Attacks:
    - Model Inversion
    - Inferring Information
    - Membership Inference Attack
    - Model Extraction using Online APIs
    - Ill-Fit
  - Evasion Attacks:
    - Adversarial Examples
    - Generative Adversarial Networks (GANs)
    - Adversarial Classification
    - Text-based Systems
    - Ill-Fit
  - Poisoning Attacks:
    - Network Intrusion Detection
    - Support Vector Machine Poisoning
    - Defensive Distillation
    - Ill-Fit
  - Ill-Fit

3 Methodology

3.1 Overview

Our process for categorizing recent research papers is as follows:

1. Collect the top 100 results from the sources, using the phrase ‘adversarial machine learning’ as the search query.
2. Review each paper.
3. Sort each paper into categories.
4. Tally the raw values of each category.

5. Tally the per-year values of each category (excluding Cornell Digital Library [https://arxiv.org/]).
6. Visualize the data with charts.

3.2 Collection

In this study we have used multiple collection sources to get a robust data set. The sources we used were ACM Digital Library, IEEE Xplore Digital Library, Springerlink, Cornell Digital Library, and Sciencedirect. Since we are focused on newer developments within this field of research, we restricted our data collection to the past ten years (2007–2017).

3.3 Data-Set Restrictions

Here we have listed the details of our collection methodology as a whole. Also, in order to get accurate results, some sources needed search criteria that was specific to them, and so those criteria are listed here, as well as any quirks specific to those sources.

General

- Collection took place: 8/28/17–9/11/17
- Language: English
- Papers published: 2007–2017
- Our university's access to the collections was used to gather the papers.
- Only those papers which were acquirable were collected (i.e. no "abstract only" or pay-to-access entries)
- Papers that are not about adversarial machine learning, but merely mention it, have been classified as "unrelated". These papers were collected, but during evaluation they have been sorted under "discarded".

IEEE Xplore Digital Library

- Only 91 papers matched the search query

Cornell Digital Library

- Experimental full text search in the subject Computer Science

Springer Link

- Searched under the discipline of Computer Science
- Did not include "Preview-Only Content"
- Sorted by relevance

3.4 Evaluation

All papers were filtered and categorized manually. We filtered out book chapters, conference posters, and those papers which were unrelated to the subject matter. We used the categories in such a way, that a paper can appear in multiple categories. Some papers were categorized as a top-level category, but did not fit into any of its sub-categories. We tallied these papers in their own subcategory that is independent of the other subcategories.

The histograms for each category exclude the results from Cornell Digital Library due to the fact that the earliest published papers are from 2015, and so including these results would heavily skew the graph.

4 Results

Here we present the data that we have collected. These have been divided into sections for "Pre-Sort", "Raw Counts", and "Trends".

4.1 Pre-sort

The first step was to refine the data set which would be sorted. For this purpose, we used the following categories:

- Disregard: disregarded papers, such as conference posters, conference abstracts, and book chapters.
- Duplicates: duplicate papers.
- Meta: papers such as surveys, literature reviews, or topic overviews.
- Tools: papers which only discussed a tool that was being demonstrated.
- Unrelated: papers that are unrelated to "Adversarial Machine Learning".
- Sorted: those papers which would be categorized.

The results of these refinements can be seen in Fig. 2.

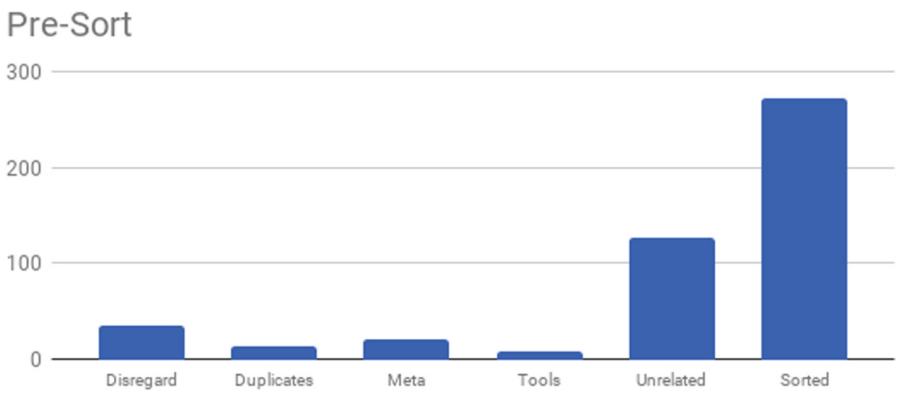

Fig. 2. Count of papers for the Pre-Sort step.

