# Introduction

With ithe iamount iof icomputing ipower iand ithe ihuge iamount iof idata iavailable ifor ithe iresearchers iand icompanies ito idraw iout iinsights iand iproviding ius iwith iinformation iabout ihow, iwhat, iwhen iand iwhere ihas itotally ilifted ithe iuse iof iAI iand iMachine iLearning ito inew iheights.

Machine iLearning ion ia ivery icoarse ilevel ihas ibeen idivided iinto i3 imajor iareas, inamely iSupervised iLearning i[1], iUnsupervised iLearning i[2], iand iReinforcement iLearning i[3].

* Supervised iLearning i[1], iit imainly ideals iwith ilabeled idata. iLabeled idata imeans ithat iwhen ithe idata iis ibeing itrained, iwe ialready ihave itold ithe ilearning ialgorithm iabout igiven iinputs iand iexpected ioutputs. iA ifew iexamples iof isupervised imachine ilearning iare iof ilinear iregression, inaïve ibayes, ilogistic iregression, iK-Nearest iNeighbor, iSupport-Vector iMachine, idecision itree, ineural inetworks ietc. iSo ithe imain irequirement ifor ithis itype iof imachine ilearning iis ilabeled idata.
* Unsupervised iLearning i[2], imainly iuses imachine ilearning ialgorithms ito ianalyze iand icluster iunlabeled idatasets. iThese ialgorithms itry ito idiscover ihidden ipatterns ior igroupings iof idata iwithout ithe ineed ifor ihuman iinterference. iThese iabilities ito idiscover isimilarities iand idifferences iin iinformation imakes iit ithe iideal isolution ifor idifferent iapproaches ilike iexploratory idata ianalysis, ibio-informatics, iclustering, ianalysis iof isequences, isegmentation iof icustomers, iimage irecognition iand imany imore. iFew iof ithe iunsupervised ilearning iapproach iare iK-Means iClustering, iAssociation iRules, iHierarchical iClustering,Neural iNetworks.
* Reinforcement iLearning i[3] ifinds iits iroot iin iPsychology. iAs iin iday ito iday ilife iwe icome iacross imany isituations iwhich iforces ius ito itake ia isome iactions. iAfter itaking isuch iactions iwe iget ieither ipositive ireward ior inegative ireward. iThis iis ithe isame iidea iand ithought iprocess ibehind ireinforcement ilearning. iYou itrain ia iagent isuch ithat ibased ion icertain istate iit’s ipresent iin, iit itakes ifew iactions iand igets irewards. iNow ibased ion iwhat iwe iwant ito ido iwith ithose irewards iis ithe imain iidea ithat iis ibeing iexplored inowadays. iReinforcement ilearning ihas istarted ito igain irecognition imainly iin ithe ifield iof irobotics, isupply ichain imanagement, iself- idriving icars, icomputer ivision, ifinance iand itrading iand imuch imore.

So istarting iwith ireinforcement ilearning, ithe ifirst iever iidea ithat istarted iwas imainly iembedded iin ipsychological ijournals. iTaking inotes iand iinspiration ifrom ihow ihumans ireact ito idifferent imethods iof ilearning. iFor iexample, iif ia ichild iis irunning ia irace iand ihe/she iwins ithe irace, ithe ichild iis irewarded iwith ia ichocolate ior ia imedal ior iany isort iof iappreciation. iThis iis iknown ias i“Positive iReinforcement”. iIf ia ichild iis iplaying iin ia iplayground iand ihe iaccidently itouches ithorny ibush, ihe iimmediately istarts icrying ibecause ithe ithorn iis icausing ia ipain. iThis iis ian iexample iof i“Negative iReinforcement”.

The ihistorical ibackdrop iof ireinforcement ilearning igenerally ideals iwith itwo imain iideologies. iOne iideology iis iconcerns iitself iwhen ione ilearns iby ithe imethod iof itrial iand ierror iwhich iinitiated iwith ithe iunderstanding iof ianimal ipsychology iof ihow ithey ilearnt iand iresponded. iThis iidea iruns ialong iwith isome iof ithe iearlier iworks iin ithe ifield iof iartificial iintelligence iwhich iproceeded ito ithe resurrection iof ithe ifield iknown ias ireinforcement ilearning iin ithe i1980s. iThe inext iideology iit imatters iitself iis iwith ithe iproblem iof ioptimal icontrol iand iwhose isolution iis iby igenerally ifound iby iusing ithe iapproach iof idynamic iprogramming iand ivalue ifunction. iFor ithe imost ipart, ithis iidea ihas inot iconsidered ilearning. iAlthough iboth ithe iideas ihave ibeen iquite iindependent, ithough ithere iare iomissions irevolving iaround ia ithird iand iless itransparent iidea iconcerned iwith itemporal-difference imethods. iAll ithe ithree iideologies iculminated itogether iin ithe ilate i1980s ito iproduce iwhat iwe iknow ias itoday’s imodern ifield iof ireinforcement ilearning.

The iterm iknown i"optimal icontrol" iarose iin iexistence ias iof ilate i1950s iwhich iwas iused ito idescribe ia idesign iproblem idealing iwith ia icontroller iwhich iminimizes ithe iparameter iof ia idynamic isystem iand iits ibehavior ia iperiod iof itime. iOne iamong imany iof isuch iways ito ithis iproblem iwas ithought iof iin ithe imid-1950s iby iRichard iBellman iand iother iresearchers iwhich iwas ijust ithe iextension iof ia itheory iproposed iby iHamilton iand iJacobi iin ithe i19th icentury i[4]. iThe iparticular iapproach iused ithe iconcept iof ithe istate iof ithe idynamic isystem iand ithat iof ia ivalue ifunction, ior iin iother iwords ian i"optimal ireturn ifunction," i[4] iwhich iwas iused ito idefine ian iequation, inow ifamously iknown ias ithe i“Bellman iequation”. iThese itypes iof iapproaches ithat iaim iat isolving ithe icontrol iproblem ioptimality iby isolving ithe i“Bellman iequation” iis iwidely ito iknown ias i“dynamic iprogramming”. iBellman ialso ipopularized ithe istochastic ias iwell ias ithe idiscrete iversion iof ithe icontrol iproblem ioptimality ifamously iknown ias iMarkovian idecision iprocesses i(MDPs), iand iRon iHoward iwas ithe ione iwho iconcluded ithe ipolicy iiteration imethod ifor iMDPs i[14]. iAll iof ithe iidea idiscussed iabove iare ithe imost iessential ielements iwhich iunderlie iin ithe itheory iand ialgorithms iof itoday’s imodern iday ireinforcement ilearning.

Dynamic iprogramming istill ito idate iis iwidely iconsidered ithe ionly ifeasible iway ito isolve igeneral istochastic ioptimal icontrol iproblems. iSuch ioptimal iproblems isuffer ifrom iwhat iBellman iused ito iidentify ias i"the icurse iof idimensionality," iwhich imeans ithat ias ithe inumber iof istate ivariable iin ithe igiven iproblem iincreased ithe icomputational irequirement ineeded ito iactually icompute ithe ivalues iof ithe ivariable iincrease iexponentially. iBut ieven iafter ithese isetbacks iit iis istill iway imore iefficient, iaccepted iand imore iwidely iapplicable ithan iany iother igeneral imethod. iDynamic iprogramming ihas ia ivery iextensive iarc iof idevelopment istarting isince ithe ilate i1950s, iwhich ialso iincludes iand iextension ito ithe inew iand iinteresting ifield iof ipartially iobservable iDPs.

Now iagain ifocusing iback ion ithe iother imajor iidea iwhich icurrently ileads ius iinto ithe icurrent iand inew iarea iof ireinforcement ilearning irevolves iaround ithe iidea ito ilearn ifrom itrial iand ierror. iThis ispecific iidea ifind iits iroot iin ipsychology, iwhere i"reinforcement" itheories iof ilearning iare iabundant iand icommon. iPerhaps ithe ifirst ione ito iexpress ithe iessence iof itrial-and-error ilearning isuccinctly iwas iEdward iThorndike, iwho iin ia inutshell isaid ithat iany iaction irewarded iwith igood ii.e. ipositive ior ibad i.e. inegative ioutcomes ihave ia itendency ito ibe ire-elected iand ichanged iaccordingly. iThorndike iproposed ithis iphenomenon ito ibe iknown ias ithe i"Law iof iEffect" ibecause iit idescribed ihow ithe itendency iof iselecting idifferent iactions iwere iactually ieffected iby ithat iof ireinforcing ievents. Now iunderstanding ithe ithird iidea, idealing iwith ithe ihistory iof ireinforcement ilearning, iconcerned iwith ithe iidea iof ilearning iknown ias itemporal-difference. iTemporal- difference ilearning imethods iare ivery iunique iin ithe iidea ithat ithey iare idriven iby ithe idifferences iin ibetween itemporally iconsecutive iestimates iof ithe isame iquantity, ilike ifor iexample ithe iprobability ito iwin ia igame iof itic-tac-toe i. iThis iapproach iis iless idistinctly iused ithan ithe iother itwo iapproaches, ibut iit ihas iplayed ia iparticularly iimportant irole iin ithis ifield imostly ibecause itemporal-difference imethods iseemingly iare iquite inew iand iunique iin ithe icurrent ifield iof ireinforcement ilearning.

The iorigin idealing iwith ithat iof itemporal-difference ilearning ican iagain ibe iin ipart ibe iattributed ito ipsychology iof ianimal ibehavior iand ilearning inotably iin ithe iconcept iknown ias i“*secondary ireinforcers”*. iA i*secondary ireinforcer i*is ia istimuli ithat iis igenerally ipaired iwith ia i*primary ireinforcer i*that ican ibe ifood ior ipain iand idue ito ithis ias ia iresult ilead ito ia iinterchange iin ithe ireinforcing iproperties. iThe ioptimal icontrol iand itemporal-difference iideas ifully icoincided itogether iin i1989 iwhen iChris iWatkins's iproposed iand ideveloped ia inovel iconcept iknows ias i“Q-learning” i[7].

