ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Սամվելյան Միքայել Էդուարդի

Խորքային բազմագործակալ ամրապնդմամբ ուսուցման արդյունավետ մեթոդների մշակում և գնահատում

ՍԵՂՄԱԳԻՐ

Ե.13.04 «Հաշվողական մեքենաների, համալիրների, համակարգերի և ցանցերի մաթեմատիկական և ծրագրային ապահովում» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի համար

Երևան - 2020

ИНСТИТУТ ПРОБЛЕМ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ НАН РА

Самвелян Микаел Эдуардович

Разработка и оценка эффективных методов глубокого многоагентного обучения с подкреплением

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук по специальности 05.13.04 “Математическое и программное обеспечение вычислительных машин, комплексов, систем и сетей”

Ереван - 2020

Ատենախոսության թեման հաստատվել է Հայ-Ռուսական համալսարանում:

Գիտական ղեկավար՝	տեխ. գիտ. դոկտոր Գ.Հ. Խաչատրյան
Պաշտոնական ընդդիմախոսներ՝	ֆիզ.-մաթ. գիտ. դոկտոր Ս.Կ. Շուքուրյան
	ֆիզ.-մաթ. գիտ. թեկնածու Հ.Հ. Խաչատրյան
Առաջատար կազմակերպություն՝	Հայաստանի ազգային պոլիտեխնիկական համալսարան

Պաշտպանությունը կայանալու է 2020թ. ապրիլի 10-ին, ժ. 16:00-ին ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող 037 «Ինֆորմատիկա» մասնագիտական խորհրդի նիստում հետևյալ հասցեով՝ Երևան, 0014, Պ. Սևակի 1:

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ ԻԱՊԻ գրադարանում:

Սեղմագիրն առաքված է 2020թ. փետրվարի 29-ին:

Մասնագիտական խորհրդի գիտական քարտուղար, ֆիզ.-մաթ.գիտ. դոկտոր՝ Հ.Գ. Սարուխանյան

Тема диссертации утверждена в Российско-Армянском университете.

Научный руководитель:	доктор технических наук Г.Г. Хачатрян
Официальные оппоненты:	доктор физ.-мат. наук С.К. Шукурян
	кандидат физ.-мат. наук Г.А. Хачатрян
Ведущая организация:	Национальный политехнический университет Армении

Защита состоится 10-го апреля 2020г. в 16:00 часов на заседании специализированного совета 037 “Информатика” Института проблем информатики и автоматизации НАН РА по адресу: 0014, г. Ереван, ул. П. Севака 1.

С диссертацией можно ознакомиться в библиотеке ИПИА НАН РА.

Автореферат разослан 29-го февраля 2020г.

Ученый секретарь специализированного совета, доктор физ.-мат. наук А.Г. Саруханян

Աշխատանքի ընդհանուր նկարագիրը

Թեմայի արդիականությունը: Վերջին տարիների ընթացքում մեքենայական ուսուցումը (ՄՈւ) դարձել է գրեթե անփոխարինելի՝ լայն տարածում ստանալով արդյունաբերության բազմաթիվ բնագավառներում: ՄՈւ մեթոդներն ունակ են հայտնաբերելու տվյալներում առկա օրինաչափությունները և օգտագործել վերջիններս ինչպես ապագա տվյալները կանխատեսելու, այնպես էլ այլ որոշումներ կայացնելու համար: Այդ իսկ պատճառով ՄՈւ ալգորիթմները հաճախ օգտագործվում են այնպիսի ծրագրերում, որոնցում ցանկալի վարքագիծը հնարավոր չէ նախապես ծրագրավորել կամ օպտիմալ լուծումը պարզապես հայտնի չէ: Չնայած մի քանի տասնյակ տարիների պատմություն ունենալուն՝ ՄՈւ ոլորտի լայնատարած հաջողությունները սկսվել են միայն 2012 թվականից հետո:

2012 թ.-ին ամենամյա ImageNet[1] մրցույթի շրջանակներում Ջոֆֆրի Հինթոնի կողմից ղեկավարած հետազոտողների խումբը կարողացավ մարզել արհեստական նեյրոնային ցանցերի (ԱՆՑ) վրա հիմնված պատկերների դասակարգման ՄՈւ մոդել, որը ցույց տվեց բեկումնային արդյունք՝ էականորեն գերազանցելով բոլոր մրցակիցների արդյունքները: Չնայած այդպիսի ԱՆՑ-եր ու դրանց մարզման ալգորիթմները մշակվել էին դեռևս 90-ական թվականներին, մեծ հաջողություններ չէին արձանագրվել ուսուցանման օրինակների սղության և սահմանափակ հաշվողական ռեսուրսների պատճառով: AlexNet կոչվող մոդելը սկիզբ դրեց խորքային ուսուցման (ԽՈւ) հեղափոխությանը:

ԽՈւ մոդելներն օգտագործում են բազմաշերտ ԱՆՑ-եր, որոնք թույլ են տալիս սովորել տվյալների տարատեսակ ներկայացումներ[2]: Սա հնարավորություն է ստեղծում առավել արդյունավետորեն օրինաչափություններ հայտնաբերել մեծ չափեր ունեցող տվյալներում՝ առանց այդ տվյալների առանձնահատկությունները ձեռքով ի հայտ բերելու կամ խնդրի վերաբերյալ որևէ գիտելիք օգտագործելու: ԽՈւ-ն վերջին տարիներին հեղաշրջում կատարեց արհեստական բանականության (ԱԲ) բնագավառում՝ գրանցելով, թերևս, լավագույն արդյունքները պատկերների և բնական լեզվի մշակման բարդ խնդիրների դասում:

ԽՈւ-ի զարգացման արդյունքում առաջխաղացում ապրեց նաև ՄՈւ-ի ենթաոլորտներից մեկը՝ ամրապնդմամբ ուսուցումը (ԱՈւ): ԱՈւ-ի խնդիրն անդրադառնում է որոշումներ կայացնող գործակալին, որը շարունակական փոխազդեցության մեջ է գտնվում շրջապատող միջավայրի հետ[3]: Գործողություններ կատարելու արդյունքում գործակալը ստանում է թվային պարգևատրումներ, որոնց ընդհանուր գումարը առավելագույնի հասցնելը հանդիսանում է գործակալի նպատակը: Այստեղ շեշտը դրվում է շրջակա միջավայրի հետ անմիջական փոխազդեցության արդյունքում գործակալի վարքն ուսանելու վրա՝ առանց հիմնվելու որևիցե վերահսկողության վրա: ԱՈւ-ն, թերևս, ՄՈւ-ի բոլոր տարատեսակներից ամենամոտն է ուսուցանման այն ձևին, որն իրականացվում է մարդկանց և կենդանիների կողմից՝ ուսանում՝ ելնելով փորձից և սխալներից, հիմնվելով պարգևատրումների և պատժամիջոցների համակարգի վրա: Չնայած ԱՈւ-ի նախկինում արձանագրած որոշակի հաջողությունների՝ դրա կիրառելիությունը սահմանափակված է եղել փոքր քանակի հնարավոր վիճակներ ունեցող խնդիրների դասով:

Խորքային ամրապնդմամբ ուսուցումը (խորքային ԱՈւ) մեթոդների այն դասն է, որտեղ օգտագործվում են ԽՈւ մեթոդներ՝ ԱՈւ խնդիրներ լուծելու նպատակով: ԽՈւ-ի զարգացումը թույլ տվեց ԱՈւ մեթոդներին ուսուցանել առավել բարդ վարքագիծ և պայմանավորեց ԱՈւ մեթոդների խոշոր նվաճումները՝ հսկայական առաջխաղացում

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1097–1105. 2012.

[2] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[3] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction, volume 1. MIT Press, Cambridge, 1998.

արձանագրելով ԱԲ-ի մի շարք բարդ, չլուծված խնդիրների լուծման համար: Այդպիսի խնդիրների դասին են պատկանում համակարգչային «Աթարի» և բարդ ռազմավարական խաղերը, սեղանի գո և շախմատ խաղերը, սպիտակուցների ֆոլդինգը, ռոբոտների և անօդաչու թռչող սարքերի կառավարումը և այլն: Խորքային ԱՈւ-ի կիրառելիությունը հսկայական է իրական աշխարհի տարաբնույթ խնդիրներում, և այն կարող է լուծել ԱԲ-ում առկա բարդ մարտահրավերները:

Չնայած նրան, որ ԱՈւ-ի հիմնական հաջողությունները գրանցվել են մեկ գործակալից բաղկացած միջավայրերում, գոյություն ունեցող մի շարք համակարգեր բաղկացած են բազմաթիվ գործակալներից, որոնք հաճախ պետք է համագործակցեն՝ ընդհանուր նպատակին հասնելու համար: Օրինակ՝ անօդաչու թռչող սարքերի[4], ինքնագնաց մեքենաների[5] և բազմաբնույթ այլ ռոբոտային սարքերի համակարգումը բնական կերպով ներկայացվում է որպես համագործակցային բազմագործակալ ԱՈւ (ԲԱՈւ) խնդիրներ: Միագործակալ ԱՈւ-ի համար մշակված մեթոդները, սակայն, հաճախ պիտանի չեն բազմագործակալ համակարգերի խնդիրների լուծման համար հետևյալ դժվարությունների պատճառով.

- *Գործողությունների մեծ բազմություն.* գործակալների համատեղ գործողությունների բազմությունն աճում է էքսպոնենցիալ արագ գործակալների քանակի աճի հետ: Խնդիրը լուծելի դարձնելու համար հաճախ անհրաժեշտ է օգտագործել ապակենտրոնացված ռազմավարություն, երբ յուրաքանչյուր գործակալ համակարգում է իր վարքագիծը՝ հիմնվելով բացառապես սեփական փորձի վրա,
- *Ոչ ստացիոնարություն.* մարզման ընթացքում յուրաքանչյուր գործակալի ռազմավարությունը փոփոխվում է՝ առանձին գործակալների տեսանկյունից միջավայրը դարձնելով ոչ ստացիոնար,
- *Ապակենտրոնացում.* հաղորդակցման և դիտարկելիության սահմանափակումները կարող են ստիպել ուսանել ապակենտրոնացված ռազմավարություններ նույնիսկ այն դեպքում, երբ համատեղ գործողությունների բազմությունը շատ մեծ չէ,
- *Ներդրումների բաշխում.* երբ գործակալները ստանում են ընդհանուր պարգևատրում, անհատ գործակալները պետք է որոշեն իրենց ներդրումը թիմի հաջողության մեջ,
- *Միջավայրի ուսումնասիրություն.* օպտիմալ ռազմավարությունը հաճախ պահանջում է բոլոր գործակալներից կամ գործակալների որևէ ենթաբազմությունից համատեղ ուսումնասիրել նոր ռազմավարություն՝ համակարգված փորձարկելով նոր գործողությունների հաջորդականություններ:

Առանձնահատուկ հետաքրքրություն ներկայացնող խնդիրների դաս է մասնակի դիտարկելի, համագործակցային բազմագործակալ ուսուցումը[6], որտեղ անհրաժեշտ է ապակենտրոնացված ռազմավարությունների մշակում: Վերջիններս հաճախ կարելի է մարզել կենտրոնացված ձևով՝ մոդելավորված միջավայրերում կամ լաբորատոր պայմաններում: Սա կարող է ապահովել միջավայրի ընդհանուր վիճակի մասին տեղեկատվության հասանելիությունը մարզման ժամանակ, որն այլապես հասանելի չէ անհատ գործակալներին, ինչպես նաև հեռացնել հաղորդակցման սահմանափակումները գործակալների միջև: Կենտրոնացված մարզում և ապակենտրոնացված կատարում[7]

[4] J. Cui, Y. Liu, and A. Nallanathan. The application of multi-agent reinforcement learning in UAV networks. 2019 IEEE International Conference on Communications Workshops (ICC Workshops), pages 1–6, 2019.