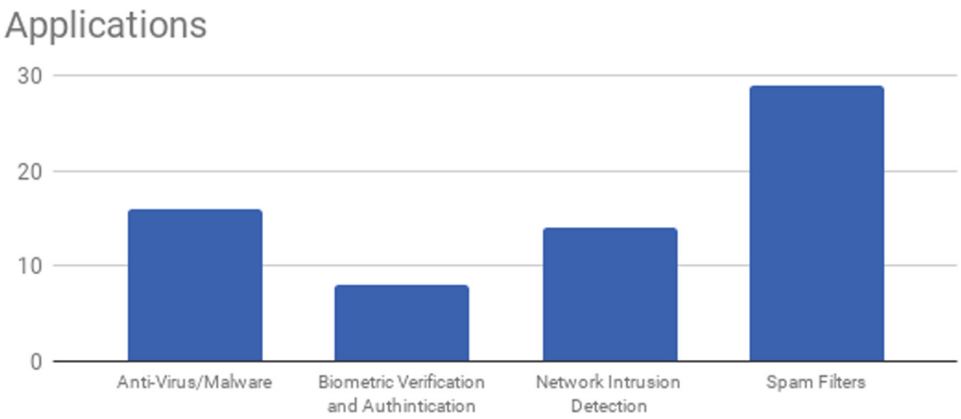

Fig. 3. Count of papers for "Applications"

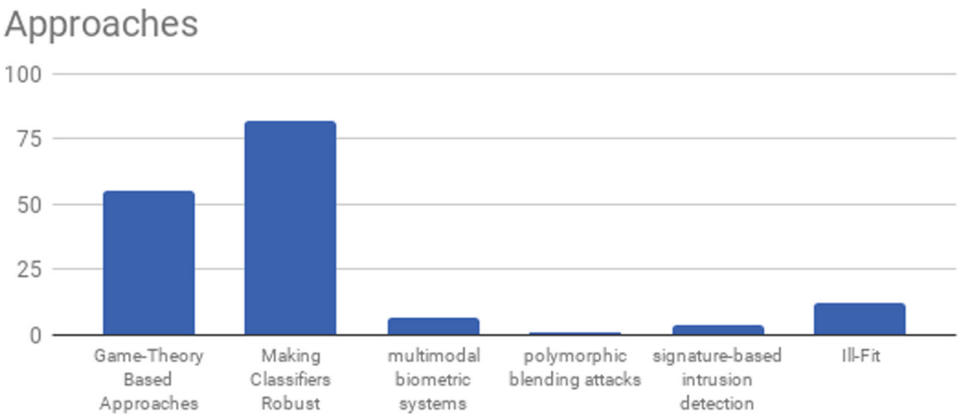

Fig. 4. Count of papers for "Approaches"

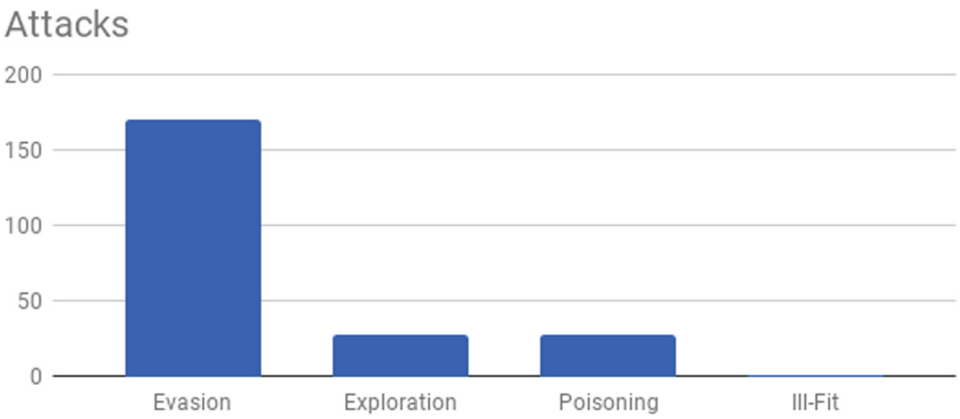

Fig. 5. Count of papers for "Attacks"