# Literature iSurvey

## Components iof iReinforcement iLearning i[4]

After ithe ibrief iintroduction iof ireinforcement ilearning, ithere iare ia ifew ikey iconcepts ithat iform ian iintegral ipart iof ireinforcement ilearning iecosystem. iFor ithat ihere iare ithe imain ior ikey icomponents iof ia ireinforcement ilearning ialgorithm:

* + 1. **Agent i[4]: i**The imain icomponent iof ia ireinforcement ilearning ialgorithm iis iagent. iThisiis ithe ientity ithat itake iactions iin ia iparticular ienvironment. iFor iexample: iAn iindustrial irobot ipicking iup inon-biodegradable iitems iin ia irecycling iplant.
    2. **Environment i[4]: i**The isurrounding iin iwhich ithe iabove imentioned iagent iwill iwork ior iperform iits iactivities. iFor iexample ifor ian iagent iin icomputer ivision ithe ipart iof iimage ienclosed iby ithe ibounding ibox ican ibe ithe ienvironment.
    3. **State i[4]: i**This iis ithe idefined ias ithe ipresent istatus iof ithe iinteracting iagent iin ia iparticular ienvironment. iFor iexample iis ithe irobot iarm iin imotion ior iit iis istationary ior iis iit igoing iup ior iis iit igoing idown.
    4. **Action i[4]: i**Action iis ithe isteps ithe iagent itakes. iMore ioften ithan inot iin ipractical iapplications icollection iof iactions icalled ias iaction ispace iis idiscrete iin inature. iFor iexample ithe irobotic iarm imoving ileft, iright, iup, idown, iclose, iopen ietc.
    5. **Reward i[4]: i**This iis idefined ias ithe iobjective ithat iis iimplicitly ior iexplicitly idefined ifor ithe iagent ito itake iactions iand imaximize.

After idiscussing iand igoing ithrough ia icouple iof imain ibuilding iblocks, ithere ihas ito ibe components ithat ibring ithese ibuilding iblocks itogether ibecause ithey ialone ican’t iact ior execute iindependently. iSo iconnection iall ithe iabove icomponents ihere iare ithe ibinding concepts iof iReinforcement iLearning. iThey iare:

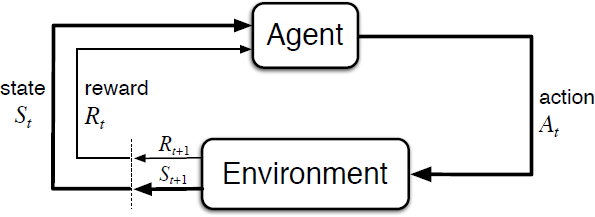
1. **Policy i[4]: i**A i*policy i*defines ilearning ithe ibehavior iof ian iagent iand iit’s iway iat ia igiven iinstance iof itime. iIn iother iwords, ia ipolicy ican ibe idefined ias ia imapping itechnique iwhich imaps ithe iagent’s ibehavior ifrom iperceived istates iof ithe ienvironment ito iactions iit imust itake iwhen iit iis ipresent iin ione iof ithose istates. iThis ican ibe icorrelated ito ipsychological iterms iknow ias iassociations ior istimulus- iresponse irule. iIn isome iof ithe icases ithe ipolicy iinvolved ican ibe ia ivery isimple ifunction ior iit ican ibe ia ilookup itable, iwhile iin iother iinstances iit ican iinvolve ivery iin-depth icomputation, ifor iexample ia iprocess ito isearch. iThe ipolicy iis iat ithe icore iof ia ireinforcement ilearning iagent iwhich imeans ithat iall iby iitself, iit iis isufficient ienough ion iit’s iown ito idetermine ithe ibehavior iof iagents. iIn igeneral isence, ipolicies ican ibe istochastic.
2. **Reward iFunction i[4]: i**A i*reward ifunction i*is ithe ione iwhich igives ithe idefinition iof ia igoal iin ia iproblem iof ireinforcement ilearning. iIn iother iwords, iit imaps ieach istate-action ipair, iin iother iwords iknows ias iperceived istate, iof ithe ienvironment iin iwhich ithe iagent iis ipresent ito ia iunique icountable inumber icalled ia i*reward*, iwhich iindicates iintrinsically ihow idesirable ia igiven istate iis. iThe isole ipurpose iand iobjective iof ia ireinforcement ilearning iagent iis ito imaximize ithe itotal ireward iit iwill ireceive iin ithe ilong irun iof ithe iexperiment iinstead iof ishort iterm igain. iThe ireward ifunction ialso idefines iwhat igood ior ibad ievents iare ifor ithe iagent iin iterms iof ithe ireward iit igains. iIn ia isystem iof ibiology, iit iwould ionly ibe iappropriate ito irelate iand ianalyze ibiological irewards iwith ieither ipleasure ior ipain. iThese iare ithe iinstantaneous iand icrucial ifeatures iof ithe iproblem ithe iagent ifaces iin ian iexperimental isetting. iThe ireward ifunction ishould imandatorily ibe iunchanged iby ithe iagent. iIt ican iinstead ibe iuse ias ibe ia ibasis ito ialter ithe iagent’s ipolicy. iFor iexample, iif ian iaction iselected iby ithe ipolicy iis ifollowed iby ilow ireward, ithen ithe ipolicy imay ibe ichanged ito iselect isome iother iaction iin ithat isituation iin ithe ifuture. iIn igeneral, ireward ifunctions ican ibe istochastic.
3. **Value ifunction i[4]: i**A i*value ifunction i*is ithe ione ithat itells ithe iagent ispecifically iabout ithe igoodness iof iwhat iit iis ifollowing ifor iit’s iown iin ithe idistant ifuture. iIn iother iwords iwe ican isay ithat ithe i*value i*of ia iparticular istate iis ithe itotal iamount iof ireward ian iagent iexpectedly iaccumulates iover ithe ifuture irun iof iit’s iexperiment, istarting ifrom ithat ispecific istate. iRewards iare ithe iparameters iwhich ihelps ithe iagent ifigure iout ithe istates iimmediate iunderlying idesirability, iwhereas ivalues ispecify ithe i*long-term i*appeal iof istates iafter itaking iinto iconsideration iof iall ithose istates ithat iare ilikely ito iappear iin ifuture iand ithe irelated irewards ithat imight ibe iavailable ifor ithose istates. iLet’s isay ifor iexample ia istate imay ialways iyield ia ilow iimmediate ireward ibut istill iyield ia ihigh ivalue ibecause ithe ifuture istates ithat ifollows ithe icurrent istate itends ito iearn iquite ihigh irewards ior ivice iversa. iLet’s itake ifor iexample ihuman ianalogy. iRewards ican ibe iequated ito ipleasures iif ithe ireward iis ihigh iand ialternatively ipain iif ithe ireward iis ilow, iwhereas ivalues icorrelating ito ia imore irefined iand ifuturistic ijudgment iof ihow isatisfied ior idissatisfied iwe iare ithat ithe ienvironment iis ipresent iin ithat iparticular istate.
4. **Model i[4]: i***Model i*can ibe ithought iof ias ione ithat itries ito icopy ithe ibehavior iof ithe ienvironment. iLet’s isay ifor iexample, iwe iknow ithe istate iand ithe iaction, ithe imodel imight ibe iable ito ipredict ithe iexpected inext istate iand icorresponding inext ireward. iModels iare igenerally iused ifor i*planning*, iwhich idecides ion ithe icourse iof iaction ithe iagent imust itake iby itaking iinto iconsideration iall ithe ifuture isituations ithat iare ipossible ibefore ithey iactually icome ito ifruition. iThe iamalgamation iof models iand iplanning iin ireinforcement ilearning isystems iis iquite irelatively ia inew idevelopment. iEarly isystems iof ireinforcement ilearning iwere istraightforward ilearners iof itrial iand ierror i, iwhich iis iactually ithe itotal i*opposite i*of iplanning. iNonetheless, islowly iand igradually iit ibecame iclearer ithat ireinforcement ilearning imethods iare iquite iclosely irelated ito idynamic iprogramming imethods iwhich iin- iturn idoes iuse imodels iand iare iin iturn iclosely irelated ito iplanning imethods iinvolving istate iand ispace. iModern ireinforcement ilearning icovers ithe iwhole ispectrum istarting ifrom ilow-level itrial iand ierror ilearning ito ihigh-level ideliberative iplanning.

Rewards iin ia iway iare iprimary iparameters, iwhereas ithe ivalues, ias ipredictions iof rewards, can ibe iunderstood ias isecondary iparameters. iWithout irewards ithere iwould inot ibe anyivalues iand ithe ione iand ionly ipurpose ito iestimate ithe ivalues iis ito inevertheless achieve ithe imost ireward. iRegardless iit iis ivalues iwith ione ishould ibe imost iconcerned about iwhen imaking iand ievaluating ithe idecisions. iValue ijudgements iare ithe ione’s ibased on iwhich iactions iare iexecuted. iThe iaim iis ito itry ito iseek iout iactions ithat iconvey iaround the istates iof ihighest ivalue ibut inot ithe imost ireward imainly ibecause isuch iactions iprocure the igreatest iamount iof ireward ifor ithe iagent iover ithe ilong iterm.