[5] Y. Cao, W. Yu, W. Ren, and G. Chen. An overview of recent progress in the study of distributed multi-agent coordination. IEEE Transactions on Industrial informatics, 9 (1):427–438, 2013.

[6] J. K. Gupta, M. Egorov, and M. Kochenderfer. Cooperative Multi-agent Control Using Deep Reinforcement Learning. Autonomous Agents and Multiagent Systems, pages 66–83. Springer, 2017.

[7] Խնդրի ֆորմալ ձևակերպումը տրված է 1.4 ենթագլխում.

(ԿՄԱԿ) մոդելը գրավիչ ուղղություն է հետազոտության համար, քանի որ նմանատիպ խնդիրներն առկա են իրական աշխարհի տարաբնույթ համակարգերում:

Վերոհիշյալ հանգամանքները դժվարացնում են ԲԱՈւ մեթոդների կիրառումը հատկապես այն բազմագործակալ համակարգերում, որտեղ առկա են երկուսից ավել գործակալներ: Այնուամենայնիվ, ԲԱՈւ մեթոդների բարելավումը շատ կարևոր է արհեստական բանականությամբ օժտված այնպիսի համակարգեր կառուցելու համար, որոնք կարող են արդյունավետորեն համագործակցել միմյանց հետ՝ կիրառական նշանակություն ունեցող բազում առաջադրանքներ լուծելու համար:

Աշխատանքի նպատակը և դիտարկված խնդիրները: Աշխատանքի նպատակն է բարդ բազմագործակալ միջավայրերի համար արդյունավետ ԲԱՈւ մեթոդների մշակումը և գնահատումը: Սույն նպատակին հասնելու համար դիտարկենք հետևյալ խնդիրները.

1. Մշակել արդյունավետ ԲԱՈւ մեթոդ, որը թույլ է տալիս լիարժեքորեն օգտագործել կենտրոնացված մարզման հնարավորությունը անհատ գործակալների ապակենտրոնացված ռազմավարություններ մշակելու համար,
2. Ստեղծել հենանիշ (benchmark) միջավայր համագործակցային ԲԱՈւ մեթոդների համար, որը թույլ կտա գնահատել և համեմատել վերջիններիս արդյունավետությունները և մատնանշել դրանց թերությունները,
3. Մշակել ԲԱՈւ մեթոդ, որը թույլ կտա ապակենտրոնացված գործակալներին իրականացնել միջավայրի լայնածավալ համակարգված ուսումնասիրություն:

Հետազոտության օբյեկտները: Սույն աշխատանքում հետազոտության օբյեկտներ են հանդիսանում համագործակցային ԲԱՈւ ալգորիթմները, խորքային նեյրոնային ցանցերը, օպտիմիզացիայի ալգորիթմները, բազմագործակալ ուսուցման, կատարման, ուսումնասիրման և ապակենտրոնացման մոտեցումները: Հետազոտության օբյեկտներ են հանդիսանում նաև ԲԱՈւ ալգորիթմների ուսուցանման և գնահատման համար հենանիշները, ծրագրային միջոցները և մեթոդոլոգիաները:

Հետազոտության մեթոդները: Հետազոտությունն իրականացվել է միագործակալ և բազմագործակալ ԱՈւ, խորքային ԱՈւ, վերահսկվող ուսուցման և օպտիմիզացիայի մեթոդների օգնությամբ: Փորձնական արդյունքների ստացման համար կիրառվել են օբյեկտակողմնորոշված և զուգահեռ ծրագրավորման մեթոդներ:

Գիտական նորույթը: Սույն աշխատության շրջանակներում առաջ են քաշվել հետևյալ գիտական նորույթները.

- Մշակվել է նոր խորքային ԲԱՈւ մեթոդ, որը թույլ է տալիս ներկայացնել գործակալների համատեղ արժեքային ֆունկցիան[8] որպես անհատ գործակալների արժեքային ֆունկցիաների ոչ գծային համադրություն,
- Առաջարկվել է համատեղ արժեքային ֆունկցիայից ապակենտրոնացված ռազմավարության արդյունավետ առանձնացման մեթոդ, որն ապահովում է կայունությունը կենտրոնացված և ապակենտրոնացված ռազմավարությունների միջև,

[8]Որոշակի ռազմավարության համար արժեքային ֆունկցիան՝ $Q(s,u)$-ը, սահմանվում է որպես սպասված պարգևատրումների գումար s վիճակում u գործողությունը կատարելիս և այնուհետև ըստ նույն ռազմավարության գործելիս:

- Կառուցվել է համագործակցային ԲԱՈւ հենանիշ միջավայր, որն ապահովում է մասնակի դիտարկելիություն, բարդ դինամիկա և վիճակների մեծ բազմություն,
- Մշակվել է նոր խորքային ԲԱՈւ մեթոդ, որը թույլ է տալիս գործակալներին տևական ժամանակահատվածում իրականացնել միջավայրի համակարգված ուսումնասիրություն:

Ստացված արդյունքների գործնական կիրառությունը: ԲԱՈւ-ն ունի մեծ կիրառելիություն իրական աշխարհի տարաբնույթ խնդիրներում: Աշխատանքի արդյունքում մշակված ալգորիթմները կարող են կիրառվել բազում բազմագործակալ համակարգերում, որտեղ գործակալները պետք է համագործակցեն ընդհանուր խնդիրը լուծելու համար: Վերջիններս կարող են կիրառվել օրինակ՝ անօդաչու թռչող սարքերի, ինքնագնաց մեքենաների և արտադրական գործարաններում բազմաբնույթ ռոբոտային սարքերի համակարգման համար: Ավելին, մշակված մեթոդները կարող են օգտագործվել երթևեկության սարքերի ղեկավարման, տրանսպորտի կառավարման, բժշկական պատկերների և բաշխված զգայական ազդանշանների վերլուծության համար: Մշակված ծրագրային գրադարանը, որն օգտատերերին հասանելի է բաց կոդով, թույլ է տալիս հեշտորեն մշակել նոր ԲԱՈւ ալգորիթմներ կամ օգտագործել գոյություն ունեցող մեթոդները նոր բազմագործակալ խնդիրներ լուծելու համար:

Ներդրումներ: Ատենախոսության շրջանակներում ստացված արդյունքները ներդրվել են մի քանի համակարգերում.

- QMIX խորքային ԲԱՈւ ալգորիթմը, որը մշակված է ատենախոսության շրջանակներում, ներդրվել է Կալիֆոռնիայի Բերքլի համալսարանում մշակված Ray նախագծի մեջ: Ray-ը բաց, գերարագ ծրագրային միջավայր է, որը թույլ է տալիս իրականացնել լայնածավալ ՄՈւ և ԱՈւ մոդելների մարզում՝ բաշխված ծրագրավորման շնորհիվ[9]: RLlib-ը, որը հանդիսանում է Ray նախագծի գրադարանը ԱՈւ-ի մեթոդների համար, ներառում է մի քանի ժամանակակից ԱՈւ ալգորիթմներ, որոնց շարքում է նաև QMIX ալգորիթմը[10]: Ray-ը օգտագործվում է հազարավոր ՄՈւ-ի օգտատերերի կողմից[11],
- QMIX ալգորիթմը նաև ներդրված է PyMARL[12] համակարգում, որը Python լեզվի բաց գրադարան է ԲԱՈւ մեթոդներ մշակելու և օգտագործելու համար՝ մշակված Օքսֆորդի համասարանում,
- Աշխատանքի արդյունքում մշակված ալգորիթմներն ու ծրագրային համակարգը ներդրվել են Highlander Technology կազմակերպությունում, ԱՄՆ, և ներկայումս գտնվում են ակտիվ կիրառության մեջ[13]: Մասնավորապես, QMIX ալգորիթմն օգտագործվում է Highlander-ի ծանրոցների համակարգման ծրագրային պլատֆորմում, որը նախատեսված է տրանսպորտի կառավարման համակարգերի համար՝ ապրանքների շարժերը պլանավորելու, իրականացնելու և օպտիմիզացնելու համար:

Պաշտպանության ներկայացվող հիմնական դրույթները: Պաշտպանության են ներկայացվում հետևյալ հիմնական դրույթները.

[9]Ray-ի կայքէջն է. https://rise.cs.berkeley.edu/projects/ray. Կոդը հասանելի է. https://github.com/ray-project/ray

[10]RLlib-ի դոկումենտացիան հասանելի է. https://ray.readthedocs.io/en/latest/rllib-algorithms.html:

[11]Ray-ը Github շտեմարանում ունի մոտ 10 հազար աստղ և 1.4 հազար ճյուղավորում:

[12]PyMARL-ը հասանելի է. https://github.com/oxwhirl/pymarl

[13]Ներդրման ակտը ներկայացված է:

1. QMIX անունով նոր ԲԱՈւ մեթոդ, որը թույլ է տալիս մարզել ապակենտրոնացված ռազմավարություններ կենտրոնացված կերպով: QMIX-ն օգտագործում է նոր տիպի նեյրոնային ցանցի ճարտարապետություն, որը հնարավորություն է տալիս առավել արդյունավետորեն մաքսիմիզացնել գործակալների համատեղ Q-արժեքը և ապահովում է կայունությունը կենտրոնացված և ապակենտրոնացված ռազմավարությունների միջև,
2. Լայնածավալ փորձարարական արդյունքներ, որոնք հաստատում են QMIX-ի զգալի արդյունավետությունը գոյություն ունեցող ԲԱՈւ մեթոդների նկատմամբ: Հիմնավոր փորձարարական արդյունքներ, որոնք հաստատում են QMIX-ի կողմից միջավայրի վիճակի վերաբերյալ հավելյալ ինֆորմացիայի օգտագործման և անհատ գործակալների Q-արժեքների ոչ գծային համադրման կարևորությունը,
3. Նոր հենանիշ SMAC (StarCraft Multi-Agent Challenge) միջավայր համագործակցային ԲԱՈւ խնդիրների համար, որը հիմնված է հանրահայտ StarCraft II ռազմավարական խաղի վրա: SMAC հենանիշի օգնությամբ գոյություն ունեցող ԲԱՈւ մեթոդների լայնածավալ գնահատում ու համեմատում,
4. Նոր խորքային ԲԱՈւ ալգորիթմ, որը անհատ գործակալներին թույլ է տալիս տևական ժամանակահատվածում իրականացնել միջավայրի համակարգված ուսումնասիրություն: Ծավալուն փորձեր՝ մշակված ուսումնասիրության մեթոդի արդյունավետությունը գնահատելու և գոյություն ունեցող մեթոդների հետ համեմատելու համար:

Ստացված արդյունքների գրաքննությունը և փորձարկումը: Ստացված արդյունքները զեկուցվել են հայկական և միջազգային մի շարք գիտաժողովներում.