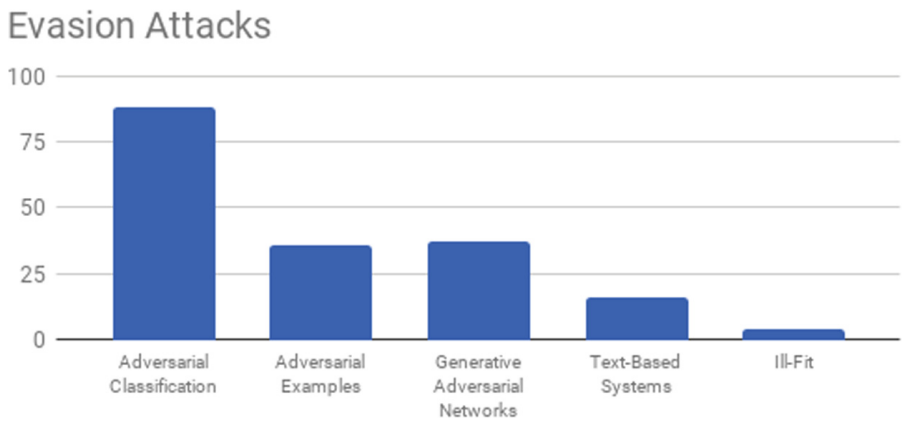

Fig. 6. Count of papers for "Evasion Attacks" sub-category

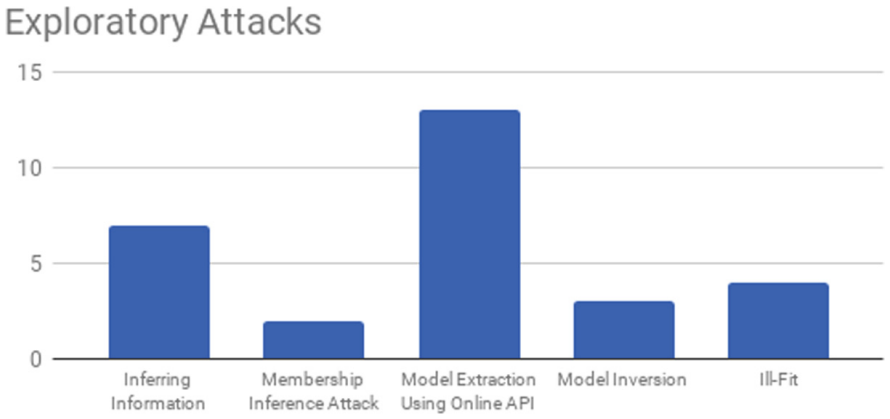

Fig. 7. Count of papers for "Exploratory Attacks" sub-category

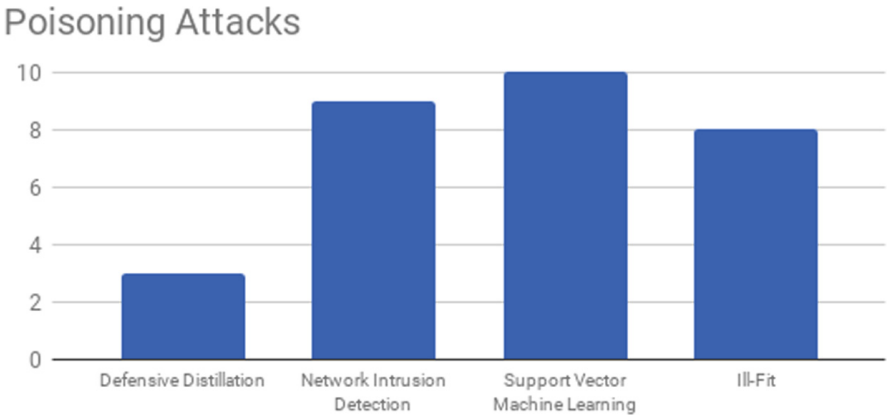

Fig. 8. Count of papers for "Poisoning Attacks" sub-category

4.2 Raw Counts

The raw paper counts for the subcategories are shown in Figs. 3, 4, 5, 6, 7 and 8. Figures 3, 4 and 5 show the counts for top-level categories, while Figs. 6, 7 and 8 show the counts for each type of attack.

4.3 Trends

The counts per year can be seen in Figs. 9, 10, 11, 12, 13, 14, 15 and 16. These figures are a subset of all the trends which were analyzed. For