In iplanning iand idecision-making ithe iborrowed iquantity iis icalled ivalue iand iit iis ione iof the imain iparameters. iUnfortunately, iit iis ieasier ito idetermine ithe irewards irather ithan determining ivalues. iRewards iare idirectly igiven iby ithe ienvironment ithe iagent iis ipresent in but ivalues iare ithe ione’s ithat ineeds ito ibe iestimated iand ire-estimated ifrom ithe isequences of iobservations ian iagent imakes iover iits ientire ilifetime. iAs ia imatter iof ifact ithe imost important iconstituent iof inearly iall ithe ireinforcement ilearning ialgorithms iis ia imethod ion how ito ieffectively iestimate ivalues. iThe ipivotal irole iof iestimation iof ivalues iis iarguably the imost iessential ithing ithat iis iknown iabout ireinforcement ilearning iover ithe ipast icouple of idecades. iAlthough iit’s inot inecessary ito iestimate ivalue ifunctions iwhen isolving ia reinforcement ilearning iproblem. iLike ifor iexample imethods iused ifor isearching isuch ias genetic ialgorithms ior igenetic iprogramming iand imuch imore, imethods ifor ioptimization have ialready ibeen iwell-used ito isolve iproblems idealing iwith ireinforcement ilearning. These methods isearch iprecisely iin ithe ipolicy ispace iwithout iever iengaging iinto ivalue ifunctions. These iare iknown ias i*evolutionary i*methods ifor ithe ivery ireason ibecause ithe iway ithey operate iis isimilar ito ithe iway ievolution ibiologically iproduces iliving iorganisms iwith skilled ibehavior ieven ithough ithey ihave inot ilearnt iit iduring itheir iindividual ilifetimes. iIf the ipolicy ispace iis iadequately ismall ior ican iin isome iway ibe istructured iso ithat igood policy iimplementations ibecome icommon ior ieffortless ito ifind ithen ievolutionary imethods can ibe ivery ieffective. iAlso iin iaddition ito ithis, ievolutionary imethods ihave iseveral advantages ion iproblems iin iwhich ithe ilearning iagent icannot iaccurately iunderstand ithe istate of iits ienvironment.

Nonetheless, iwhat iit imeans iis ithat ireinforcement ilearning ilearns iduring iit’s iinteraction with ithe ienvironment iwhich ievolutionary imethods ido inot ido. iEvolutionary imethods generally itends ito idiscount imuch iof ithe iuseful istructures iof ia iproblem idealing iin reinforcement ilearning. iSuch methods ido inot iconsider ithe ifact ithat ithe ipolicy ibeing searched ifor iis iactually ia ifunction imapping ifrom istates ito iactions iand ido inot iconsider which istates ian iagent ivisits ithrough iduring iits ilifetime ior iwhich iactions ithe iagent selects. iIn isome icases isuch iinformation ican imislead i(e.g., iwhen istates iare misinterpreted).

## Markov iDecision iProcess i[4]



**Fig. iAgent-Environment iInteraction iin ia iMarkov iDecision iProcess i[4]**

Markov iDecision iProcess i(MDPs) i[4] iin isimple iwords iare inothing imore ibut ia idirect framing iup iof ia ilearning iproblem ifrom iinteractions ito iachieve ia ispecific igoal. iThe imain entity ior ilearner iand ialso ithe idecision imaker iis iknown ias ian i*agent*. iThe ithing i*agent* interacts iwith ievery isingle itime iwhich icomprises ieverything ioutside ithe i*agent*, iis icalled an i*environment*. iThey iboth iinteract iagain iand iagain iwith ithe iagent iselecting iactions iand the ienvironment iresponding ito ithese i*actions i*and ipresenting inew isituations ibefore ithe *agent*. iThe i*environment i*also igives irise ito i*rewards*, iwhich iare inothing ibut isome ispecial numerical ivalues ithat ithe i*agent i*tries ito i*maximize i*over itime ithrough iits ichoice iof *actions*.

If ithe istates iand iaction ispaces iare ifinite, ithen ithe iproblem iso iformed iis iknown ias ia finite imarkov idecision iprocess i(fMDP). iFinite iMDPs iare ivery iimportant ifor ireinforcement learning i[3] iproblems iand imost iof iliteratures iout iin ithe ischolarly iworld ihave iassumed that ithe ienvironment iis ia ifinite iMDP iin itheir iworks.

Any ireinforcement ilearning iproblem ican ibe imodeled ias ia iMarkov iDecision iProcess i[14]. Markov iDecision iProcesses iare ia iclassic iformulation iof isequential idecision imaking, where iactions itend ito iinfluence inot ijust iimmediate irewards, ibut ialso isubsequent situations ior istates, iand ithrough ithose, ifuture irewards i[4].

MDPs iinvolves idelayed ireward iand iputs ion ia iheavy iemphasis ion ithe ineed ito iadjust ibetween iimmediate iand idelayed irewards. iConversely iin iproblems icontextualized ias ibandit iproblems, ithe ivalue i*q\*(a) i*of ieach iaction i*a i*is iestimated, iwhereas iin iMDPs iwe itry ito ipredict ithe ivalue i*q\*(s, ia) i*of ieach iaction i*a i*in ieach istate i*s*, ior iwe itend ito ipredict ithe ivalue i*v\*(s) i*of ieach istate igiven ioptimal ipreference iof iactions. iThese istate-dependent iquantities iare ivery icrucial iin iorder ito iaccurately iassign icredit ifor ilong-term iimportance ito iindividual ielection iof iactions.

The iagent iand ienvironment iinteract iwith ieach iother iat ieach iof ithe idiscrete itime isteps idenoted iby i*t* = i0, i1, i2, i3,… iAt ieach idiscrete itime istep i*t*, ithe iagent ireceives isome isort iof irepresentational iinformation iof ithe ienvironment’s icurrent i*state i*given iby *iSt, i*and ion ithe ibasis iof ithe istate iinformation iit ithen iselects ian i*action i*given iby i𝐴𝑡 i∈ i𝐴(𝑠)*. i*Now iafter ione itime istep ireflecting ithe irepercussion iof ian i*action*, ia ireward iis ipresented ito ithe iagent iwhich iis igiven iby i*Rt+1 i*∈ i*R i i*and i ithus ilands iitself iin ia inew istate idenoted iby i*St+1. i*Thus iall ithis iinformation iabove igives irise ito ia isequence ithat ilooks ilike i*S0, iA0, iR1, iS1, iA1, iR2, iS2, iA2, iR3,… i*[4]

Now iin ia i*finite i*MDP, ithe imultiple iset iof istates, iactions, iand irewards i(*S,A,R) i*have iall ibut i*finite i*number iof ielements iin ithem. iIn isuch icases, ithe ireward i*Rt i*and istate i*St i*both ihave iwell idefined idiscrete iprobability idistributions iwhich iis ireliant ionly ion ithe ipreceding istate iand icorresponding i*action*. iNow ifor ia ispecific ivalue iof ithe irandom ivariables, i*s’ i*and i*r*, ithere iis ia iprobability iof ithese ivalues ioccurring iat ia igiven itime i*t i*due ito ithe ivalues iparticularly igiven ifrom ithat iof ithe ipreceding istates iand icorresponding iactions iwhich ican ibe igiven iby ithe iequation imentioned ibelow:

i[4]

for iall i*s’, is i iS, ir iR, i*and i*a* i i*A(s).* i

The ifunction i*p i*here irepresents ithe isystem idynamics ior ithe imeasurements iof ithe ispecific iMDP. The idot ion ithe iequals isign iin ithe iequation iconveys ithe iinformation ithat iit iis ia idefinition, iparticularly i iin i ithis i icase, i ithe i ifunction i i*p i i*rather ithan i ia ifact i ithat i ifollows i ifrom i igiven iprevious idefinitions. iThe ifunction iof idynamics igiven iby p i i[0, i1] iis ian iordinary ideterministic ifunction iwhich iconsists iof ifour ielements. iThe i‘|’ iis iborrowed ifrom ithe inotation iwhich idenotes iconditional iprobability, ibut iin ithis icontext iit ijust ipoints iout ito ius ithat i*p i*indicates ia iprobability idistribution ifor ieach ichoice iof i*s i*and i*a, i*that iis igiven iby ibelow i[4],



In ia iMarkov idecision iprocess, ithe iprobabilities iconveyed iby i*p i*comprehensively icharacterizes ithe ienvironment’s idynamic inature iwhich iin iother iwords ican ibe iunderstood ias ithe iprobability iof ievery iachievable ivalue ifor i*St i*and i*Rt i*depending ionly ion ithe iimmediate ipreceding istate iand iit’s icorresponding iaction, igiven iby i*St−1 i*and i*At−1 i*respectively iand igiven iboth iof ithe ivalues iand inot iat iall ion iearlier istates iand/or iactions. iThis ican ibe iviewed ias ia irestriction ion ithe istate, ibut inot ion ithe idecision iprocess iitself. iThe istate igenerally iincludes iall ithe inecessary iinformation iabout iall iaspects iof iall ithe iprevious iagent-environment iinteractions iwhich imakes ialways imakes ia idifference iin ithe ifuture. iSo iif ithe iall iof ithe iabove iconditions iare imet iand ifollowed, ithen ithe istate iis iunderstood ias ito ifollow ia i*Markov iProperty. i*[4] iDiscount ifactors iare iassociated iwith itime ihorizons ias igiven ibelow i[4].

C:\Users\anind\Downloads\1_NKX4zhTk9YKx4J5b0PKbBA.png

The iMDP iframework iis iquite iconceptual ibut ias iwell ias ivery imalleable iand ifinds iit’s iapplication iin imany idistinct iproblems iin ivaried iways. iLike ithe itime isteps iare inot ineeded ito ihave ifixed iintervals iof ireal itime; ithey iinstead ican ihave iarbitrary iconsecutive idecision imaking istages iand iperforming ian iaction. iThe iactions ican ibe iof ilow-level icontrols ilike ithe ivoltages iapplied ito ithe imotors iof ia irobotic iarm, ior ihigh-level idecisions ilike iwhether ior inot ito igo iand iwatch ia imovie ior ito igo ito ia iseminar ior inot. iSimilar ito iactions, istates ican itake ia ilot iof idifferent iforms. iThey ican ieither ibe icompletely iresolved iby ilow-level isensations, ilike idirect isensor ireadings, ior ithey ican ibe ihigh-level iand iabstract iin ia isense ias ithat iof ia isymbolic idescription iof iobjects ipresent iin ia iroom.

