به نام خدا

تکلیف اول روش پژوهش و ارائه

انتخاب موضوع تحقیق

محمد جواد زندیه ۹۸۳۱۰۳۲

۲۸ بهمن ۱۴۰۰

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

بخش اول

عنوان

عنوان فارسی

"بررسی حملات خصمانه و دفاع در یادگیری عمیق"

عنوان انگلیسی

"Adversarial Attacks and Defenses in Deep Learning"

بیان مسئله

در کنار توجه به گسترش تکنیک های یادگیری عمیق در حوزه هوش مصنوعی، باید موضوع قدرتمندی این الگوریتم ها در برابر حملات خصمانه[1] را هم مورد توجه قرار داد. این موضوع از آن جهت دارای اهمیت است که کارایی این تکنیک ها در عمل و در محیط های غیر شبیه سازی شده به چالش کشیده می شود و اگر سیاست دفاعی مناسبی را در برابر حملات مخرب از خود نشان ندهند پذیرفته نخواهند بود. علاوه بر هوش مصنوعی این موضوع در حوزه های دیگری نیز مطرح می باشد، از این رو متخصصان حوزه های مختلف در تلاش برای گسترش آن می باشند.

از جمله عوامل موثر در حل این مشکل میتوان به بهبود قدرتمندی شبکه عصبی با آموزش آن با نمونه های ورودی خصمانه پرداخت تا شبکه یاد بگیرد که ورودی های مخرب را شناسایی کند و در آموزش از آنها استفاده نکند.

در این تحقیق برخی از تکنیک های حمله شناخته شده در حوزه پردازش تصویر و متن[2] را بیان می کنیم، در ادامه نیز به بررسی تعدادی از تکنیک های دفاع در برابر این حملات می پردازیم و همچنین برخی از مسائل باز[3] و چالش های آن در این حوزه را مطرح می کنیم. در انتها نیز مقایسه ای خواهیم داشت میان تکنیک های حمله و دفاع تا وضعیت کنونی[4] این موضوع و میزان پیشرفت در هر یک از این دو بخش را مورد بررسی قرار دهیم.

مراجع مربوطه

مراجع اصلی

[1] Ren, K., Zheng, T., Qin, Z. and Liu, X., 2020. Adversarial attacks and defenses in deep learning. Engineering, 6(3), pp.346-360.

[2] Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L. and Jain, A.K., 2020. Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17(2), pp.151-178.

[1] Adversarial attacks
[2] Image and text processing
[3] Open problems
[4] state-of-the-art

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[4] Ozdag, M., 2018. Adversarial attacks and defenses against deep neural networks: a survey. Procedia Computer Science, 140, pp.152-161.

[5] Akhtar, N. and Mian, A., 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access, 6, pp.14410-14430.

[6] Jin, W., Li, Y., Xu, H., Wang, Y., Ji, S., Aggarwal, C. and Tang, J., 2020. Adversarial Attacks and Defenses on Graphs: A Review, A Tool and Empirical Studies. arXiv preprint arXiv:2003.00653.

مراجع فرعی

[7] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D., 2018. Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.

[8] Li, Y., Jin, W., Xu, H. and Tang, J., 2020. Deeprobust: A pytorch library for adversarial attacks and defenses. arXiv preprint arXiv:2005.06149.

[9] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J., 2018. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9185-9193).

بخش دوم

اهداف پروژه

هدف آرمانی

ارائه راهکار به منظور بهبود دقت[5] در یکی از الگوریتم های حمله و یا دفاع

هدف کلی

شناخت انواع الگوریتم های حمله و دفاع در یادگیری عمیق به همراه روش های پیاده سازی آنها

اهداف ویژه

شناخت مبانی حملات خصمانه و دفاع در یادگیری عمیق، شناخت قدرتمندی و دقت الگوریتم ها، مروری بر مسائل باز مطرح شده و میزان پیشرفت در این حوزه

اهداف کاربردی

افزایش میزان دقت در پردازش تصویر، پردازش متن و بینایی ماشین[6] در کاربرد های پزشکی، امنیتی و غیره.