the sake of readability and brevity, we have included the most noteworthy trend charts in this paper, and excluded those which had sparse or sporadic paper counts.

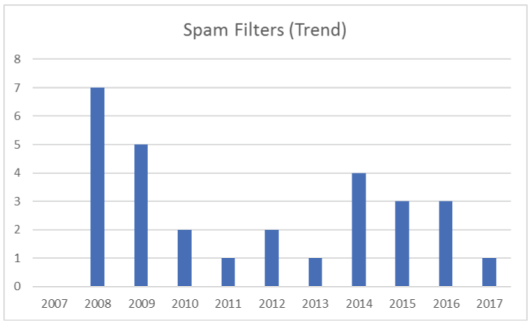

Fig. 9. Paper counts by year for "Spam Filters".

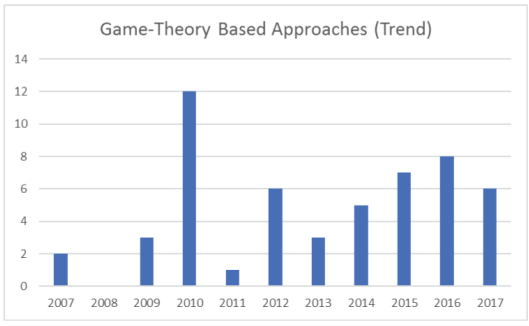

Fig. 10. Paper counts by year for "Game-Theory Approaches".

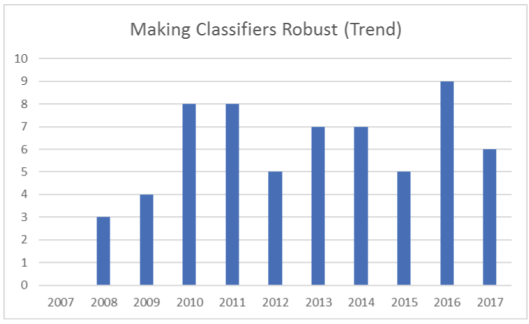

Fig. 11. Paper counts by year for "Making Classifiers Robust".

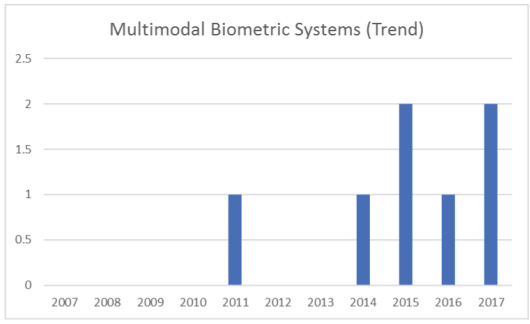

Fig. 12. Paper counts by year for "Multimodal Biometric Systems".

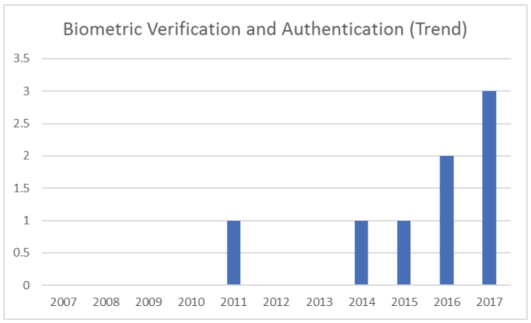

Fig. 13. Paper counts by year for "Biometric Verification and Authentication".