The iMDP iframework ican ibe ithought iof ias ia iproblem ithat ideals iwith ia idirected imethod iof ilearning iwith ithe igoal ito ilearn ifrom ivarious iinteractions. iIt icomes iup iwith ithe iidea ithat iwhatever ithe idetails iof ithe isensors, imemory, iand icontrol isetup iare ithere iand iwhichever iobjective ione iis itrying ito isolve, iany isuch iproblem iof ilearning ia igoal-specific iactions ican ialways ibe ireformulated ias ian iensemble ithree isignals icommunicating iin ibetween ian iagent iand iits ienvironment: ione iof ithe isignal iwill irepresent ithe ichoice ian iagent imakes i(i.e. iactions), ione isignal iwill irepresent ithe imeasure ion iwhich ithe ichoices iare ibeing imade i(i.e. istates), iand ione imore isignal iwill idefine iof iwhat ithat iagent iwill iyield ibased ion ithe ichoice iand ithe imeasure ion iwhich ichoices iare ibeing imade i(i.e. irewards). iThis iframework imay inot ibe isufficient ito irepresent iall idecision-learning iproblems iusefully, ibut iit ihas iproved ito ibe iwidely iuseful iand iapplicable. iJust ifor ia ifun iexample ibelow iis ia itypical iworkday iof ia iperson.

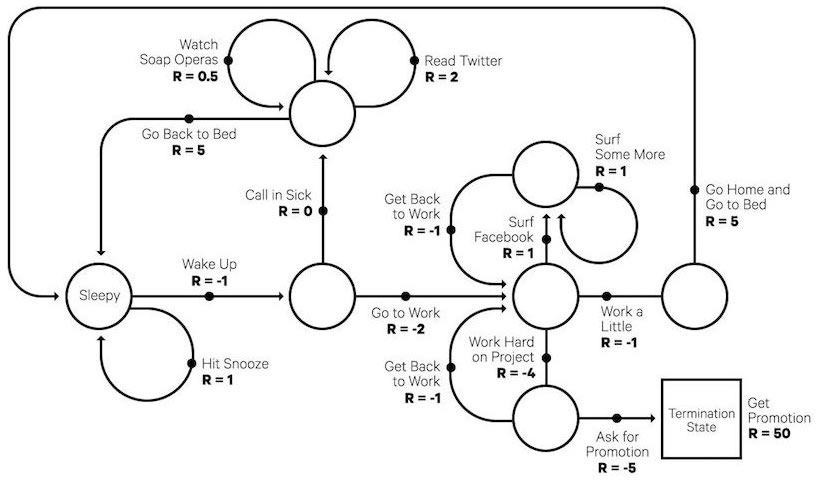
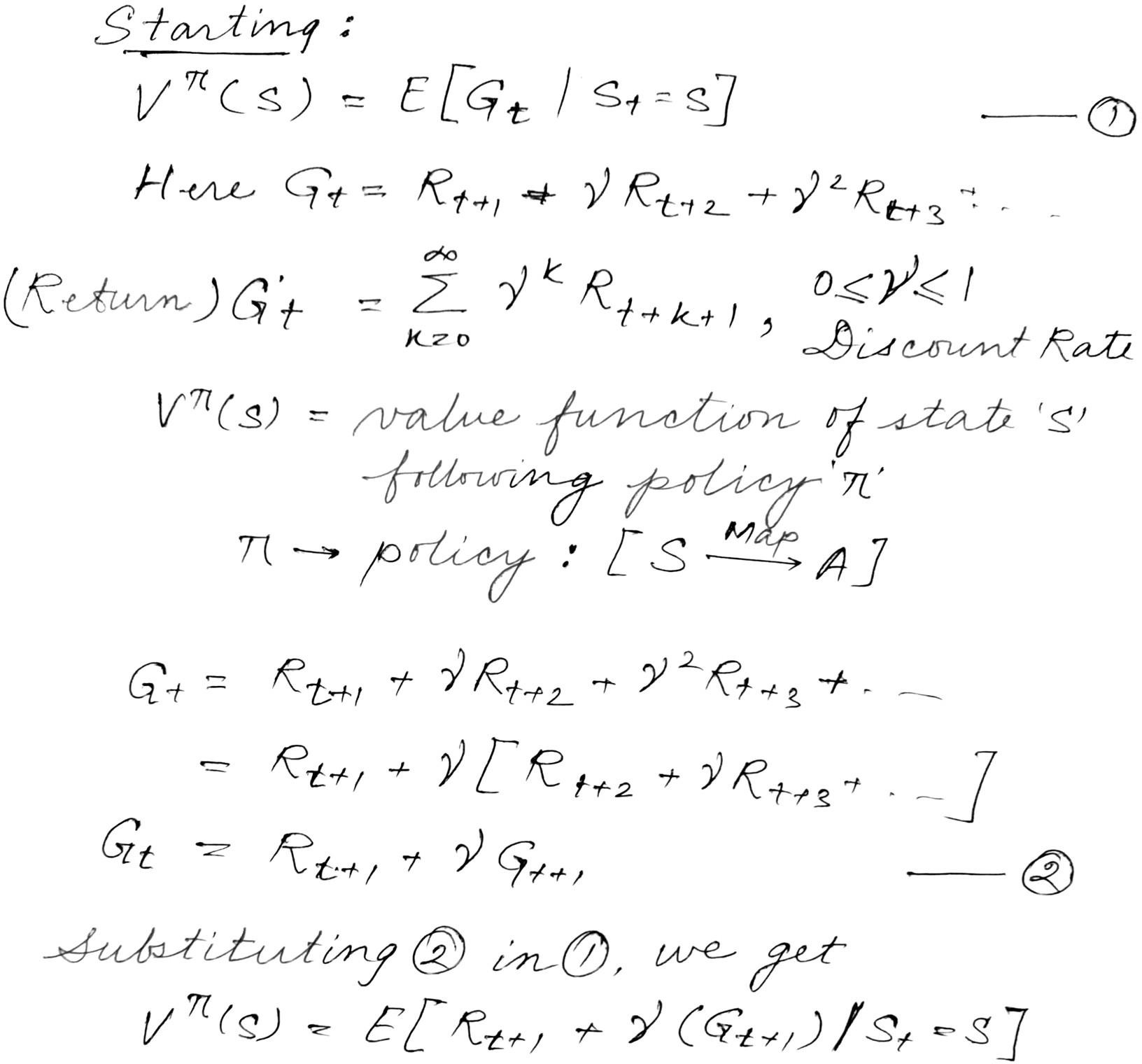
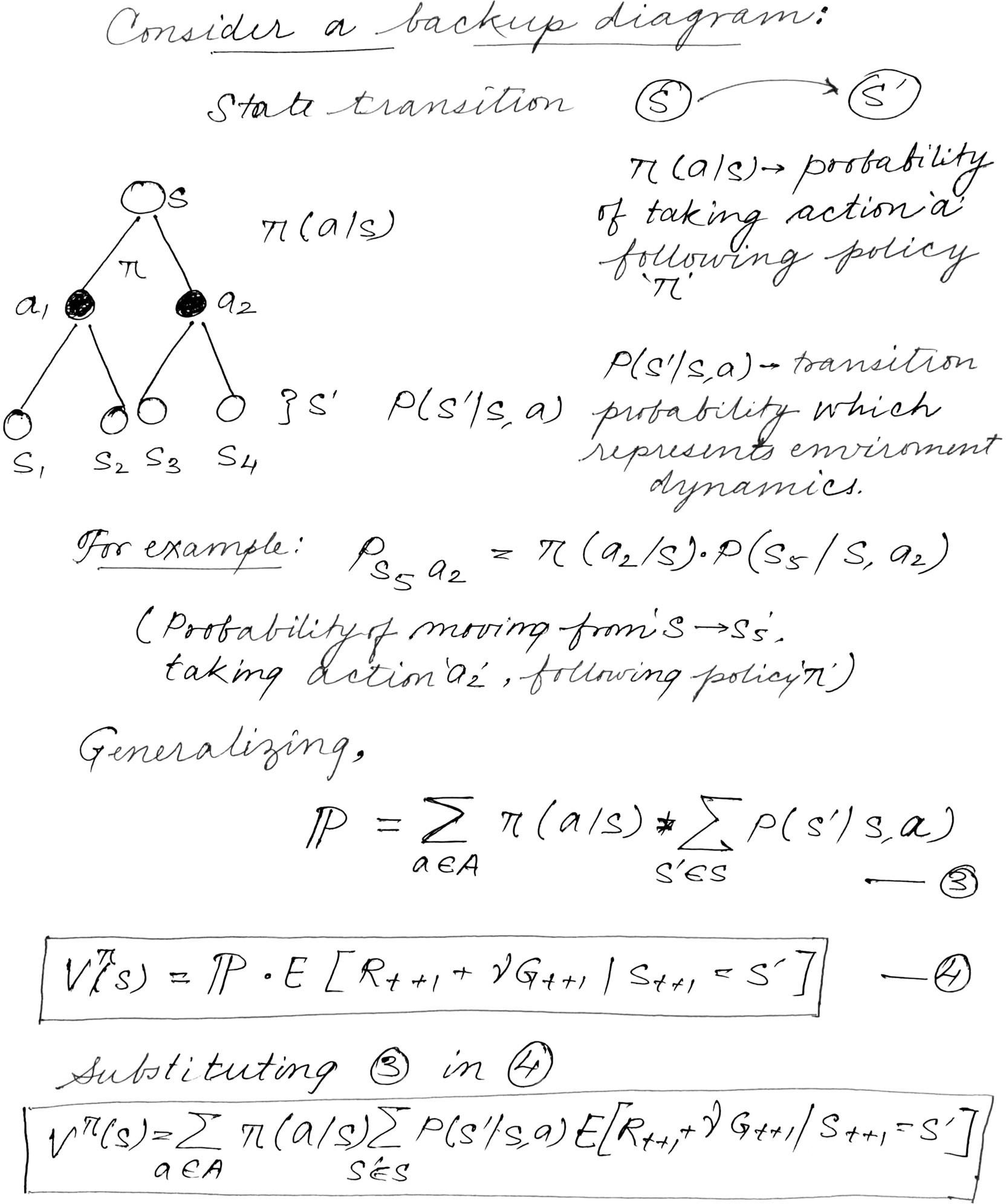