1. Восьмая годичная научная конференция РАУ, Ереван, Армения, 4-8 декабря 2017г.,
2. BMVA Symposium on Reinforcement Learning in Computer Vision, London, UK, May 9, 2018,
3. 35th International Conference on Machine Learning, Stockholm, Sweden, July 10-15, 2018,
4. Девятая годичная научная конференция РАУ, Ереван, Армения, 3-7 декабря 2018г.,
5. 18th International Conference on Autonomous Agents and MultiAgent Systems, Montreal, Canada, May 13-17, 2019,
6. 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, December 8-12, 2019,
7. 5th Deep Reinforcement Learning Workshop, Vancouver, Canada, December 13-14, 2019:

Աշխատանքի արդյունքները քննարկվել են Հայ-Ռուսական համալսարանում, Երևանի պետական համալսարանում և ՀՀ ԳԱԱ ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում անցկացված սեմինարների ընթացքում:

Հրապարակումները: Ատենախոսության թեմայի վերաբերյալ տպագրվել է 5 գիտական աշխատանք:

Աշխատանքի ծավալը և կառուցվածքը: Աշխատանքի ծավալը կազմում է 116 էջ: Աշխատանքը բաղկացած է ներածությունից, երեք գլուխներից, եզրակացությունից և գրականության ցանկից: Գրականության ցանկը ներառում է 142 աշխատություն: Աշխատանքը ներառում է 27 նկար և 8 աղյուսակ:

Աշխատանքի համառոտ նկարագրությունը

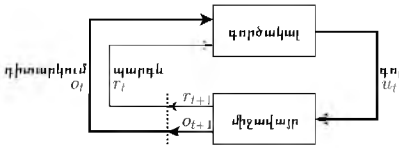
Ներածություն բաժնում հիմնավորվում է թեմայի արդիականությանը, ձևակերպվում են աշխատանքի նպատակը, գիտական նորույթը, կիրառական նշանակությունը, ինչպես նաև հակիրճ մատնանշվում է աշխատանքի էությունը:

Գլուխ 1-ը նվիրված է ուսումնասիրվող խնդիրների ընդհանուր քննարկմանը: Քանի որ խորքային ԲԱՈւ-ի մեթոդները հիմնվում են դասական ԱՈւ-ի, ԽՈւ-ի և խորքային ԱՈւ-ի մեթոդների վրա, հանգամանորեն ուսումնասիրվում են վերջիններիս տեսությունները, նկարագրվում են խնդիրները, դրվածքներն ու գոյություն ունեցող ալգորիթմները:

1.1 ենթագլուխը նկարագրում է ԱՈւ խնդիրն ու դասական ԱՈւ ալգորիթմները: ԱՈւ-ի նպատակն է սովորել՝ ինչպես է անհրաժեշտ իրավիճակներին համապատասխանեցնել այնպիսի գործողություններ, որ առավելագույնի հասցվի թվային պարգևատրումը[3]:

ԱՈւ-ի խնդիրը պարզագույն դեպքում ներառում է միայն մեկ ուսումնառող և որոշում կայացնող, որին անվանում են **գործակալ** (agent): Գործակալից դուրս ամեն ինչ կոչվում է **միջավայր** (environment): Ժամանակի ամեն դիսկրետ t քայլին գործակալը ստանում է միջավայրից o_t **դիտարկում** (observation) և կատարում է u_t **գործողությունը** (action): Ժամանակի հաջորդ քայլին գործակալը ստանում է թվային r_t **պարգևատրումը** (reward)՝ որպես կատարած գործողությունների արդյունք: Գործակալին չի հաղորդվում, թե որ գործողությունները պետք է կատարվեն: Փոխարենը, փորձելով տարբեր գործողություններ, նա պետք է ինքնուրույն բացահայտի, թե որոշակի իրավիճակներում որ գործողություններն են առավելագույն չափով պարգևատրվում: Նկար 1-ում ներկայացվում է գործակալի և միջավայրի միջև փոխազդեցությունը ԱՈւ-ում:

Լիարժեքորեն դիտարկելի միջավայրերում գործակալն ուղղակիորեն դիտարկում է միջավայրի ներկա s_t վիճակը՝ $o_t = s_t$: Մասնակիորեն դիտարկելի միջավայրերում գործակալի դիտարկումները պարունակում են սահմանափակ տեղեկություն միջավայրի s_t վիճակի մասին:



Նկար 1: Գործակալի և միջավայրի փոխազդեցությունը ԱՈւ-ում

Լիարժեքորեն դիտարկելի ԱՈւ-ի խնդիրը կարելի է ձևակերպել որպես Մարկովյան որոշումների գործընթաց (ՄՈԳ) (Markov Decision Process): ՄՈԳ-ը ներկայացվում է $\langle \mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ հնգյակի միջոցով, որտեղ $\mathcal{S}$-ը վիճակների վերջավոր բազմությունն է, իսկ $\mathcal{U}$-ն՝ գործողությունների վերջավոր բազմությունը: $\mathcal{P} : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to [0, 1]$ ներկայացնում է վիճակների անցման հավանականային բաշխումը, որտեղ $\mathcal{P}(s, u, s') = \mathbb{P}[s_{t+1} = s' | s_t = s, u_t = u]$, իսկ $\mathcal{R} : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$-ն՝ պարգևատրման ֆունկցիան, որտեղ $\mathcal{R}(s, u) = \mathbb{E}[r_{t+1} | s_t = s, u_t = u]$. $\gamma \in [0, 1]$-ն ծառայում է որպես պարգևատրման զեղչման գործակից:

Ժամանակի ամեն t քայլին գործակալի նպատակն է ընտրել այն գործողությունը, որն առավելագույնի կհասցնի սպասվող **զեղչված շահույթը**, որը տրվում է հետևյալ բանաձևով՝ $R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{n=0}^{\infty} \gamma^n r_{t+n+1}$: γ գործակիցը որոշում է ապագա պարգևատրումների առկա արժեքը:

Գործակալի վարքը նկարագրվում է π **ռազմավարությամբ** (policy), որն արտապատկերում է վիճակների բազմությունից դեպի գործակալի հնարավոր գործողություններն ընտրելու հավանականությունը: Ռազմավարությունը կարող է լինել դետերմինացված, որտեղ գործակալը s վիճակում մշտապես ընտրում է նույն գործողությունը՝ $u = \pi(s)$, կամ ստոխաստիկ, որտեղ գործողություններից յուրաքանչյուրին համախատասխանեցվում է որոշակի ընտրության հավանականություն՝

$\pi(u|s) = \mathbb{P}[u_t = u|s_t = s]$: Այսպիսով, ԱՈւ խնդրի նպատակն է մշակել մի ռազմավարություն, որն առավելագույնի է հասցնում սպասվող զեղչված շահույթը բոլոր $s \in \mathcal{S}$ վիճակների դեպքում:

π ռազմավարության **արժեքային ֆունկցիան** (value function) սահմանվում է որպես սպասված շահույթ s վիճակում u գործողությունը կատարելիս և այնուհետև ըստ ռազմավարության գործելիս՝ $Q_\pi(s,u) = \mathbb{E}_\pi[R_t|s_t = s, u_t = u]$: Օպտիմալ արժեքային ֆունկցիան սահմանվում է որպես $Q_*(s,u) = \max_\pi Q_\pi(s,u)$,: Ունենալով օպտիմալ Q-ֆունկցիան՝ կարող ենք մշակել օպտիմալ $\pi*$ ռազմավարությունը, որը միշտ ագահաբար է գործում Q_*-ի նկատմամբ՝ $\pi_*(s) = \text{argmax}_{u\in\mathcal{U}} Q_*(s,u)$: Ժամանակի ընթացքում առաջարկվել են բազում մեթոդներ ԱՈւ խնդիրը լուծելու համար՝ մոտարկելով Q-ֆունկցիան: ԱՈւ-ում առաջխաղացմանը նպաստած ալգորիթմներից է Q-ուսուցումը[14], որը բոլոր $s \in \mathcal{S}$ և $u \in \mathcal{U}$ զույգերի համար պահում է $Q(s,u)$-ի արժեքները համապատասխան աղյուսակում: Ուսուցանման ընթացքում, երբ գործակալն առկա ռազմավարությամբ գործողություններ է կատարում միջավայրում և ստանում դրանց համապատասխան պարգևատրումներն, աղյուսակի արժեքները թարմացվում են: Որոշակի պայմանների բավարարելու դեպքում Q-ուսուցումը զուգամիտում է օպտիմալ Q_*-ին:

1.2 ենթագլուխը ներկայացնում է ԽՈւ-ն վերահսկվող ուսուցման խնդիրների շրջանակներում: Ենթագլխում հանգամանորեն ուսումնասիրվում են ԱՆՑ-եր, ցանցերի տարատեսակ ճարտարապետություններ և մարզման ժամանակակից մեթոդներ:

1.3 ենթագլուխը միավորում է ԽՈւ-ն ԱՈւ խնդրի հետ և ուսումնասիրում խորքային ԱՈւ մեթոդների դասը, որոնք օգտագործում են խորքային ԱՆՑ-եր ԱՈւ խնդիրների լուծման համար: ԱՈւ դասական (աղյուսակային) մեթոդները նպատակահարմար չեն այն ՄՈԳ-երում, որտեղ վիճակների $\mathcal{S}$ բազմությունը հսկայական է: Օրինակ՝ նարդի խաղում վիճակների բազմության հզորությունը 10^{20} կարգի է, գո խաղում՝ 10^{170}, իսկ «Աթարի» համակարգչային խաղերում՝ $256^{84\times84\times4}$ [15]: Հնարավոր լուծումներից մեկն է Q-ֆունկցիայի մոտարկումն ԱՆՑ-երի միջոցով:

Խորքային ԱՈւ-ի առաջին հաջողված մոդելը խորքային Q-ցանցերն են (DQN), որոնցում Q-ֆունկցիան ներկայացված է որպես ԱՆՑ՝ $Q(s,u;\theta)$: Մարզման ժամանակ գործակալն ընտրում է գործողությունները՝ ելնելով ուսումնասիրության (exploration) որոշակի մեթոդից, ինչպիսին է ϵ-ագահ ուսումնասիրությունը, երբ $1-\epsilon$ հավանականությամբ ընտրվում է օպտիմալ գործողությունը՝ $\text{argmax}_{u\in\mathcal{U}} Q(s,u;\theta)$, իսկ ϵ հավանականությամբ ընտրվում է պատահական գործողություն:

1.4 ենթագլուխն ուսումնասիրում է ԲԱՈւ խնդիրը և գոյություն ունեցող ԲԱՈւ ալգորիթմները: Համագործակցային ԲԱՈւ խնդիրը կարելի է ձևակերպել որպես ապակենտրոնացված մասնակի դիտարկելի ՄՈԳ (decentralised partially observable MDP)[16]՝ կազմված հետևյալ ությակից՝ $\langle S, U, P, R, Z, O, n, \gamma\rangle$: $s \in S$ նկարագրում է միջավայրի վիճակը: Ժամանակի ամեն t քայլին յուրաքանչյուր $a \in A \equiv \{1,...,n\}$ գործակալ ընտրում է $u^a \in U$ գործողություն՝ կազմելով գործակալների համատեղ $\mathbf{u} \in \mathbf{U} \equiv U^n$ գործողությունը: Գործողությունը կատարելիս միջավայրի վիճակը փոփոխվում է՝ ելնելով վիճակների անցման հավանականային բաշխումից՝ $P(s'|s,\mathbf{u}) : S\times\mathbf{U}\times S \to [0,1]$: Բոլոր գործակալները կիսում են ընդհանուր պարգևատրումը՝ ըստ $R(s,\mathbf{u}) : S\times\mathbf{U}\to\mathbb{R}$ ֆունկցիայի: $\gamma \in [0,1]$-ը պարգևատրման զեղչման գործակիցն է:

Դիտարկենք մասնակի դիտարկելի միջավայրերը, որտեղ յուրաքանչյուր գործակալ ստանում է $z \in Z$ դիտարկում՝ ելնելով դիտարկումների $O(s,a) : S\times A\to Z$ ֆունկցիայից: Յուրաքանչյուր գործակալ ունի գործողություն-դիտարկում զույգերի պատմություն $\tau^a \in$

[14] C. Watkins and P. Dayan. Q-learning. Machine Learning, pages 279–292, 1992.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.

[16] F. A. Oliehoek and C. Amato. A Concise Introduction to Decentralized POMDPs. SpringerBriefs in Intelligent Systems. Springer, 2016.

$T \equiv (Z \times U)^*$, ըստ որի կարելի է մշակել $\pi^a(u^a|\tau^a) : T \times U \to [0,1]$ ռազմավարություն: Համատեղ π ռազմավարության համար կարող ենք սահմանել համատեղ արժեքային ֆունկցիա՝ $Q_\pi(s_t, \mathbf{u}_t) = \mathbb{E}_\pi\left[R_t|s_t, \mathbf{u}_t\right]$, որտեղ $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$:

ԿՄԱԿ պայմաններում միջավայրի s վիճակը, ինչպես նաև բոլոր գործակալների դիտարկումներն ու գործողությունները հասանելի են կենտրոնացված մարզման ընթացքում: Գործակալների ապակենտրոնացված ռազմավարությունները մարզումից հետո կարող են հիմնվել միայն սեփական դիտարկումների և գործողությունների վրա:

Այսպիսի բազմագործակալ խնդրի համար գործակալների գործողությունների արդյունքները պատշաճ կերպով ներկայացնելը պահանջում է կենտրոնացված արժեքային ֆունկցիա՝ Q_{tot}, որը հիմնվում է միջավայրի $s \in S$ վիճակի և գործակալների համատեղ գործողության $\mathbf{u} \in \mathbf{U}$ վրա: Սակայն նման ֆունկցիա դժվար է սովորել, երբ առկա են մեծ թվով գործակալներ, քանի որ $\mathbf{U}$ բազմության հզորությունը էքսպոնենցիալ աճում է գործակալների n թվի աճի հետ: Նույնիսկ եթե այն կարելի է սովորել, պարզ չէ՝ ինչպես կարելի է դրանից առանձնացնել ապակենտրոնացված ռազմավարություն յուրաքանչյուր գործակալի համար:

Պարզագույն լուծումն է հրաժարվել կենտրոնացված արժեքային ֆունկցիայից, և թույլ տալ յուրաքանչյուր a գործակալին սովորել անհատական արժեքային Q_a ֆունկցիա անկախ կերպով, ինչպես անկախ Q-ուսուցում մեթոդում (IQL)[17]: Այս մոտեցումն, այնուամենայնիվ, անտեսում է ԲԱՈւ խնդիրներում գործակալների փոփոխվող ռազմավարությունների արդյունքում առաջացած ոչ ստացիոնարությունը:

Հնարավոր է նաև սովորել լիովին կենտրոնացված արժեքային Q_{tot} ֆունկցիա, որը հիմնված է համատեղ $\mathbf{u}$ գործողության վրա, և օգտագործել այն՝ ապակենտրոնացված ռազմավարությունների առանձնացման համար, ինչպես օրինակ COMA մեթոդում [18]: Սակայն այդպիսի արժեքային Q_{tot} ֆունկցիայի մարզումը անիրագործելի է, եթե գործակալների թիվը շատ մեծ է:

VDN մեթոդը[19] փորձում է սովորել համատեղ արժեքային $Q_{tot}(\boldsymbol{\tau}, \mathbf{u})$ ֆունկցիա, որտեղ $\boldsymbol{\tau} \in \mathbf{T} \equiv \mathcal{T}^n$-ն համատեղ դիտարկում-գործողություն զույգերի պատմությունն է: VDN-ը ներկայացնում է Q_{tot}-ը որպես առանձին գործակալների արժեքային $Q_a(\tau^a, u^a; \theta^a)$ ֆունկցիաների գումար՝ $Q_{tot}(\boldsymbol{\tau}, \mathbf{u}) = \sum_{i=1}^{n} Q_i(\tau^i, u^i; \theta^i)$, որոնք հիմնված են միայն սեփական դիտարկումների և գործողությունների պատմության վրա: VDN-ը թույլ է տալիս ներկայացնել միայն ֆունկցիաներ, որոնք ներկայացվում են որպես առանձին Q_a ֆունկցիաների գծային համադրություն: Այնուամենայնիվ, խնդրի բարդ էության և մեծ թվով գործակալների առկայության պատճառով բազում իրական միջավայրեր պահանջում են որպեսզի գործակալները մշակեն Q_{tot}-ի առավել բարդ, ոչ գծային ֆակտորացում: Բացի այդ, VDN-ն անտեսում է կենտրոնացված ուսուցման ժամանակ հասանելի $s \in S$ վիճակի վերաբերյալ տեղեկատվությունը, որը կարող է նպաստել առավել ճշգրիտ Q_{tot}-ի ուսուցանմանը:

Գլուխ 2-ում ԲԱՈւ խնդիրների լուծման համար առաջարկվում է QMIX մեթոդը, որը լրացնում է գոյություն ունեցող ԲԱՈւ ալգորիթմներում առկա մի շարք բացեր: Սույն գլխում տրվում է QMIX-ի մանրակրկիտ նկարագիրը, և քննարկվում է վերջինիս առավելությունը գոյություն ունեցող մեթոդների նկատմամբ: Գլուխը ներառում է լայնածավալ փորձարարական արդյունքներ, որոնք հաստատում են QMIX-ի զգալի արդյունավետությունը գոյություն ունեցող ԲԱՈւ մեթոդների նկատմամբ:

[17] M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. Proceedings of the Tenth International Conference on Machine Learning, pages 330–337, 1993.

[18] J. N. Foerster, G. Farquhar, T. Afouras, et al. Counterfactual multi-agent policy gradients. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[19] P. Sunehag, G. Lever, A. Gruslys, et al. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. International Conference on Autonomous Agents and Multiagent Systems, pages 2085-2087, 2017.

2.1 ենթագլխում նկարագրված է ԲԱՈւ-ում առկա մարտահրավերներից մեկը՝ ինչպես ներկայացնել և օգտագործել գործողությունների արժեքային ֆունկցիան, որը կիրառվում է ԱՈւ ալգորիթմների մեծ մասի կողմից: Գործակալների գործողությունների արդյունքները պատշաճ կերպով ներկայացնելը պահանջում է կենտրոնացված արժեքային ֆունկցիա՝ Q_{tot}, որը հիմնվում է միջավայրի վիճակի և գործակալների համատեղ գործողության վրա: Նման ֆունկցիան դժվար է սովորել մեծ թվով գործակալների առկայության պարագայում, և նույնիսկ այն սովորելու դեպքում պարզ չէ՝ ինչպես կարելի է յուրաքանչյուր գործակալի համար առանձնացնել ապակենտրոնացված ռազմավարություն: Ուսումնասիրության կարևոր խնդիր է հասկանալը, թե արդյոք կենտրոնացված մարզումը թույլ կտա սովորել համատեղ Q_{tot}, որը կարող է ֆակտորացվել ըստ գործակալի Q_a ֆունկցիաների այնպես, որ ապահովվի կայունությունը կենտրոնացված և ապակենտրոնացված ռազմավարությունների միջև:

2.2 ենթագլուխն անդրադառնում է մի շարք ալգորիթմների, որոնք փորձել են լուծել համագործակցային ԲԱՈւ խնդիրը: Այստեղ ներկայացվում են վերջիններիս սահմանափակումները, և մատնանշվում են դրանց և QMIX մեթոդի տարբերությունները:

2.3 ենթագլխում առաջարկվում է QMIX ալգորիթմը ԲԱՈւ խնդիրների լուծման համար: Ալգորիթմի հիմքում ընկած է այն դատողությունը, որ VDN մեթոդում Q_{tot}-ի խիստ ֆակտորացումը անհրաժեշտ չէ, որպեսզի մշակվեն ապակենտրոնացված ռազմավարություններ, որոնք համահունչ են համատեղ կենտրոնացված ռազմավարութան հետ: Դրա համար անհրաժեշտ է միայն հետևյալ պայմանի բավարարումը՝

$$\underset{\mathbf{u}}{\operatorname{argmax}}\, Q_{tot}(\boldsymbol{\tau}, \mathbf{u}) = \begin{pmatrix} \operatorname{argmax}_{u^1} Q_1(\tau^1, u^1) \\ \vdots \\ \operatorname{argmax}_{u^n} Q_n(\tau^n, u^n) \end{pmatrix}. \tag{1}$$