[5] Accuracy
[6] Computer vision

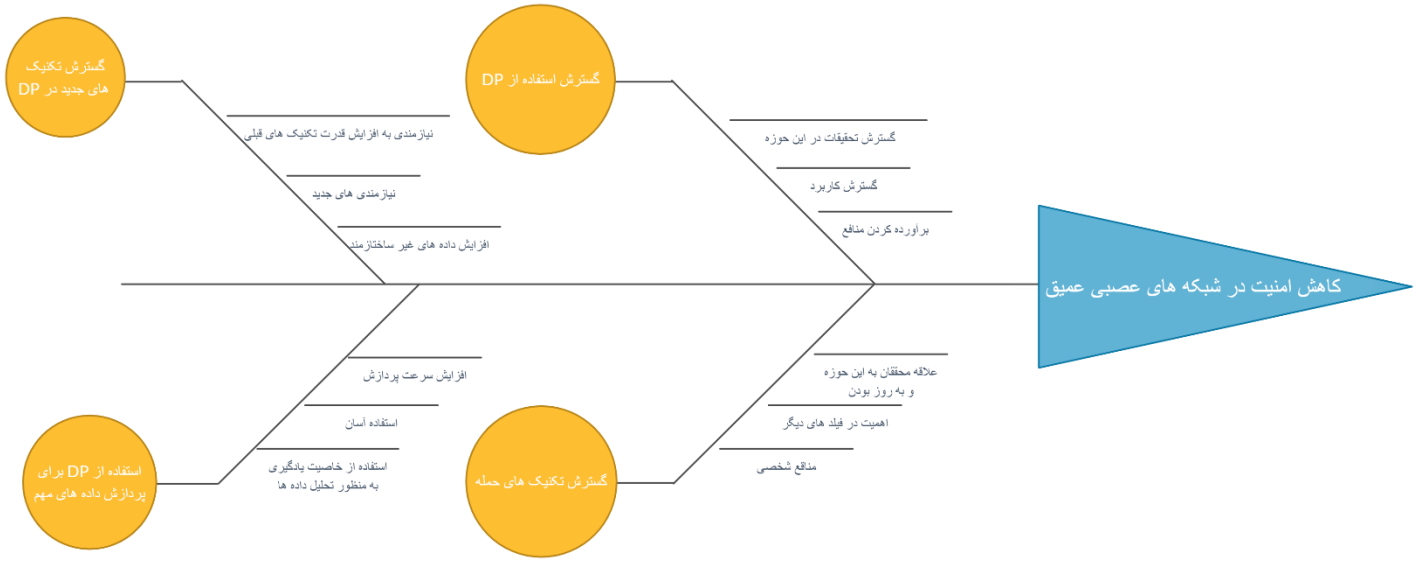
بخش سوم

تحقیق مورد نظر قصد برطرف کردن چه مشکلی را دارد؟

در این تحقیق قصد داریم تا به اهمیت و به روز بودن این مسئله بپردازیم و بیان کنیم که در صورت عدم پیشرفت در این حوزه، تمامی بخش ها و جنبه های هوش مصنوعی و فیلد هایی که در تعامل با آن هستند دچار مشکلات امنیتی[7] خواهند شد و کارایی خود را از دست خواهند داد.

پس مشکل اصلی که قصد برطرف کردن آنرا داریم، **کاهش امنیت** در مسائل مربوط به هوش مصنوعی و یادگیری عمیق می باشد.