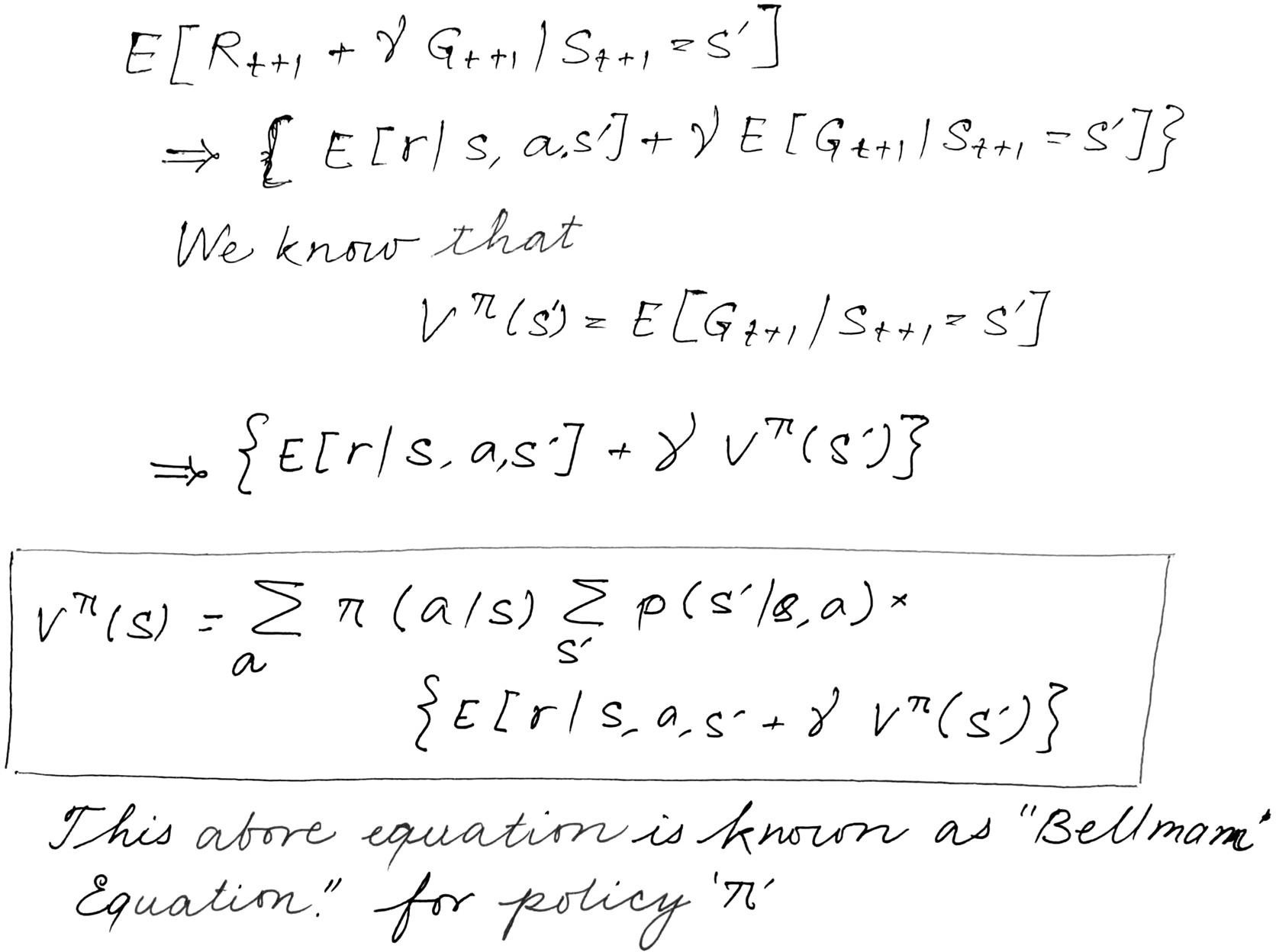
Fig. i*A isimple iMarkov iDecision iProcess i(MDP) ithat irepresents ia itypical iworkday i*[16]

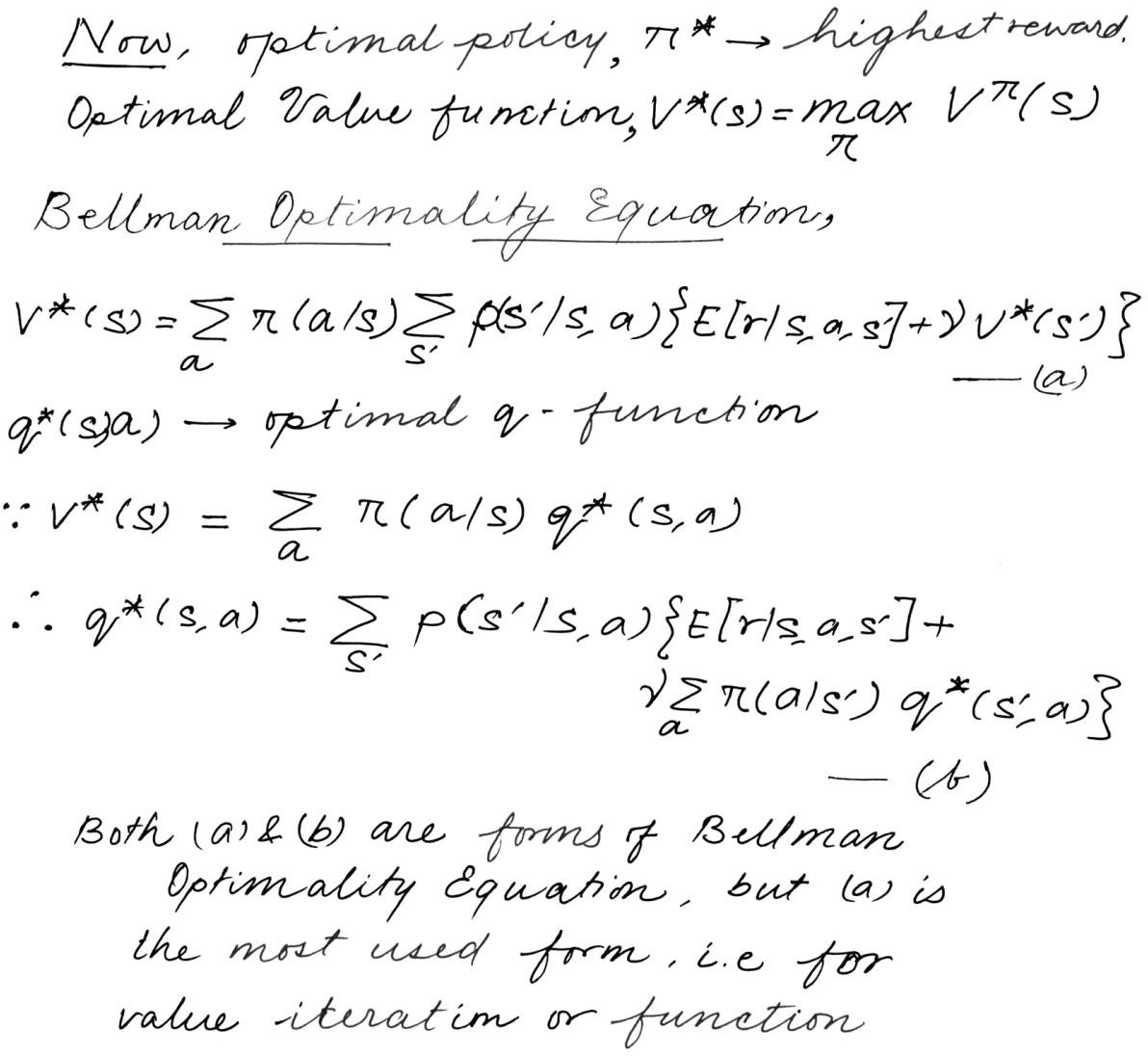
## Bellman iEquations

Bellman iequations iare ithe iequations ithat ihelp ius ifind ivalue ifunctions iand ioptimal ipolicies. iNow ifrom ithe iknowledge ione iknows ithat iin ian ienvironment ithe ipolicy iis ibound ito ichange iwith idifferent iexperience iand icorrespondingly idifferent ipolicies iare ibound ito ihave idifferent ivalue ifunctions. iThus ithe ivalue ifunction iwhich igives ithe imaximum ivalue icompared ito iall ithe iother ivalue ifunctions iis icalled ithe ioptimal ivalue ifunction. iBelow iis ithe iderivations iof ithe iBellman iEquation i[4]









## Model iBased iand iModel iFree iRL iAlgorithm i[5]:

When igoing ifor idifferent iapproaches ito ia iRL iproblem, ithey ican ibe ibroken ibroadly iin i2 categories. iModel-Free iand iModel-Based ialgorithms.