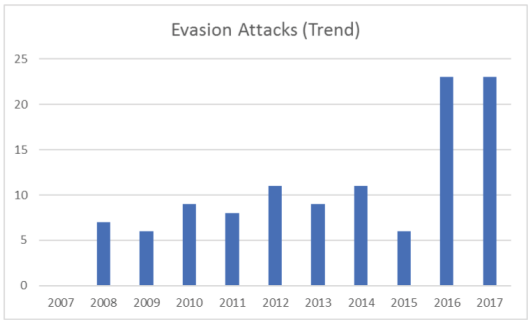

Fig. 14. Paper counts by year for "Evasion Attacks".

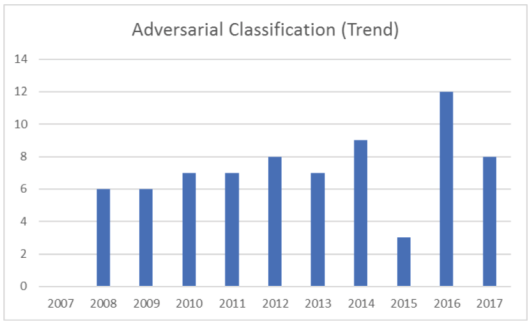

Fig. 15. Paper counts by year for "Adversarial Classification".

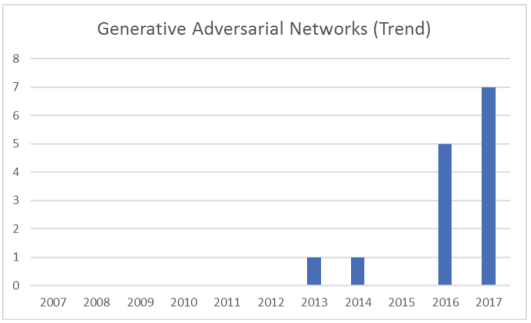

Fig. 16. Paper counts by year for "Generative Adversarial Networks".

5 Discussion

5.1 Raw Count Data Analysis

From Figs. 3, 4, 5, 6, 7 and 8 we can see the following:

- "Spam Detection" is by far the most popular application, with ∼41% of papers exploring the topic.
- "Making Classifiers Robust" and "Game-Theory Based Approaches" are the most common approaches, making up approximately 51% and 34% of the papers, respectively.
- ∼76% of papers discussing attacks focus on "Evasion" attacks.
- Regarding evasion attacks, the most explored topic was "Adversarial Classification", have ∼49% of papers which discuss it. There were also the topics of "Adversarial Examples" and "Generative Adversarial Networks" which were somewhat popular, at ∼20%, each.
- Regarding exploratory attacks, the most explored aspect was "Model Extraction Using Online APIs", which made up ∼45% of papers in that category.

- For the "Ill-Fit" category, we see that "Poisoning Attacks" has the largest proportion of ill-fit papers, with ∼27% falling into the sub-category. However, it is the "Approaches" category which has the highest raw count, at 12 ill-fitting papers.

These are some of the notable topics that were discovered during evaluation but were designated as "ill-fit":

- *Applications*: Privacy
- *Approaches*: Feature Squeezing
- *Approaches*: Domain Adaptation
- *Poisoning Attacks*: Online Neural Networks.

5.2 Trends Data Analysis

By looking at the breakdown of each category's paper-counts by year, we can see the following trends in Figs. 9, 10, 11, 12, 13, 14, 15 and 16:

- "Spam Filters" had a large spike in 2008, before declining, and then rising slightly in popularity.
- "Game-Theory Based Approaches" had a large spike in 2010, before lowering and remaining stable.
- "Making Classifiers Robust" has been a consistently explored topic since 2008.
- Figures 12, 13 show that AdvML in biometric systems has become more popular in recent years.
- Evasion attacks have had steady interest since 2008, with a sharp rise within the past 2 years.
- "Adversarial Classification" has had steady interest since 2008.
- "Generative Adversarial Networks" have garnered a large amount of interest in just the past few years.