(1)-ը թույլ է տալիս արդյունավետորեն ընտրել օպտիմալ համատեղ $\mathbf{u}$ գործողությունը, որը համընկնում է յուրաքանչյուր a գործակալի կողմից ըստ սեփական Q_a ֆունկցիայի օպտիմալ u^a գործողության ընտրության հետ:

VDN-ի ներկայացումը բավարարում է (1)-ին: Այուամենայնիվ, QMIX-ն ունակ է ներկայացնել ֆունկցիաների շատ ավելի հարուստ ընտանիք, որոնք ևս բավարարում են (1)-ին: Մոնոտոնությունը կարող է պարտադրվել՝ Q_{tot}-ի և յուրաքանչյուր Q_a միջև՝ հետևյալ պայմանը դնելով.

$$\frac{\partial Q_{tot}}{\partial Q_a} \geq 0, \ \forall a \in A: \tag{2}$$

(2) պայմանին բավարարելու համար QMIX-ն օգտագործում է Q_{tot}-ի մի ճարտարապետություն, որն ընդգրկում է հետևյալ երեք բաղադրիչները՝ *գործակալների ցանցեր* (agent networks), *խառնիչ ցանց* (mixing network) և հիպերցանցերի[20] (hypernetwork) բազմություն: Նկար 2-ում պատկերված են QMIX-ի բաղադրիչ մասերը:

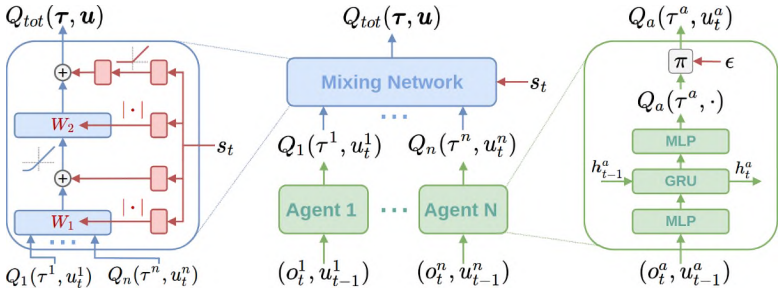
Յուրաքանչյուր a գործակալի համար գոյություն ունի մեկ գործակալի ցանց, որը ներկայացնում է նրա արժեքային $Q_a(\tau^a, u^a)$ ֆունկցիան: Այն ներկայացված է որպես DRQN ցանց[21], որը հիմնված է ռեկուրենտ ՆՑ-ի վրա[22] և ժամանակի ամեն քայլին որպես մուտք ստանում է անձնական o_t^a դիտարկումն ու գործակալի վերջին u_{t-1}^a գործողությունը (տես Նկար 2գ):

Խառնիչ ցանցն իրենից ներկայացնում է ուղիղ տարածման (feedforward) ԱՆՑ, որը մուտքում ստանում է գործակալների ցանցերի ելքերը, և «խառնում» է վերջիններիս

[20] D. Ha, A. Dai, and Q. V. Le. HyperNetworks. International Conference on Learning Representations, 2017.

[21] M. Hausknecht and P. Stone. Deep Recurrent Q-Learning for Partially Observable MDPs. AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents, 2015.

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.



Նկար 2: (ա) Խառնիչ ցանցի կառուցվածքը. կարմիր գույնով պատկերված են հիպերցանցերը, որոնք արտածում են խառնիչ ցանցի պարամետրերը (վերջիններս պատկերված են կապույտ գույնով): (բ) QMIX-ի ընդհանուր ճարտարապետություն: (գ) Գործակալների ցանցերի կառուցվածքը:

մոնոտոն կերպով՝ արտածելով Q_{tot}-ի արժեքը (Նկար 2ա): Որպեսզի բավարարվի մոնոտոնության (2) պայմանը, խառնիչ ցանցի կշիռները պետք է լինեն ոչ բացասական:

Խառնիչ ցանցի կշիռներն արժեքավորվում են հիպերցանցերի միջոցով: Յուրաքանչյուր հիպերցանց մուտքում ստանում է միջավայրի s վիճակը և արտածում խառնիչ ցանցի մեկ շերտի կշիռները: Բոլոր հիպերցանցերը կազմված են մեկական գծային շերտերից, որոնց հաջորդում է բացարձակ արժեք ֆունկցիան՝ կշիռների ոչ բացասականությունը ապահովելու համար: Նկար 2ա-ում պատկերված են խառնիչ ցանցն ու հիպերցանցերը:

QMIX-ը մարզվում է մինիմիզացնելով հետևյալ կորստի ֆունկցիան՝ $\mathcal{L}(\theta) = \mathbb{E}\left[\left(r + \gamma \max_{\mathbf{u}'} Q_{tot}(\boldsymbol{\tau}', \mathbf{u}', s'; \theta^-) - Q_{tot}(\boldsymbol{\tau}, \mathbf{u}, s; \theta)\right)^2\right]$, որտեղ θ^--ն Q_{tot} ցանցի θ պարամետրերի բազմության պատճենն է, որը «սառեցվում» է ուսուցանման ժամանակ և թարմացվում միայն որոշակի քանակի իտերացիաներից հետո: QMIX-ի ճարտարապետությունը թույլ է տալիս մարզել բոլոր կոմպոնենտները (խառնիչ, գործակալների և հիպերցանցերը)՝ մինիմիզացնելով միայն այս կորստի ֆունկցիան և օգտվելով սխալների հետադարձ տարածման (backpropagation) մեթոդից [23]: (1) պայմանի բավարարվածությունը թույլ է տալիս մաքսիմիզացնել Q_{tot}-ը գծային ժամանակում՝ ըստ գործակալների n թվի:

2.4 ենթագլխում գնահատվում են QMIX, VDN և այլ մեթոդներով ներկայացվող արժեքային ֆունկցիաների ունիվերսալությունը պարզ մատրիցային խաղի շնորհիվ, որը պարունակում է երկու գործակալ: Փորձնական ճանապարհով հաստատվում է, որ QMIX մեթոդը կարողանում է սովորել առավել ճշգրիտ Q-ֆունկցիայի ներկայացում, որի արդյունքում մշակվում է խնդրի օպտիմալ ռազմավարությունը: VDN-ին և մյուս մեթոդներին չի հաջողվում ուսուցանել ճշգրիտ Q-ֆունկցիա, որի արդյունքում մշակվում է ոչ օպտիմալ բազմագործակալ ռազմավարություն:

2.5 ենթագլխում նկարագրվում է StarCraft II միջավայրը, որի շնորհիվ գնահատվում է QMIX մեթոդը և համեմատվում VDN և IQL ալգորիթմների հետ: StarCraft II-ը հանրահայտ ռազմավարական (strategy) խաղ է, որտեղ խաղացողները զինվորներից բաղկացած բանակներ են կազմում և պատերազմում միմյանց դեմ: StarCraft II-ում կենտրոնանում ենք *ապակենտրոնացված միկրոկառավարման* խնդրի վրա, որտեղ յուրաքանչյուր գործակալ ղեկավարում է բանակի մեկ զինվոր: Սույն գլխում դիտարկում ենք սցենարներ, որտեղ քարտեզի հակառակ կողմում գտնվում են երկու բանակներ, որոնցից մեկը ղեկավարվում

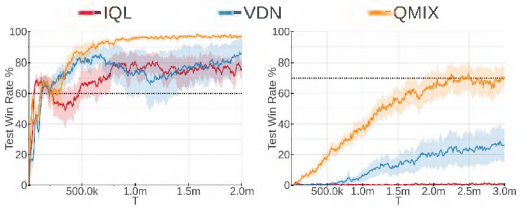
[23] D. Rumelhart, G. Hinton, and R. Williams, Learning representations by back-propagating errors. Nature, 323(6088):533–536, 1986.

է ԱՈւ գործակալների, իսկ մյուսը՝ համակարգչային ծրագրի կողմից, որը դրսևորում է ծրագրավորված մարտավարություն: Փարձարկումների ընթացքում օգտագործվում են 3-ից 9 զինվոր պարունակող 6 սցենար: Նկար 3-ում պատկերված է *5m* սցենարը: Միջավայրն առավել հանգամանորեն նկարագրվում է հաջորդ գլխում:

Նկար 3: 5m սցենար:

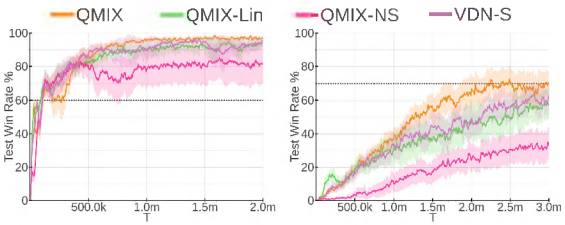
2.6 ենթագլխում ներկայացվում են QMIX, VDN և IQL մեթոդների ցուցադրված արդյունքները՝ StarCraft II միջավայրում: Որպեսզի գնահատվի յուրաքանչյուր մեթոդի արդյունավետությունը, կատարվում են հետևյալ քայլերը. յուրաքանչյուր մեթոդի յուրաքանչյուր ուսուցանման ժամանակ իրականացվում է մեթոդի մարզում 100 մարտ երկարությամբ, որին հաջորդում է 20 փորձնական մարտ (որոնց ընթացքում կիրառվում են ագահ մարտավարություններ): Մարտերի այն տոկոսը, որում մեթոդին հաջողվում է հաղթահարել թշնամուն, կանվանենք հաղթանակի գործակից (win rate):

Նկար 4-ում պատկերված են ցուցաբերված միջին հաղթանակի գործակիցները QMIX, VDN և IQL մեթոդների ուսուցանման ժամանակ 2 սցենարների համար (95% վստահության միջակայքի հետ միասին): Յուրաքանչյուր մեթոդ յուրաքանչյուր սցենարում փորձարկվել է 20 անգամ:



Նկար 4: IQL, VDN և QMIX մեթոդների հաղթանակի գործակիցները:

Բոլոր 6 սցենարներում QMIX-ին հաջողվում է ցույց տալ լավագույն արդյունքը: Այն ավելի արագ և հաջող է կարողանում սովորել հաղթել հակառակորդին քան VDN-ը: IQL մեթոդի ցույց տրված արդյունքները վատագույնն են, քանի որ այս մեթոդին չի հաջողվում հաղթահարել ԲԱՈւ-ում առկա միջավայրի ոչ ստացիոնարության խնդիրը:



Նկար 5: QMIX, QMIX-Lin, QMIX-NS և VDN-S մեթոդների հաղթանակի գործակիցները:

Կատարված են նաև որոշ փորձեր, որոնք գնահատում են կենտրոնացված ուսուցանման ընթացքում միջավայրի վիճակի մասին հավելյալ ինֆորմացիայի օգտագործումը, ինչպես նաև QMIX-ում անհատ գործակալների Q-արժեքների ոչ գծային համադրման կարևորությունը: Արդյունքները տրված են Նկար 5-ում, որտեղ QMIX-NS-ը (QMIX no state) QMIX մեթոդի առանց միջավայրի s վիճակի օգտագործման մեթոդն է, իսկ QMIX-Lin-ը օգտագործում է Q-ֆունկցաների գծային համադրություն (ոչ գծային համադրության փոխարեն): Բացի այդ, ուսումնասիրված է VDN-S մեթոդը, որը VDN ալգորիթմի այն ձևափոխումն է, որն օգտագործում է միջավայրի s վիճակի մասին ինֆորմացիան: Ստացված արդյունքներից կարելի է եզրակացնել, որ և՛ s-ի օգտագործումը, և՛ ոչ գծային համադրությունը անհրաժեշտ են առավել լավ ռազմավարություն մշակելու համար:

2.7 ենթագլխում ամփոփված են գլուխ 2-ում ստացված արդյունքներն ու հակիրճ

շարադրված են հետագա հետազոտությունների նպատակները:

Գլուխ 3-ում առաջարկվում է նոր հենանիշ միջավայր ԲԱՈւ խնդիրների լուծման համար, որի անունն է StarCraft-ի բազմագործակալ մարտահրավեր (StarCraft Multi-Agent Challenge) կամ պարզապես SMAC, որը հիմնված է հանրահայտ StarCraft II խաղի վրա:

3.1 ենթագլուխը հիմնավորում է ԲԱՈւ խնդիրների համար նոր հենանիշ միջավայրի անհրաժեշտությունը: ԲԱՈւ ոլորտի առաջխաղացմանը խոչընդոտող առանցքային խնդիր է ուսուցանման և գնահատման համար ստանդարտացված հենանիշ (benchmark) միջավայրի բացակայությունը: Չնայած վերջին տարիներին համագործակցային ԲԱՈւ խնդրի լուծման համար մեծ թվով աշխատությունների հրապարակմանը՝ հեղինակների կողմից, որպես կանոն, առաջարկվում են մեկանգամյա միջավայրեր, որոնք չափազանց պարզ են կամ հարմարեցված են առաջարկվող ալգորիթմներին: Միագործակալ ԱՈւ-ում ALE[24] և MuJoCo[25] ստանդարտ միջավայրերը մեծ առաջընթացի հիմք են հանդիսացել: Այնուամենայնիվ, խորքային ԲԱՈւ-ի համար արկա չեն նմանատիպ միջավայրեր:

Հիմք ընդունելով վերոգրյալը՝ մեծ անհրաժեշտություն է առաջանում ստեղծելու դժվարին հենանիշ, որը կարող է օգտագործվել ԲԱՈւ ալգորիթմներ մշակելու, մարզելու, համեմատելու և առաջընթացը գնահատելու համար:

3.2 ենթագլխում ներկայացված է հակիրճ ակնարկ գոյություն ունեցող այն միջավայրերի մասին, որոնք օգտագործվել են ԲԱՈւ ալգորիթմներ մշակելու համար: Ենթագլխում մանրամասն ներկայացված են դրանց թերություններն ու սահմանափակումները:

3.3 ենթագլխում հակիրճ ներկայված է աշխատանքի համար մշակված ծրագրային պլատֆորմը, որը բաց հասանելի է Github շտեմարանում[26]: Այն պարունակում է QMIX, VDN, COMA և IQL մեթոդների իրականացումները: Պլատֆորմը մշակված է Python լեզվով և օգտագործում է հանրահայտ PyTorch գրադարանը նեյրոնային ցանցերի հետ աշխատանքի համար: Պլատֆորմն օգտագործում է օբյեկտակողմնորոշված ծրագրավորման հիմունքները: Վերջինիս բաղադրիչներն ունեն մոդուլար կառուցվածք, ինչը թույլ է տալիս հեշտորեն օգտագործել ԲԱՈւ ալգիրիթմների տարատեսակ կարգավորումներ, ինչպես նաև հեշտորեն մշակել նոր ալգորիթմներ: Պլատֆորմը թույլ է տալիս ծրագրի աշխատանքը կատարել NVIDIA գրաֆիկական քարտերի վրա՝ օգտվելով CUDA տեխնոլոգիայից, ինչն ապահովում է մեթոդների գերարագ մարզումը համապատասխան սարքավորումների առկայության դեպքում:

3.4 ենթագլխում մանրամասն նկարագրվում է SMAC հենանիշը: SMAC-ը հիմնված է StarCraft II ռազմավարական խաղի միկրոկառավարման վրա, որը խաղի կարևագույն բաղադրիչներից է և անդրադառնում է անհատ զինվորների ղեկավարմանը: Մշակված հենանիշ միջավայրին բնորոշ են մի շարք հատկություններ, որոնք առկա են բազում իրական բազմագործակալ միջավայրերում.

- Միջավայրի մասնակի դիտարկելիություն. ինչպես մի շարք իրական խնդիրներում գործակալները միայն սահմանափակ տեղեկություն են ստանում միջավայրի մասին (օրինակ՝ անօդաչու թռչող սարքերի և ինքնագնաց մեքենաների ղեկավարում),
- Գործակալների մեծ բազմություն. առկա է մինչև 27 գործակալ պարունակող 14 սցենար,
- Երկարաժամկետ ռազմավարություններ մշակելու անհրաժեշտություն. հաղթանակ տանելու համար պահանջվում է մի շարք հաջորդական մարտավարական

[24]M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279, 2013.

[25]E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026–5033, 2012.

[26]https://github.com/oxwhirl/pymarl

որոշումների կայացում,

- Վիճակների մեծ բազմություն. տարատեսակ զինվորների առկայությունն ու քարտեզների մեծ չափերը նպաստում են հնարավոր վիճակների հսկայական բազմության ստեղծմանը,
- Գործողությունների մեծ բազմություն. ժամանակի ամեն քայլին գործակալներից յուրաքանչյուրը կարող է իրականացնել մինչև 70 գործողություն,
- Համակարգված թիմային աշխատանքի անհրաժեշտություն. զինվորների ղեկավարումը պահանջում է անհատ գործակալներից դրսևորել խիստ համակարգված ռազմավարություն, ինչը չափազանց կարևոր է ԲԱՈւ խնդիրների համար,
- Ստոխաստիկություն. հակառակորդի իրականացրած ռազմավարությունները փոփոխվում են, իսկ զինվորների փամփուշների վերալիցքավորման ժամանակահատվածը պատահական մեծություն է:

SMAC-ն առաջին ԲԱՈւ հենանիշ միջավայրն է, որն օժտված է վերոհիշյալ բոլոր հատկություններով և թույլ է տալիս առավել լիակատար գնահատել ԲԱՈւ մեթոդների արդյունավետությունը:

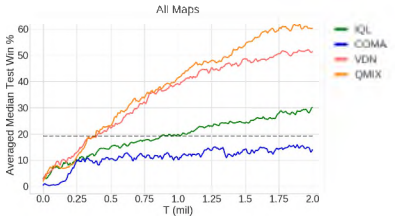
Նկար 6: SMAC հենանիշի որոշ սցենարներ:

SMAC-ի սցենարները հատուկ մշակված են թիմային աշխատանքի մի շարք հնարքների ի հայտ գալը ստուգելու համար: Միկրոկառավարման կարևորագույն հնարք է, օրինակ՝ հավաքական կրակ արձակելը (focus fire), որի ժամանակ բոլոր զինվորները համակարգված կրակ են արձակում նույն հակառակորդի վրա, որի մահից հետո համակարգված կրակ են արձակում հաջորդ հակառակորդի վրա և այդպես շարունակ: Մեկ այլ կարևոր հնարք է քայթինգը (kiting), որի ժամանակ առավել արագ վազող գործակալը ստիպում է հակառակորդին վազել իր ետևից, որի ընթացքում նա ժամանակ առ ժամանակ պտտվում է, վնաս հասցնում թշնամուն, ինչից հետո շարունակ փախուստի դիմում: SMAC-ում առկա են բազում քարտեզներ, որտեղ հակառակորդի բանակի չափերը էականորեն մեծ են գործակալների բանակի չափերից, ինչը ստիպում է գործակալներին դրսևորել կատարյալ համակարգված թիմային աշխատանք: Նկար 6-ում պատկերված է SMAC-ի մի քանի սցենար:

SMAC-ում գործակալները որպես դիտարկում ստանում են վեկտորի տեսքով տեղեկություն որոշակի հեռավորության վրա գտնվող թիմակիցների և թշնամիների վերաբերյալ: Այն պարունակում է հետևյալ տեղեկությունները. հեռավորություն, կյանքի մնացորդ, զրահի մնացորդ, զինվորի դասը: Մահացած, ինչպես նաև հեռու գտնվող թիմակիցների մասին տեղեկություն հասանելի չէ գործակալներին: Միջավայրի վիճակը, որը հասանելի է միայն մարզման ընթացքում, պարունակում է բոլոր զինվորների կորդինատները, կյանքի ու զրահի մնացորդները, փամփուշտների լիցքավորվածությունը և այլն:

Գործակալները կարող են շարժվել 4 ուղղությամբ, կանգ առնել, ինչպես նաև հարձակվել յուրաքանչյուր հակառակորդի վրա, ով գտնվում է բավականաչափ մոտ: Հաճախ հակառակորդին տեսնելուց հետո անհրաժեշտ է մոտենալ նրան կրակ արձակելուց առաջ: Բոլոր գործակալները ստանում են ընդհանուր պարգևատրում հակառակորդին վնաս պատճառելու, հակառակորդ զինվորներին ոչնչացնելու կամ ամբողջ մարտը հաղթելու արդյունքում:

SMAC-ը հաղորդակցվում է StarCraft II-ի շարժիչի հետ SC2LE գրադարանի[27] միջոցով:



Նկար 7: Հաղթանակի միջին գործակիցները SMAC-ի բոլոր սցենարներում:

3.5 ենթագլխում ներկայացված է SMAC հենանիշի կիրառման մեթոդաբանությունը: Ներառված է հենանիշի օգտագործման արդյունքում ստացված արդյունքների ներկայացման կարգը: Նաև նկարագրված են միջավայրի պարամետրերի թույլատրելի փոփոխությունները, ինչպես նաև անփոխարինելի կարգավորումները, որոնք նպաստում են մեթոդների արդար համեմատմանը:

3.6 ենթագլխում քննարկվում է SMAC հենանիշով գնահատված գոյություն ունեցող ԲԱՈւ ալգորիթմների արդյունավետությունը: Օգտագործվել են QMIX, VDN, IQL և COMA մեթոդները: Նկար 7-ում պատկերված են բոլոր 14 սցենարում մեթոդների փորձարկման արդյունքները, որոնք փաստում են, որ QMIX մեթոդը ցույց է տալիս առավել լավ արդյունք այլ մեթոդների համեմատ: Ենթագլխում նաև աղյուսակի տեսքով ներառված է բոլոր սցենարներում նույն մեթոդների ցույց տրված արդյունքները: Մարզման ավարտից հետո QMIX մեթոդի ցուցադրած արդյունավետությունը գերազանցում է երկրորդ լավագույն արդյունքն արձանագրած VDN մեթոդին 22.2% տոկոսով:

Ենթագլխում նաև մանրամասն ներկայացված են փորձարկված ԲԱՈւ ալգորիթմների արդյունքները յուրաքանչյուր սցենարի համար: Նկարագրված են մեթոդների հաջողությունների գրավականները կոնկրետ սցենարներում, ինչպես նաև մատնանշվում են դրանց սահմանափակումները (օրինակ՝ միջավայրի ոչ բավարար ուսումնասիրությունը): **3.7 ենթագլուխը** հակիրճ ամփոփում է SMAC հենանիշ միջավայրը՝ ընգծելով վերջինիս կարևությունը և սահմանելով հետագա բարելավման ծրագրերը:

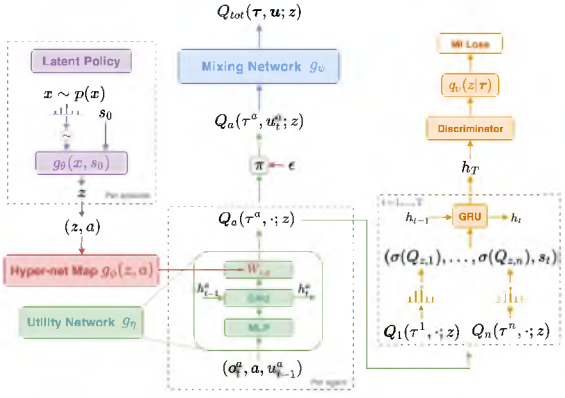
Գլուխ 4-ում առաջարկվում է նոր ԲԱՈւ մեթոդ, որը թույլ է տալիս տևական ժամանակահատվածում իրականացնել միջավայրի համակարգված ուսումնասիրություն:

4.1 ենթագլխում ներկայացվում է ԲԱՈւ-ում նշանակալի մի խնդիր՝ վիճակների և գործողությունների գերմեծ բազմությունների ուսումնասիրությունը: Բանն այն է, որ ժամանակակից ԲԱՈւ մեթոդներն օգտագործում են անհատ գործակալների կողմից իրականացվող պարզ, չուղղորդված հետազոտություններ: Այնուամենայնիվ, օպտիմալ ռազմավարությունը հաճախ պահանջում է բոլոր գործակալներից կամ գործակալների որևէ ենթաբազմությունից համատեղ ուսումնասիրել նոր ռազմավարությունը բազում քայլերի կամ ամբողջական էպիզոդների ընթացքում՝ համաձայնեցված ընտրելով պատահական գործողություններ:

Ցավոք, գոյություն ունեցող մեթոդներից ոչ մեկը զինված չէ միջավայրի ուսումնասիրության նմանատիպ ռազմավարությամբ, որի բացակայության արդյունքում ԲԱՈւ մեթոդները հաճախ սովորում են ոչ օպտիմալ ռազմավարություններ:

4.2 ենթագլխում տրվում է MAVEN (multi-agent variational exploration) մեթոդի նկարագիրը բազմագործակալ միջավայրների համակարգված ուսումնասիրության համար: MAVEN մեթոդում գործակալների Q-ցանցերը հիմնվում են z փոփոխականի վրա, որը յուրաքանչյուր էպիզոդից առաջ պատահականորեն ընտրվում է լատենտային տարածությունից օգտագործելով միջավայրի առկա վիճակը: Սա կատարվում է հիպերցանցերի շնորհիվ և ϵ-ագահ մեթոդին պարգևում է համակարգված ուսումնասիրություն իրականացնելու հնարավորություն, քանի որ բոլոր գործակալների համար z-ն ընդհանուր է:

[27] O. Vinyals, T. Ewalds, et al. StarCraft II: A New Challenge for Reinforcement Learning. arXiv:1708.04782, 2017.

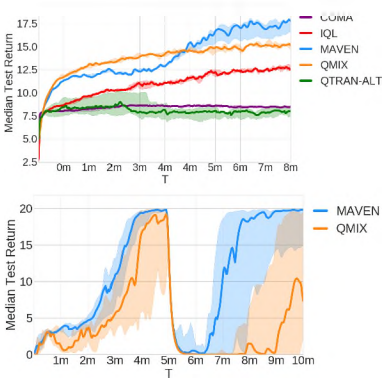


Նկար 8: MAVEN-ի ճարտարապետությունը:

Որպեսզի լատենտային z փոփոխականի յուրաքանչյուր արժեքին համապատասխանի միջավայրի յուրահատուկ համատեղ ուսումնասիրություն, օպտիմիզացվող ֆունկցիային գումարվում է լրացուցիչ անդամ՝ դիտարկված τ հետագծերի և համապատասխան լատենտային z փոփոխականների միջև ընդհանուր ինֆորմացիան (mutual information) մաքսիմիզացնելու համար:

Մշակված մեթոդն իր ճկունության շնորհիվ կարող է օգտագործվել ժամանակից յուրաքանչյուր խորքային ԲԱՈՒ մեթոդի հետ, ինչպիսինն են QMIX-ն ու VDN-ը: Նկար 8-ում պատկերված է MAVEN-ի ճարտարապետությունը՝ օգտագործված QMIX-ի հետ:

4.3 ենթագլխում ներկայացված են լայնածավալ փորձարկումների արդյունքներ, որոնք հաստատում են MAVEN մեթոդի առավել արդյունավետությունը COMA, IQL, QMIX և QTRAN[28] մեթոդների նկատմամբ: Փորձարկումներն արված են SMAC հենանիշի գերբարդ սցենարների վրա, որոնցում QMIX մեթոդին չի հաջողվում օպտիմալ ռազմավարություն մշակել: Ներկայացված են նաև պարզ մատրիցային խաղի և SMAC-ի նոր սցենարների արդյունքները, որոնց մշակման նպատակն է հատկապես միջավայրի ուսումնասիրությունը գնահատելը: Արդյունքներից կարելի է եզրակացնել, որ MAVEN մեթոդն արդյունավետությամբ գերազանցում է գոյություն ունեցող մեթոդները բարդ միջավայրի լայնածավալ ուսումնասիրություն պահանջող սցենարներում: Նկար 9-ում ներկայացված են մեթոդների ցուցադրած արդյունքները 6h_vs_8z և 2 corridors գերբարդ սցենարներում:



Նկար 9: Միջին շահույթ մարտի ընթացքում:

4.4 ենթագլխում նկարագրված են ժամանակից ԲԱՈՒ ալգորիթմներում կիրառվող ուսումնասիրության սկզբունքները, ինչպես նաև տրված է հակիրճ ակնարկ միագործակալ ԱՈՒ մեթոդներում կիրառվող ուսումնասիրության մեթոդների մասին:

4.5 ենթագլխում ամփոփված են գլուխ 4-ում առկա արդյունքներն ու հակիրճ շարադրված են հետագա հետազոտությունների նպատակները:

[28] K. Son, D. Kim, W.J. Kang, et al, QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning, Proceedings of the 36th International Conference on Machine Learning, PMLR 97, pages 5887-5896, 2019.

Աշխատանքի հիմնական արդյունքները

1. Մշակվել է խորքային ԲԱՈւ QMIX մեթոդը համագործակցային բազմագործակալ խնդիրների լուծման համար: QMIX-ը թույլ է տալիս ուսուցանել համատեղ արժեքային ֆունկցիաների հարուստ դաս, որի արդյունքում հնարավոր է առանձնացնել ապակենտրոնացված ռազմավարություններ, որոնք համահունչ են կենտրոնական, համատեղ ռազմավարության հետ: StarCraft II-ի ապակենտրոնացված միկրոկառավարման բարդ միջավայրում լայնածավալ փորձարարական արդյունքերը ցույց են տալիս, որ QMIX մեթոդի արդյունավետությամբ զգալիորեն գերազանցում է գոյություն ունեցող ԲԱՈւ ալգորիթմներին [1-3]:
2. Համագործակցային ԲԱՈւ մեթոդների մարզման և գնահատման համար մշակվել է SMAC բաց հենանիշ միջավայրը: Այն հիմնված է StarCraft II ռազմավարական խաղի վրա և պարունակում է 14 տարատեսակ սցենարներ, որոնք պահանջում են ԲԱՈւ մեթոդներին հաղթահարել միջավայրի մասնակի դիտարկելիության ու վիճակների մեծ բազմության մարտահրավերները: Ստեղծվել է PyMARL բաց պլատֆորմը՝ ժամանակակից ԲԱՈւ մեթոդների հետազոտման, ինչպես նաև նոր մեթոդների մշակման համար: Լայնածավալ փորձերի արդյունքում SMAC-ի օգնությամբ ներկայացվել են ժամանակակից ԲԱՈւ ալգորիթմների համեմատությունները [2, 3]:
3. Մշակվել է խորքային ԲԱՈւ MAVEN ալգորիթմը, որը թույլ է տալիս իրականացնել բազմագործակալ միջավայրի համաձայնեցված և համակարգված ուսումնասիրություն: MAVEN մեթոդն արձանագրում է բարձր արդյունավետություն SMAC հենանիշի գերբարդ սցենարներում, որտեղ գոյություն ունեցող մեթոդներին չի հաջողվել ուսուցանել օպտիմալ ռազմավարություն [4]:
4. Մանրազնին հետազոտելով ԱՈւ խնդիրն ու գոյություն ունեցող ԱՈւ մեթոդները՝ մատնանշվել են վերջիններիս օգտագործման խնդիրները բազմագործակալ միջավայրերում: Ակնարկ է արվել ժամանակակից ԲԱՈւ ալգորիթմների վերաբերյալ, և մատնանշվել են դրանց սահմանափակումները: Նկարագրվել են հեռանկարային մի քանի ուղղություններ հետագա աշխատանքների կատարման համար [5]:

Հրապարակված աշխատանքների ցանկ

1. T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, S. Whiteson, "QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning", *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, pp. 4295–4304, 2018.
2. M. Samvelyan, T. Rashid, C. Schroeder, G. Farquhar, N. Nardelli, T. Rudner, C.-M. Hung, P. Torr, J. Foerster, S. Whiteson, "The StarCraft Multi-Agent Challenge", *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pp. 2186–2188, 2019.
3. M. Samvelyan, T. Rashid, C. Schroeder, G. Farquhar, N. Nardelli, T. Rudner, C.-M. Hung, P. Torr, J. Foerster, S. Whiteson, "The StarCraft Multi-Agent Challenge", *Workshop on Deep Reinforcement Learning, NeurIPS 2019*, 11 pages, 2019.
4. A. Mahajan, T. Rashid, M. Samvelyan, S. Whiteson, "MAVEN: Multi-Agent Variational Exploration", *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 7611–7622, 2019.
5. Մ. Սամվելյան, «Համագործակցային խորքային բազմագործակալ ամրապնդմամբ ուսուցում. դժվարություններ, սկզբունքներ և բաց խնդիրներ», *ՀՀ ԳԱԱ և ՀԱՊՀ տեղեկագիր. տեխնիկական գիտությունների սերիա*, 72(3), է. 301–313, 2019:

Development and Evaluation of Efficient Deep Multi-Agent Reinforcement Learning Methods

Abstract

Machine learning (ML) has recently become a practical technology with widespread industrial usage across numerous domains. ML methods can automatically learn patterns from data and use them for making predictions or other outcomes of interest. ML algorithms are therefore commonly used in applications where the desired behaviour cannot be preprogrammed or the optimal solution is unknown.

Reinforcement learning (RL) is a subfield of ML concerned with how agents ought to act in an environment to maximise the rewards received from the environment. The recent success of deep learning, a family of methods for learning representations, allowed RL methods to advance the state-of-the-art of Artificial Intelligence on a number of challenging tasks, such as computer gaming, board games and robotic control.

Although most of the success of RL has been in the single-agent domain, many real-world environments consist of multiple agents that need to cooperate to achieve a common goal. For example, the coordination of autonomous drones, self-driving cars, and multi-robot systems in production factories is becoming increasingly critical. Nonetheless, methods for single-agent RL are poorly suited for such tasks due to several unique challenges, such as large action space, decentralisation, nonstationary and credit assignment. These challenges make it hard to successfully scale multi-agent RL (MARL) with a large number of agents which is crucial for building artificially intelligent systems that can productively cooperate to solve difficult tasks in a wide range of real-world applications.

A particularly challenging class of problems in MARL is partially observable, cooperative, multi-agent learning, in which a group of agents must learn to coordinate their behaviour while conditioning only on their private observations. This is an attractive research area since such problems are relevant to a large number of real-world systems. Partial observability and communication constraints often necessitate the learning of *decentralised policies*, which condition only on the local action-observation history of each agent. Fortunately, decentralised policies can often be learned in a centralised fashion in a simulated or laboratory setting. This could grant access to additional state information, otherwise hidden from agents, and remove inter-agent communication constraints. The paradigm of centralised training with decentralised execution has recently attracted attention in the RL community. However, many challenges surrounding how to best utilise the centralised training opportunity in the presence of many learning agents remain open.

The Goal and Problems Considered

The goal of this work is to develop and evaluate efficient MARL methods for complex multi-agent environments. To this end, we consider the following objectives:

1. Develop an efficient MARL method that takes full advantage of centralised training opportunity for training decentralised policies,
2. Build a challenging benchmark for MARL methods that allows to evaluate and compare their effectiveness and identify their limitations,
3. Develop a MARL method that allows decentralised agents to perform coordinated exploration of the environment.

Practical Applications

Cooperative MARL has tremendous real-world applicability. The algorithms designed within the scope of this work can be straightforwardly applied to numerous multi-agent systems are comprised of a group of agents that need to cooperate to solve a common task. For instance, they can be used for coordination problems for self-driving cars, unmanned aerial vehicles and other multi-robot systems. Furthermore, they can also be used for traffic routing, medical image analysis, fleet management and a great number of other multi-agent coordination tasks.

The open-source MARL framework developed within the scope of this work allows for out-of-the-box development and can be easily used for training on new multi-agent environments.

Provisions Presented for Defence

The following provisions are presented for defence:

1. A novel MARL method called QMIX that can train decentralised policies in a centralised end-to-end fashion. QMIX features a novel neural architecture that allows tractable maximisation of the joint action-value and guarantees consistency between the centralised and decentralised policies,
2. An extensive empirical evaluation of QMIX which illustrates that QMIX significantly outperforms the existing MARL methods. Thorough ablation studies that investigate the influence of the inclusion of extra state information and the necessity of non-linear transformations of agent Q-values in QMIX,
3. A novel benchmark environment for deep MARL called the StarCraft Multi-Agent Challenge (SMAC) which is based on the popular real-time strategy game StarCraft II and focuses on decentralised unit micromanagement. Rigorous evaluations of the existing MARL algorithms using the SMAC benchmark and revelations of their limitations that require algorithmic improvements,
4. A novel deep MARL method that achieves committed, temporally extended exploration of multi-agent environments. An extensive empirical evaluation of the developed exploration method against existing algorithms.

Thesis Structure

The total length of the thesis is 116 pages. The thesis is comprised of an introduction, four chapters, a conclusion and bibliography. The work includes 27 figures and 8 tables.

Разработка и оценка эффективных методов глубокого многоагентного обучения с подкреплением

Резюме

Машинное обучение (МО) в последнее время стало практичной технологией и широко используется во многих областях промышленности. Методы МО могут автоматически извлекать закономерности из данных и использовать их для прогнозирования или других решений, представляющих интерес. Поэтому алгоритмы МО обычно используются в приложениях, где желаемое поведение не может быть запрограммировано или оптимальное решение неизвестно.

Обучение с подкреплением (ОП) - это область МО, связанная с тем, как агенты должны действовать в окружающей среде, чтобы максимизировать вознаграждения, получаемые от окружающей среды. Недавние успехи глубокого обучения (семейство методов для обучения представлений) позволили методам ОП продвинуть область искусственного интеллекта в таких областях, как компьютерные игры, настольные игры и управление роботами.

Хотя большая часть успеха ОП была достигнута в средах с одним агентом, многие реальные среды состоят из нескольких агентов, которые должны сотрудничать между собой для достижения общей цели. Например, координация автономных беспилотных летательных аппаратов, автомобилей с автоматическим управлением и систем с несколькими роботами на производственных предприятиях преобретает все более важное значение. Все же методы одноагентного ОП плохо подходят для решения таких задач из-за нескольких уникальных проблем, таких как большое пространство действий, децентрализация, нестационарность и распределение кредитов. Эти проблемы затрудняют успешное масштабирование многоагентного ОП (МОП) с большим количеством агентов, которое имеет большое значение для создания искусственных интеллектуальных систем, способных эффективно взаимодействовать для решения сложных задач в широком спектре реальных приложений.

Особенно сложным классом проблем в МОП является частично наблюдаемое, кооперативное, многоагентное обучение, в котором группа агентов должна научиться координировать свое поведение, опираясь только на свои личные наблюдения. Это очень привлекательная область исследований, поскольку такие проблемы актуальны для ряда реальных систем. Частичная наблюдаемость и коммуникационные ограничения часто требуют изучения децентрализованной стратегии, которая опирается только на локальной истории наблюдений и действий каждого агента. К счастью, децентрализованную стратегию можно изучать централизованно в моделируемой или лабораторной обстановке. Это может предоставить доступ к дополнительной информации о состоянии среды, иначе скрытой от агентов, и устранить ограничения связи между агентами. Парадигма централизованного обучения с децентрализованным исполнением недавно привлекла внимание сообщества ОП. Тем не менее, многие проблемы, связанные с тем, как максимально использовать возможности централизованного обучения в присутствии многих обучающих агентов, остаются открытыми.

Цель и рассматриваемые задачи

Целью данной работы является разработка и оценка эффективных методов МОП для сложных многоагентных сред. Для этого мы рассмотрим следующие задачи:

1. Разработка эффективного метода МОП, который полностью использует возможности централизованного обучения для подготовки децентрализованных

стратегий,

2. Создание бенчмарка для методов МОП, который позволит оценить и сравнить их эффективность и выявить их ограничения,
3. Разработка метода МОП, который позволяет децентрализованным агентам выполнять скоординированное исследование окружающей среды.

Практические применения

Кооперативное МОП имеет огромное практическое применение. Алгоритмы, разработанные в рамках данной работы, могут быть непосредственно применены к многочисленным многоагентным системам, включающим группу агентов, которые должны взаимодействовать для решения общей задачи. Например, они могут быть использованы для решения проблем координации для автомобилей с автоматическим управлением, беспилотных летательных аппаратов и других мульти-робототехнических систем. Кроме того, они также могут использоваться для маршрутизации трафика, анализа медицинских изображений, управления транспортных средств и множества других задач координации нескольких агентов. Программное обеспечение с открытым кодом, разработанное в рамках данной работы, также дает возможность с легкостью применять методы МОП для обучения в новых многоагентных средах.

Положения представленные для защиты

Для защиты представлены следующие положения:

1. Новый метод МОП под названием QMIX, который может обучать децентрализованной стратегии централизованным способом. QMIX обладает новой нейронной архитектурой, которая позволяет добиться эффективной максимизации совместных действий и обеспечивает согласованность между централизованной и децентрализованной стратегиями,
2. Обширные эмпирические результаты для метода QMIX, которые показывают, что QMIX значительно превосходит существующие методы МОП. Тщательные исследования иллюстрируют важность включения информации о состоянии среды и необходимости нелинейных преобразований Q-значений агентов в QMIX,
3. Новая среда тестирования для глубокого МОП под названием StarCraft Multi-Agent Challenge (SMAC), которая основана на популярной стратегической игре StarCraft II и ориентирована на задаче микроуправления, где каждая единица контролируется независимым агентом и должна действовать на основе локальных наблюдений. Обширное исследование и сравнение существующих алгоритмов МОП с помощью SMAC, чтобы понять их недостатки, требующие улучшения алгоритмов,
4. Новый метод глубокого МОП, который позволяет нескольким агентам продолжительное времемя совершать систематическое исследование многоагентной среды. Тщательное эмпирическое исследование и сравнение разработанного метода с существующими алгоритмами МОП.

Структура диссертации

Общий объем диссертации составляет 116 страниц. Диссертация состоит из введения, четырех глав, заключения и списка литературы. Работа включает 27 рисунков и 8 таблиц.