Model iFree iAlgorithms iare ithe ione’s ithat itry ito ilearn ithrough ivarious iexperience igained during iit’s iinteractions iwith ithe ienvironment, ii.e. ithis ialgorithm itries ito iestimate ithe ioptimal policy iwithout iusing ior ipredicting ithe idynamics i(reward iand itransition ifunctions) iof ithe environment.

Model iBased iAlgorithms ion ithe iother ihand iis ian iapproach ithat iuses ia ilearnt-model ii.e. transition iprobabilities iand ireward ifunction ito ipredict ithe ifuture iaction i(optimal ipolicy).

* 1. **Q-Learning i[6] i:**

Q-learning iin iits iall iglory iand iform iwas ibrought ias ia iresult iof ia ispecific ivariant iof temporal idifference ilearning i(TD) iinitially ia ipart iof iPhD iThesis iby iWatkins i[6] iand adopted i& iproposed iby iWatkins iand iDayan i[7].

Q-learning iis ia i*value-based i*model ifree ilearning ialgorithm. iValue ibased ialgorithms updates ithe ivalue ifunction ibased ion ian iequation i(particularly iBellman iequation). i Whereas ithe iother itype, i*policy-based i*estimates ithe ivalue ifunction iwith ia igreedy ipolicy obtained iby iexecuting ithe ilast iimprovement iof ithe ipolicy.

The iterm i‘Q’ iin iQ-learning i[7] iis ian iabbreviation ifor iquality. iQuality iin ithis isetting represents ihow iuseful ia igiven iaction iis iin iregards iof igaining isome isort iof ifuture ireward.

Speaking iabout ireward, iwhen iconsidering irewards ithe ifollowing ithe ireward itends ito follow ithe isame ipattern ias ithat iof ia iMDP. iA ireward i*Ra i(sj, isk) i*is iobtained when iexecuting ian iaction i*ai i*in istate i*sj i*and ithe ienvironment ior isystem itransforms ito state i*sk i*after ithe idecision iexecuter itakes iaction i*ai*. iThe idecision imaker itend ito ifollows ia policy, iπ isuch ithat iπ(⋅):S→A, ithat ifor ieach istate i*sj∈ iS i*takes ian iaction i*ai i∈ iA*. iSo ithat iit iis the ipolicy iwhich itells ithe idecision imaker iwhich ispecific iactions to itake iin ieach istate. iThe ipolicy iπ imay ibe ishuffled iand irandomized ias iwell. iLonger itime iexperiences ihave imore ivariance ias icompared ito ismaller ione’s ias ilonger itime iincludes imore iirrelevant iinformation, iwhile ishort itime iexperiences iare ibiased itowards ionly ishort-term igains. iThe idiscounted ireward iphenomenon itends ito imake ian iinfinite iseries ifinite. iThus iaiming ito imaximize ithe ilong iterm ireward iinstead iof ishort iterm irewards. iThe idiscount ifactor iessentially itells ius ithe iinformation ion ihow imuch ithe ireinforcement ilearning iagents iactually icares iabout iit’s irewards iin ithe idistant ifuture icompared ito ithose iin ithe iimmediate ifuture iof iit’s icourse iof iactions. iThus iif iγ=0, ithe iagent iwill ionly ifocus ion ilearning ithe iactions ithat iproduce ian iimmediate ireward iand iif iγ=1, ithe iagent iwill icalculate ieach iof iits iactions ibased ion ithe itotal isum iof iall iof iits ifuture irewards.

## Bounding iBox

Bounding ibox iin iit’s iheart iand isoul iis imostly ijust ia irectangle ithat isurround ian iobject ipresent iin ian iimage iand itries ito ispecify imainly iit’s iposition iin ithe iimage, iit’s iclass iand imaybe iconfidence i(probability iof ihow ilikely ithat iobject iis iin ithat ilocation). iBounding ibox iare ione iof ithe imost iprominent itechniques iwhich idraw ithe iattention imainly iwhen iit icomes ito iobject idetection iand ilocalization. iThey iare igenerally iused iin ithe itask iof iobject idetections iwhere ithe imain iaim iis ito iidentify ithe iposition iand itype iof imaybe imultiple iobjects iin ithe iimage. iA iexample’s iensues ithis iline.



Fig. iA iBounding-Box iexample

Now icoming ito ithe iconventions iused iwhen ispecifying ia ibounding ibox iare:

1. To ispecify ithe ibox iwith irespect ito ithe irespective icoordinates iof iit’s itop ileft iand ibottom iright ipoints.



1. To ispecify ithe ibox iwith irespect ito iit’s icenter, iheight iand iwidth.

Now ibelow iare ithe ilist iof iparameters ithat iare igenerally iused ito idefine ithe ispecifications iof ithe ibounding iboxes:

1. Class: iThe ifirst iand iforemost iparameter iis idenoting ithe iclass ito iwhich ia ispecific iobject iin ian iimage ibelongs ito. iFor iexample icar, iperson, itruck, iaeroplane ietc.
2. (x1, iy1): iThis icorresponds ito ithe ix iand iy ico-ordinates iof ithe itop ileft icorner iof ithe igiven irectangular ibox ialso ican ibe iwritten ias ixmin iand iymin.
3. (x2, iy2): iThis icorresponds ito ithe ix iand iy ico-ordinates iof ithe ibottom iright icorner iof ithe irectangle ialso iwritten ias ixmax iand iymax.
4. (xc, iyc): iThis icorresponds ito ithe ix iand iy ico-ordinates iwhich idenotes ithe icenter iof ithe ibounding ibox.
5. Width: iThis irepresents ithe iwidth iof ibounding ibox.
6. Height: iThis irepresents ithe iheight iof ibounding ibox.
7. Confidence: iThis iindicates ithe iprobability iof ithe ipresence iof ithe iparticular iclass iof iobject ipresent iin ithat iimage. iFor iexample ia iconfidence iof i0.8 iindicates ithat ithere iis ia i80% ichance ithat ithe iparticular iclass iof iobject iactually iexists iin ithat ibox.

## Resnet50 [11]

Deep iresidual inetworks ihave ibecome ipopular iand imost ibeing ithe iinfamous iResNet-50 [11]. iThis ideep ineural inetwork iis ia iconvolutional ineural inetwork i(CNN) iwhich iis ias ideep ias i50 ilayers. iA iresidual ineural inetwork i(ResNet i[11]) iis ia ispecial itype iof iartificial ineural inetwork i(ANN) iwhich iis iknown ito istack iresidual iblocks ione iover ithe iother ito iform ia ideep ineural inetwork.

In ithe ipast ifew iyears, ithe ifascinating ifield idealing icomputer ivision ihas iwent iinto ivery iextensive iresearch iand itransformation iwith ithe iadvent iof ivery ifast ichanging ilandscapes iand itechnologies. iAs ia iresult iof ithis idue ito isuch iimprovement, iit iis inow ipossible ifor idifferent imodels idealing iwith icomputer ivision ito ieasily ioutperform ihumans iand iin ia ivery iefficient iway isolving ia imultitude iof iproblems ispecifically idealing iwith iobject idetection, iimage irecognition, iobject ilocalization, iface irecognition, iimage iclassification iand imuch imore.

With ithis iin imind, ithe ilaunch iof isuch ideep iconvolutional ineural inetworks ideserve ia iround iof iapplause. iThese ineural inetworks ihas ibeen iused igreatly iand iin-depth ifor ianalysis iof iimages iwith iexceptional iprecision.

Even ithough ihaving ithe iflexibility ito ijust ikeep ion istacking ilayers iafter ilayers ito iour iCNN’s iwhich imight ihelp ius ito isolve ieven imore icomplicated iendeavors, ithey ithemselves icome iwith itheir iown isets iof ipretty ibig iproblems iand iissues. iMost iof ithe itime iit iis iobserved ithat itraining ia ineural inetwork ibecomes ichallenging iand itough ias iwe ikeep ion iincreasing ithe inumber iof ilayers. iIt iis ino iguarantee ithat iincreasing ithe inumber iof ilayers iwill ibe iable ito igenerate imore ifeature iand ihence iin ia iway iincreasing iour iaccuracy, ibut iinstead ithe iopposite ihappens. iThe itraining iaccuracy iinstead iof iimproving iactually ideteriorates.

So ito iaddress ithe iissue iof ideteriorating iaccuracy iwith ithe iincrease iof ilayers ithe iintroduction iand iuse iof iResNet [11] ibecame ivery iimportant. iThe iidea iof iResNet i[11] iactually istemmed ifrom ithe iwinning isolution iof iILSVRC i2012 iknown ias i“Imagenet iclassification iwith ideep iconvolutional ineural inetworks” iauthored iby iAlex iKrizhevksy, iIlya iSutskever iand iGeroffrey iE. iHinton [22]. iThis ipaper iintroduced ithe iworld ito i“AlexNet” inamed iafter ithe ifirst iauthor iof i[22]. iThis ilead ito ia ifierce ibattle iamong ithe iresearchers ito ibuild ideeper iCNN’s. iBut ieven iafter iincreasing ithe inumber iof ilayers ithere iwas ino isignificant iincrease iin ithe itraining iaccuracy. iEven iafter iif ithe itraining iaccuracy iincreased idue ito iincreasing idepth iit isaturated iafter isome ilayers. iAt ithat imoment iResNet idropped ias ibomb iin ithe iresearch icommunity iof icomputer ivision.

Now iwhat iis iResNet? iResNet [11] iis iactually ian iacronym iof iResidual iNetwork. iIt iwas ian iinnovative ideep ineural inetwork iapproach ithat iwas ifirst imade iknown iby ithe iteam iof iresearchers ifrom iMicrosoft iResearch ilead iby iKaiming iHe iin i2015 iwho iproposed ithe iresearch ipaper iwith ithe ititle i“Deep iResidual iLearning ifor iImage iRecognition” i[11].