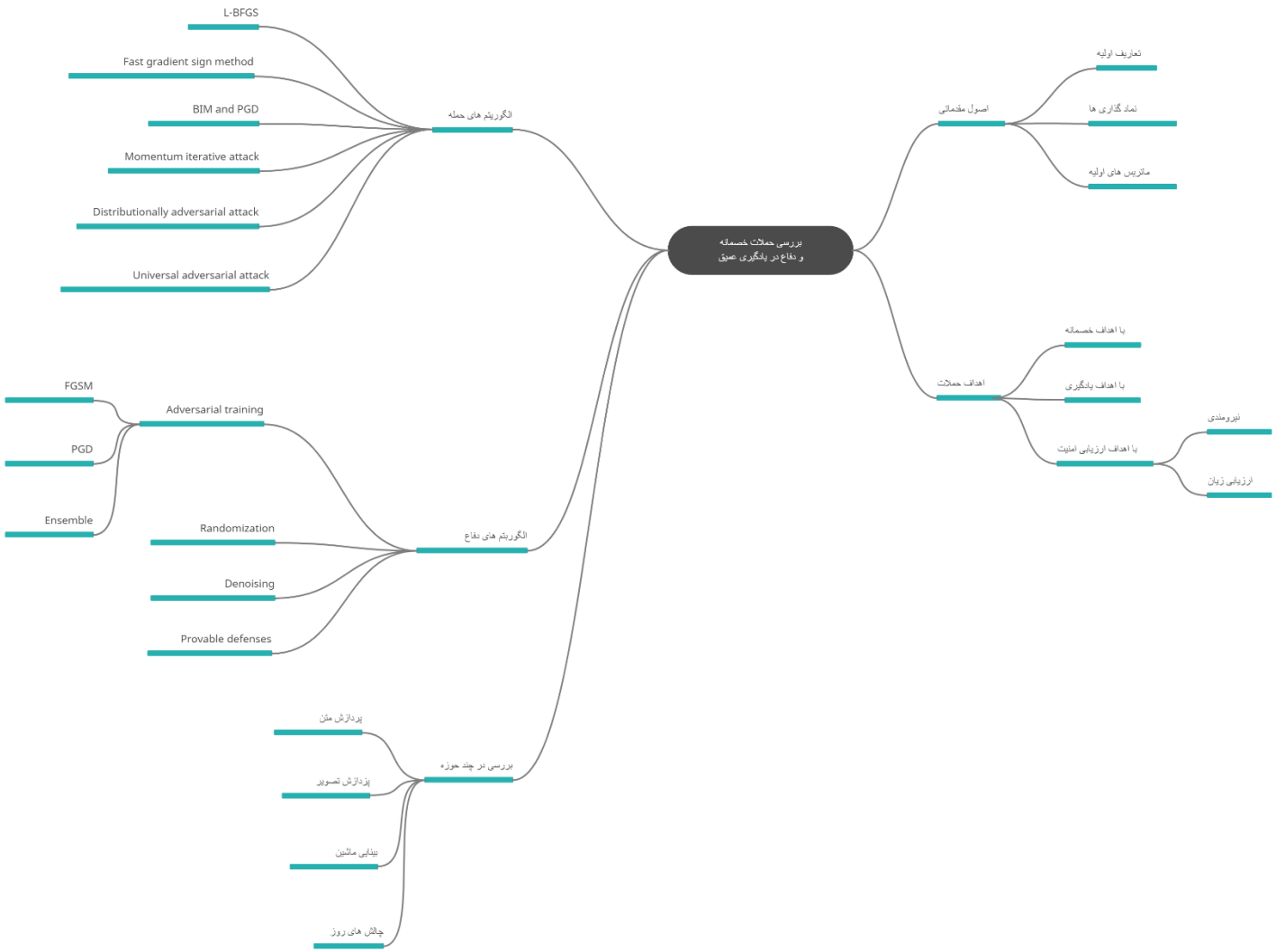
دیاگرام استخوان ماهی



[7] Security

بخش چهارم

نقشه ذهن



بخش پنجم

دلیل انتخاب موضوع

با توجه به علاقه به حوزه های هوش مصنوعی و امنیت، موضوعی که در اشتراک با این دو حوزه میباشد انتخاب شده است. به ویژه اینکه این موضوع از موضوعات تحقیقاتی روز[8] در حوزه هوش مصنوعی می باشد و به نوعی تمامی بخش های آنرا تحت تاثیر قرار میدهد.

[8] Hot research topic