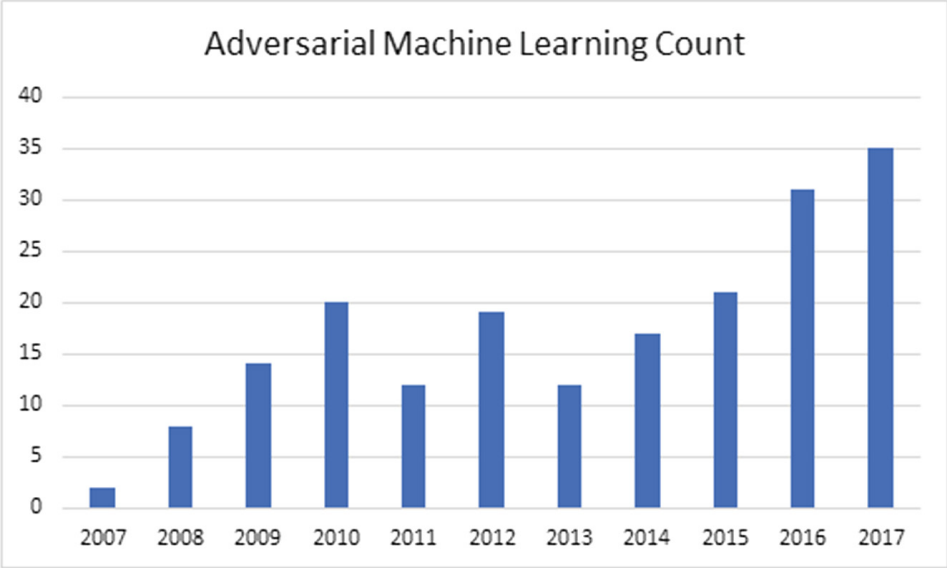

Fig. 17. Paper counts by year for all papers in the field of Adversarial Machine Learning

And finally, we see in Fig. 17 the 10-year trend in the topic of Adversarial Machine Learning as a whole.

Together, these charts show an initial wave of interest, which peaks in 2010. Interest in the field levels off for a few years, but then begins to steadily gain more and more interest from 2013 onward.

6 Conclusion

In this paper, we have presented data on the current state and trends of the field of "Adversarial Machine Learning". To this end, we collected, sorted, and analyzed 475 research papers from five different sources.

We found that certain topics are more popular than others, and that some topics have only recently started to gain interest within this field.

Notably, we have identified that, while there has been steady interest in the subject for a while, AdvML is rapidly gaining in popularity, and that this popularity is evident in the trends of lesser-researched topics within the field.

We also identified four topics within the field which should be regarded when categorizing the subject. These topics are privacy, feature squeezing, domain adaptation, and online neural networks.

References

1. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I.P., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security Artificial Intelligence, pp. 43–58 (2011)
2. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016, pp. 582–597 (2016)
3. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. CoRR abs/1605.07277 (2016)
4. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. CoRR abs/1602.02697(2016)
5. Biggio, B., Fumera, G., Roli, F.: Adversarial pattern classification using multiple classifiers and randomisation. In: da Vitoria Lobo, N., et al. (eds.) SSPR /SPR 2008. LNCS, vol. 5342, pp. 500–509. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89689-0_54
6. Biggio, B., Corona, I., Fumera, G., Giacinto, G., Roli, F.: Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In: Sansone, C., Kittler, J., Roli, F. (eds.) MCS 2011. LNCS, vol. 6713, pp. 350–359. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21557-5_37
7. Villacorta, P.J., Pelta, D.A.: Exploiting adversarial uncertainty in robotic patrolling: a simulation-based analysis. In: Greco, S., et al. (eds.) IPMU 2012 Part IV. CCIS, vol. 300, pp. 529–538. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31724-8_55
8. Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: Proceedings - 2016 IEEE European Symposium Security Privacy, EURO S P 2016, pp. 372–387 (2016)
9. Kumar, A., Mehta, S.: A survey on resilient machine learning. CoRR, abs/1707.03184 (2017)