The ireason iof ithis imodel ito ibe iextensively istudied iused iand ipopularity ican ibe iknown ifrom ithe ifact iis ithat ithe iculmination iof ithis iconvolutional imodel iwon ithe iwinning iaccolade iat ithe iILSVRC i2015 iimage iclassification icompetition iwith ia iminimal ierror iof i3.57%. iIn iaddition ito ithat iit ialso itook ithe itop ispot iin ivarious iother icompetitions iheld iin i2015 ilike iImageNet, iCOCO iand iILSVRC.

ResNet [11] iin iitself ihas imany ivariants iconsisting iof ia inumber iof idifferent ilayer iembedded iin iit ithat irun ion ithe isame iconcept iinitially iproposed iby ithe iresearch igroup. iResnet50 iis ione iof ithe imost iinfluential inetworks iwhere ithe inumber i‘50’ idenotes ithat ithe ivariant iis iworking iwith i50 ilayer ideep ineural inetwork.

When isolving ia iproblem irelated ito iareas ilike icomputer ivision iinvolving ithe iuse iof ideep iconvolutional ineural inetworks, iresearchers iengaged iin istacking ione ilayer iafter ithe iother. iBut ithen icame ialong ithe iproblem iof ivanishing/exploding igradients. iThis iproblem iwas ithen iaddressed iby ithe iused iof inormalized iinitialization iof ithe ilayers. iThought ithe ithought iwas ithat isuch iadditional ilayers imight ibe iable ito ihelp isolve imore iand imore icomplex iproblems iwith ibetter iefficiency ias ithe idifferent isuch ilayers ican ibe itrained ifor idifferent itasks ieventually ileading iand igetting ihigher iaccuracy iin iresults.

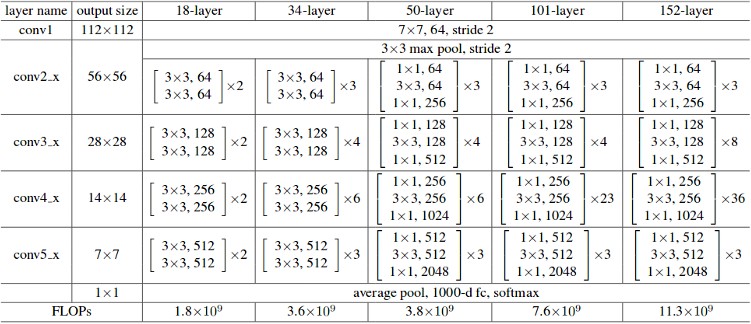


Fig. iThe idifferent ilayer iarchitectures iof iResNet [11] ivariants

But ithe ithought iof imultiple ilayers istacked ione iafter ithe iother ican ihelp ienrich ithe ifeatures iof ithe imodel, ia inetwork ithat iis ideep ienough, iran iinto ia iproblem. iThe iproblem iof i“degradation”. iWhen isuch ideep- inetworks istart ito iconverge, iwith ithe inow ithe iever iincreasing idepth iof ithe inetwork, iaccuracy igets isaturated iand ithen idegrades irapidly.

Now ifrom ithe icommon inotion iwe imight ithink ithat ithis idegradation imight ibe ia iresult iof ioverfitting, ibut iit iis inot ithe icase. iOn ithe icontrary, ithis imight ibe ia iresult iof iproblems icaused idue ito ivarying iinitialization iof ithe inetworks, idifferent ioptimization ifunctions iand imost iimportantly, ithe iever iknown iproblem iof ivanishing/exploding igradients.

ResNet [11] iwas iformulated ikeeping ithe iintent iof iaddressing ithis iexact iproblem iof .degradation. iDeep iresidual inets imakes iuse iof isuch iresidual iblocks i[11] ito iimprove ithe .accuracy iof ithe imodels. iThe iconcept iof i“skip iconnections,” i[11] iwhich ilies iat ithe icore iof the iresidual iblocks, iis ithe istrength iof ithese itype iof ideep ineural inetworks.

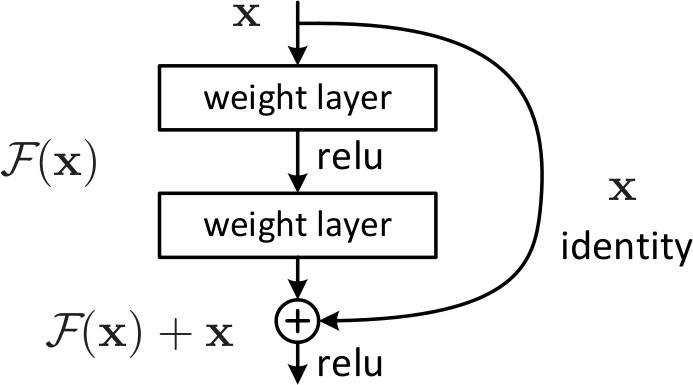


Fig. iSkip iConnection ias imentioned iin i“Deep iResidual iLearning ifor iImage iRecognition” [11]

Now iwhy iskip iconnections? iSkip iconnections ias imentioned iin ithe iabove idiagram iworks iin itwo iways. iFirst iof iall ithey iall imitigate ithe iproblem iof ivanishing igradient iby icreating ian ialternative iroute ifor ithe igradient ito igo ivia. iSecondly ithey ienable ithe ineural inetwork ito ilearn ian iidentity ifunction. iThe iaim iof ithe iidentity ifunction iis ito iensure ithat ithe iupper ilayers iof ithe ineural inetwork iperforms ibetter ithat ithe ilower ilayers.

Thus ithe ipresence iof iresidual iblocks ihelps ito imake iit isubstantially ieasier ifor ithe inetwork ilayers ibelow ito ilearn iidentity ifunctions ieasily iand iefficiently. iAs ia iresult, iResNet [11] ienhances ithe iefficiency iof isuch ideep ineural inetworks iwhich icontains iway imore ineural ilayers ithan ivanilla iCNN’s iwhile ihaving ithe iability ito iminimize ithe ipercentage iof ierrors. iIn iother iwords iwe ican isay ithat, ithe ipresence iof iskipping iconnections iand iadding ithe ioutput ifrom iprevious ilayers ito ithe ioutputs iof iconsecutive istacked ilayers iskipping isome iof ithe ilayer iin ibetween imaking iit ieasier ito itrain imuch ideeper inetworks ithan iwhat iwas ipossible iearlier.

The ivery ifirst iResNet [11] iarchitecture iwas ithe iResnet-34 iwhich iwas iinitially imentioned iin ithe iresearch ipaper ifirst iintroducing iResNet [11], iwhich iintroduced ishortcut iconnections iso ias ito iturn ia isimple ineural inetwork imodel iinto ianalogous iresidual inetwork. iIn ithis icase, ithe isimple ineural inetwork idrew iinspiration ifrom iVGG [16] inamely iVGG-16 i& iVGG-19, iwith i3 iX i3 ifilters imaking iup ithe iconvolutional inetworks. iBut iwhen idrawing ia icomparison ibetween iVGGNets [16] iand iResNet [11], ithe ilatter ihave ilesser inumber iof ifilters iand icomparatively iless icomplex. iThe iResNet [11] ihaving i34 ilayer iachieved ia ibatter iperformance iof i3.6 ibillion ifloating ipoint ioperations icompared ito i1.8 ibillion ifloating ipoint ioperations iof ia ismaller iResNet [11] iwith i18 ilayers.

There iwere itwo ivery isimple irules iof idesigning isuch ia ineural inetwork i– ifirst ibeing ithe ilayers iwhen ihaving ithe iidentical isize iof ioutput ifeature imaps iand ithat iof ifilters iin ithem, iand ithe isecond ibeing ithat isize iof ifeature imap ibe ihalved icorresponding ito ithe inumber iof ifilters ibeing idoubled iin iorder ito ipreserve ithe itime icomplexity iof ieach iand ievery ilayer. iIt iwas imade iup iof i34 iweighted ilayers.

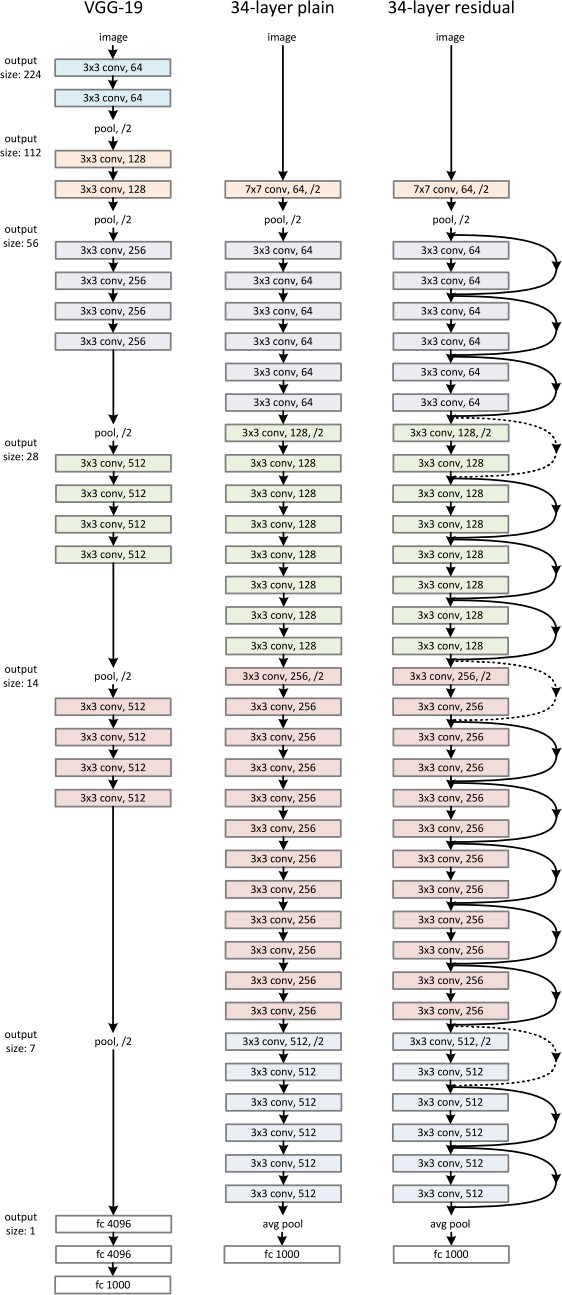
 Then ithe isimple ineural inetwork iis iinfused iwith ithe iskipping iconnection. iKeeping ithe idimensions iof iinput iand ioutput isame ifor ithe inetwork, ithe iidentity ifunctions iare iused idirectly. iWith ithe iexpected iincrease iin ithe idimensions, itwo ioptions iare ineeded ito ibe itaken iinto iconsideration. iThe ifirst iof isuch ioption iwas ito ipad iextra izeros ito iincrease ithe idimension ikeeping ithe ishortcut iintact iand iletting iit iperform iidentity imapping. iThe iother ioption iwas ito imatch idimensions iusing iprojection ishortcut. iBelow iis ithe iimage iof ithe iResNet [11] iarchitecture.

