به نام خدا

تکلیف دوم روش پژوهش و ارائه

یافتن، اعتبارسنجی و پالایش منابع

محمد جواد زندیه ۹۸۳۱۰۳۲

۲۸ بهمن ۱۴۰۰

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

بخش اول

مراحل جست و جو مقالات

موضوع تحقیق

"بررسی حملات خصمانه و دفاع در یادگیری عمیق"

"Adversarial Attacks and defenses in Deep Learning"

بررسی موضوع در چند جمله

۱. مفاهیم چه هستند؟ مقصود اصلی چیست؟ در حوزه یادگیری عمیق به دنبال یافتن الگوریتم هایی هستیم که می توانند به عنوان حمله به شبکه عصبی شناخته شوند یعنی فرایند یادگیری[1] و تشخیص[2] در شبکه را مختل کنند. همچنین به دنبال الگوریتم هایی هستیم که بتواند این حملات را خنثی کرده و یا اثرات آن ها را کاهش دهند.

۲. آیا موضوع بخشی از یک موضوع وسیع تر است؟ بله، این موضوع در حیطه یادگیری عمیق و شبکه های عصبی می باشد.

۳. آیا یک حالت خاص تر از موضوع وجود دارد؟ بله، بررسی هر یک از الگوریتم ها به صورت جداگانه میتواند یک حالت خاص از این موضوع باشد.

۴. کلمات متمایز کننده(کلید واژه ها) در موضوع چه هستند؟ یادگیری عمیق/ شبکه عصبی عمیق/بینایی ماشین/پردازش تصویر/ حملات/ خصمانه/ دفاع

ترکیب کلید واژه ها

نکته: در جست و جو باید ابتدا از ترکیب های بسیار خاص و جزئی شروع به جست و جو کنیم و سپس به جست و جوی ترکیب های کلی تر بپردازیم.

۱. حملات در یادگیری عمیق
۲. دفاع در یادگیری عمیق
۳. حملات خصمانه در یادگیری عمیق
۴. حملات خصمانه و دفاع در یادگیری عمیق
۵. الگوریتم های پرکاربرد در حملات خصمانه به شبکه های عصبی عمیق
۶. الگوریتم های پرکاربرد دفاعی در برابر حملات خصمانه به شبکه های عصبی عمیق
۷. حملات خصمانه به روز در برابر الگوریتم های جدید دفاعی در شبکه های عصبی عمیق
۸. بررسی الگوریتم های دفاع در بینایی ماشین به منظور تعیین نقاط ضعف در برابر حملات خصمانه

[1] Learning
[2] Classification /Regression

۹. مقایسه میزان کارایی الگوریتم های دفاعی در برابر حملات خصمانه به شبکه های عصبی عمیق در کاربرد های تشخیص چهره و پردازش تصویر

جست و جوی کلید واژه ها در موتور های جست و جوی مختلف

با توجه به اینکه موتور های جست و جو مختلف پایگاه های اطلاعاتی و الگوریتم های رنک بندی متفاوتی دارند در نتیجه نیاز است تا ترکیبات کلید واژه ها در موتور های مختلف جست و جو شوند.(ترکیبات بالا به صورت انگلیسی جست و جو شده اند)

از موتور های **جست و جوی آکادمیک** استفاده شده است تا نتایج بهتری حاصل شود.(البته به غیر از Excite که موتور جست و جوی آکادمیک نیست)

1. Google Scholar[3]

* Search: Comparison of the efficiency of defense algorithms against adversarial attacks on deep neural networks in face recognition and image processing applications

R01: "The secret revealer: Generative model-inversion attacks against deep neural networks":

https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhang_The_Secret_Revealer_Generative_Model-Inversion_Attacks_Against_Deep_Neural_Networks_CVPR_2020_paper.pdf

* Search: Investigate defense algorithms in computer vision to determine weaknesses against adversarial attacks

R02: "Defense against adversarial attacks using high-level representation guided denoiser":
https://openaccess.thecvf.com/content_cvpr_2018/papers/Liao_Defense_Against_Adversarial_CVPR_2018_paper.pdf

* Search: up-to-date adversarial attacks against new defense algorithms in deep neural networks

R03: "On Evaluating Adversarial Robustness":

https://arxiv.org/pdf/1902.06705.pdf

2. Excite[4]

* Search: Frequently used defense algorithms against adversarial attacks on deep neural networks

R04: "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks":

https://arxiv.org/pdf/1511.04508v2.pdf

[3] https://scholar.google.com/
[4] https://www.excite.com/

* Search: Frequently used adversarial attacks on deep neural networks

R05: "Simple Black-Box Adversarial Attacks on Deep Neural Networks":

https://openaccess.thecvf.com/content_cvpr_2017_workshops/w16/papers/Kasiviswanathan_Simple_Black-Box_Adversarial_CVPR_2017_paper.pdf

3. BASE[5]

* Search: Adversarial attacks and defenses in deep learning

R06: "Adversarial Attacks and Defenses in Deep Learning":

https://reader.elsevier.com/reader/sd/pii/S209580991930503X?token=B1693B587122805195CF5414FD8A449A5766E34ACD48932D1A88455F52BD0A5896BF72D7A268A4EAF71BE8BB40D21D01&originRegion=eu-west-1&originCreation=20220301070703

* Search: Adversarial attacks in deep learning

R07: "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey":

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8294186

4. PubMed[6]

* Search: Defense in deep learning

R08: "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review":

https://link.springer.com/content/pdf/10.1007/s11633-019-1211-x.pdf

* Search: Attacks in deep learning

R09: "Towards Deep Learning Models Resistant to Adversarial Attacks":

https://arxiv.org/pdf/1706.06083.pdf%E4%B8%AD%E6%9C%89%E4%BD%93%E7%8E%B0%EF%BC%8C%E4%BB%A5%E5%90%8E%E8%AF%B4%E5%88%B0CW%E6%94%BB%E5%87%BB%E5%86%8D%E7%BB%86%E8%AF%B4%E3%80%82

R10: "Boosting Adversarial Attacks with Momentum":

https://openaccess.thecvf.com/content_cvpr_2018/papers/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.pdf

[5] https://www.base-search.net/
[6] https://pubmed.ncbi.nlm.nih.gov/

دقت به معیار های اعتبار سنجی مقالات

به منظور انتخاب و یا رد مقالات به عنوان مرجع، باید به شاخص ها دقت شود که در **بخش دوم** ذکر شده اند.

انواع مقالات موجود در زمینه تحقیق

منابع استخراج شده طبق مراحل ذکر شده در قسمت قبل:

R01: "The secret revealer: Generative model-inversion attacks against deep neural networks"

R02: "Defense against adversarial attacks using high-level representation guided denoiser"

R03: "On Evaluating Adversarial Robustness"

R04: "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks"

R05: "Simple Black-Box Adversarial Attacks on Deep Neural Networks"

R06: "Adversarial Attacks and Defenses in Deep Learning"

R07: "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey"

R08: "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review"

R09: "Towards Deep Learning Models Resistant to Adversarial Attacks"

R10: "Boosting Adversarial Attacks with Momentum"

بخش دوم

اعتبارسنجی منابع با معیار های (Impact Factor, Number of Citations, journal h-index)

Research	Year	Journal Impact Factor	Citations	Journal	Journal h-index[7]	SJR	Total Cites (3years)
R01	2020	45.17	102	J01	406	4.658	132568
R02	2018	45.17	496	J01	406	4.658	132568
R03	2019	-	483	J02	-	-	-
R04	2016	19.03	2313	J03	132	2.407	5481
R05	2017	20.97	193	J04	279	4.133	57027
R06	2020	-	184	J05	20	-	-
R07	2018	4.48	1200	J06	127	0.587	116691
R08	2020	3.17	248	J07	36	0.655	665
R09	2017	-	5050	J02	-	-	-
R10	2018	45.17	1132	J01	406	4.658	132568

[7] https://scholar.google.com/citations?view_op=top_venues&hl=en

Journals List:

J01: “Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition”

J02: “arXiv pre-print[8]”

J03: “IEEE symposium on security and privacy (SP)”

J04: “Proceedings of the IEEE International Conference on Computer Vision”

J05: “Engineering”

J06: “Ieee Access”

J07: “International Journal of Automation and Computing”

این اطلاعات از سایت های https://www.scimagojr.com/ و https://www.resurchify.com/ گرفته شده است.

بخش سوم

دیگر معیار های ارزیابی کمی منابع

SJR (SCImago Journal Rank):

اندازه گیری تاثیر علمی ژورنال ها می باشد، به این صورت که هم تعداد استناد(citation) های دریافت شده توسط یک ژورنال و هم اهمیت یا اعتبار ژورنال را که استناد ها از آنجا آمده اند، محاسبه می کند.

Total Cites (3years):

تعداد استناد های دریافت شده در سال انتخابی توسط یک مجله به اسناد منتشر شده در سه سال قبل.

اعتبارسنجی منابع با معیار های جدید

در جدول بخش دوم اعتبارسنجی با معیار های جدید SJR و Total Cites(3 years) قرار داده شده است.

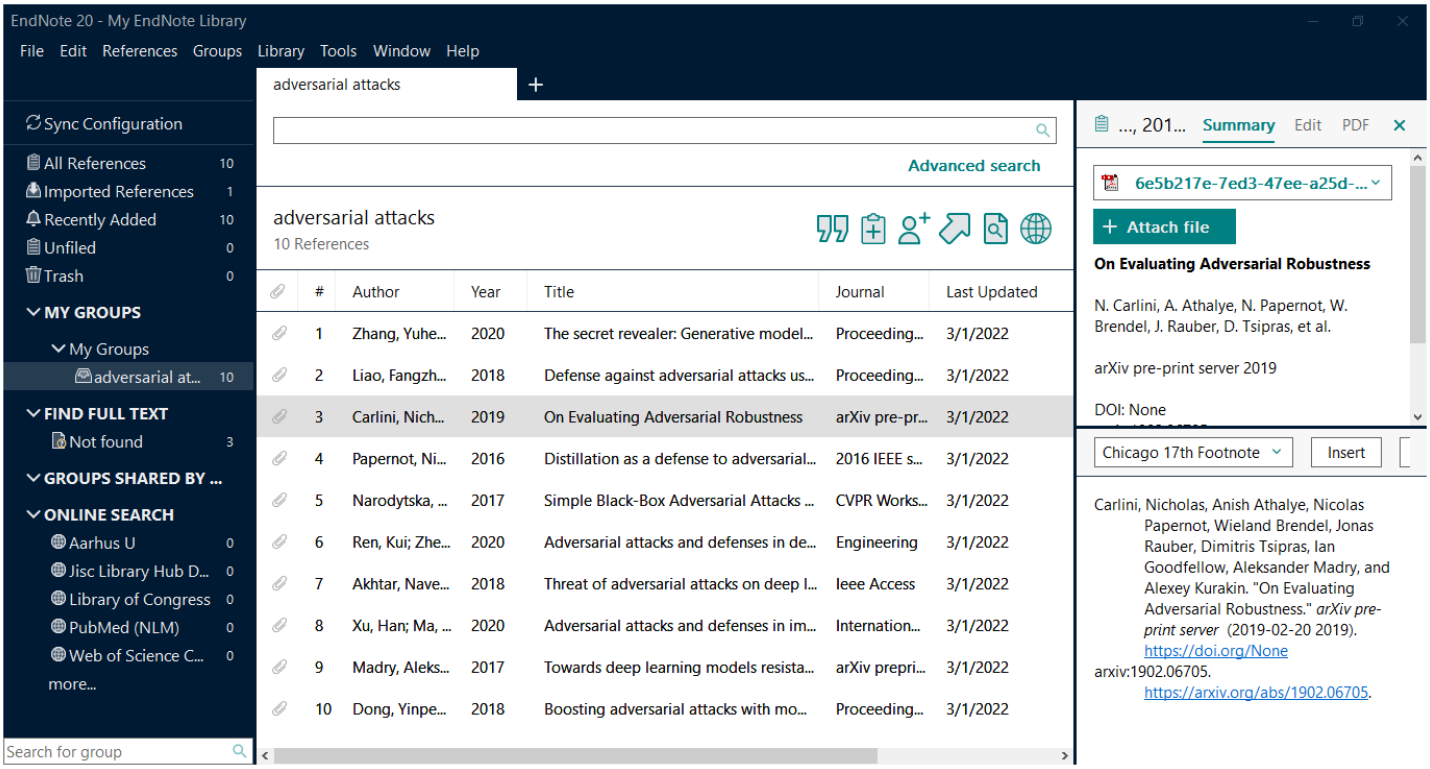
بخش چهارم

لیست مقالات منتخب بر اساس اولویت

مدیریت و لیست کردن مقالات توسط EndNote صورت گرفته است و از Chicago 17th Footnote به عنوان style نمایش منابع استفاده شده است.

منابع اصلی و فرعی به ترتیب اولویت شماره گذاری شده اند.

[8] https://en.wikipedia.org/wiki/ArXiv



منابع اصلی

[1] Akhtar, Naveed, and Ajmal Mian. "**Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.**" Ieee Access 6 (2018): 14410-30.

[2] Papernot, Nicolas, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "**Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks**." Paper presented at the 2016 IEEE symposium on security and privacy (SP), 2016.

[3] Ren, Kui, Tianhang Zheng, Zhan Qin, and Xue Liu. "**Adversarial Attacks and Defenses in Deep Learning**." Engineering 6, no. 3 (2020): 346-60.

[4] Xu, Han, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. "**Adversarial Attacks and Defenses in Images, Graphs and Text: A Review**." International Journal of Automation and Computing 17, no. 2 (2020): 151-78. https://link.springer.com/content/pdf/10.1007/s11633-019-1211-x.pdf.

[5] Narodytska, Nina, and Shiva Prasad Kasiviswanathan. "**Simple Black-Box Adversarial Attacks on Deep Neural Networks.**" Paper presented at the CVPR Workshops, 2017.

[6] Liao, Fangzhou, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. "**Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser.**" Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[7] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "**Towards Deep Learning Models Resistant to Adversarial Attacks.**" arXiv preprint arXiv:1706.06083 (2017).

منابع فرعی

[8] Dong, Yinpeng, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. "**Boosting Adversarial Attacks with Momentum**." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[9] Zhang, Yuheng, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. "**The Secret Revealer: Generative Model-Inversion Attacks against Deep Neural Networks**." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[10] Carlini, Nicholas, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. "**On Evaluating Adversarial Robustness**." arXiv pre-print server (2019-02-20 2019).