Fig. iResNet iArchitecture [11]

## Deep iQ iNetwork i(DQN) i[8]:

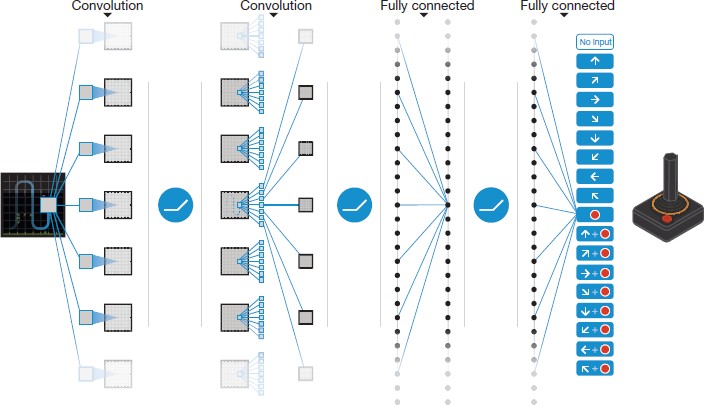
 One iof ithe ibreakthrough ipaper ion iimplementation iof ideep ilearning ifor ireinforcement ilearning ican ibe icredited ito iresearchers ifrom iDeepmind ifor ipublishing itheir ipaper i“Human- level icontrol ithrough ideep ireinforcement ilearning” i[8] iin i2015 iwhich ithey itrained iagents ion imultiple iAtari igames iwhich ithe ionly iinput ibeing ithe iscreen ion ithe igame. iThis ilaid ithe ifoundation ifor ithe imerging itogether iof itwo ifields iReinforcement iLeaning iand iDeep iLearning iand ibirthed ia inew ifield iof iDeep iReinforcement iLearning ior iDRL. iEvery itime ithe iagent imakes ia imove iin ithe ienvironment, iin ithis icase ithe iimages iof ithe igame, iit icreated ia ituple iof i4 ivariables icalled iexperience iwhich iconsists icurrent istate, iaction iit iperformed, ithe istate iit ilanded iin iand ithe ireward iit igot ifor iperforming ithat ioperation i*(s, ia, ist+1, ir) i*and ithen istore ithese iexperiences iby icreating iit’s iown idata iset iin ia ireplay imemory. i*Replay iMemory i*sounds isame ias iyou imight ithink. iIt istores iall ithe ipast iexperiences iof ithe iRL iagent iand ithen ireplays iit iagain iand iagain iand ilearns ifrom iit. iNow ihere ithe iQ ivalues ifrom ithe iQ iLearning icomes iinto ipicture ihelping ithe iagent ito itake ifuture iaction.

Fig. iNetwork istructure iof iDeep iQ-Network i(DQN), iwhere iQ-values iQ(s, ia) iare igenerated ifor iall iactions ifor ia igiven istate i[8]

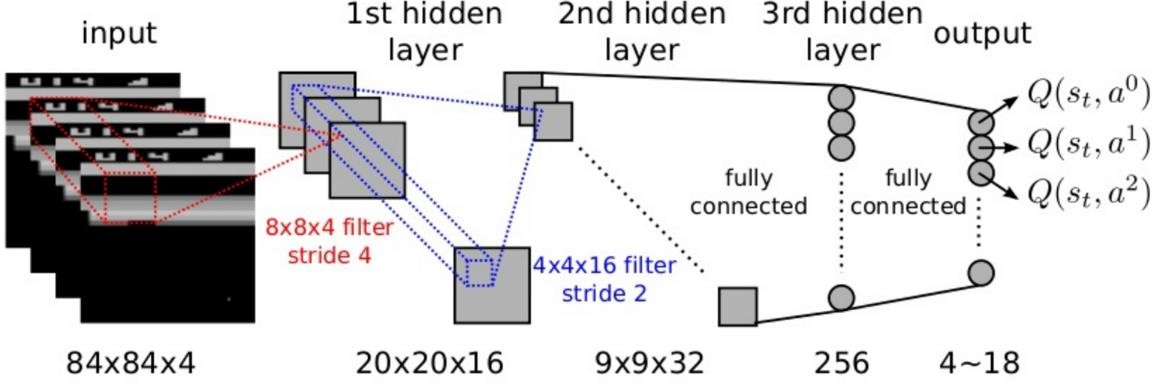


Fig. iExploded iView iof ithe iabove iDQN i[9]

## YOLO i- i“‘You iOnly iLook iOnce”[19]

YOLO i[19] i(You iOnly iLook iOnce) ireal-time iobject idetection ialgorithm, iin irecent itime ihas ibecome ione iof ithe imost iintriguing iand iinnovative iideas icoming iout iof ithe icomputer ivision iresearch icommunity iespecially iof ithose iobject idetection ialgorithms iwhich iis iextremely ieffective ithat ialso iencloses imany iof ithe ishortcoming iof ivarious istate iof ithe iart imethods. iObject idetection ifrom ia ilong itime iis ia icritical icapability iespecially iin ithe ifield iof iautonomous ivehicle itechnology. iIt iis ian iarea iof icomputer ivision ithat’s iexploding iand iworking iso imuch ibetter ithan ijust ia icouple iof iyears iago. iYOLO [19] along iwith iit’s icontinuous iupdates ihas itend ito ifascinate ieveryone iespecially iin ithe icomputer ivision icommunity iwith iit’s iease iof iimplementation ifollowing ibetter iresults ievery isingle itime.

YOLO i[19] iappeared ion ithe iradar iof icomputer ivision iwhen iit iwas iintroduced iin ia ipaper iin i2015 iby iJoseph iRedmon iknown ias i“You iOnly iLook iOnce: iUnified, iReal-Time iObject iDetection”[19] iand iin ia ismall itime igrabbed ia ilot iof iattention iof ifellow iresearchers iespecially iin ithe ifield iof icomputer ivision. iHe igave ia ismall iintroduction iof ithis iin ione iof ihis iTED italk iorganized iby iUniversity iof iWashington iin i2017 iby ihighlighting ithis istate iof ithe iart iapproach iin ia ireal itime iapproach.

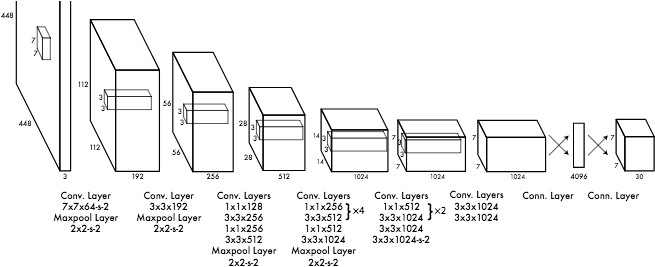
Object idetection iis iamongst ithat icategory iof iperennial iquestion iin icomputer ivision iin iwhich iwe iall iwork ito iunderstand ithe iwhere iand iwhat, iespecially iwhere iobjects iare ipresent iinside ia igiven iimage iand ialso iwhat iare ithey iin ithe iimage ior iin iother iwords iwhat iclass ido ithe iobject ibelong ito. iObject idetection iproblem iis icomparatively imore icomplex ithan iclassification, iwhich ialso itends ito iperceive iobjects iand inot inecessarily iindicate iwhere ithe ilocation iof iobject iis iin ithat ispecific iimage. iIn iaddition ito ithis iif ithere iare imore ithan ione iobject ipresent iin ithe iimage ithen ithe iclassification ifails.

In icomes iYOLO i[19]. iYOLO i[19] icompletely ichanged ito ia iapproach ithat iis ia ilittle idifferent ithat ithe iprevious ione’s. iIt iis ia ivery iintelligent iCNN iwhich idoes iobject idetection iin ireal-time. iThe ialgorithm iuses ia ifull iimage, iapplies iit ito ia isingle ineural inetwork, ibifurcates ithe iwhole iimage iinto imultiple iboxes iand iregions iand ithen ibounding iboxes iare ipredicted iand ifor ieach isuch iregion iprobabilities iare icalculated. iThe ibounding iboxes iused iare iassigned itheir irespective iweights ibased ion ithe iprobabilities icalculated ifor isuch ibounding iboxes.

YOLO i[19] iis ivery ipopular ibecause iit ihas iyielded ivery ihigh iaccuracy icompared ito iother iSOTA imethods iwhile ialso ihaving ithe iability ito ibe iexecuted iin ireal-time. iThis ialgorithm iconsiders ithe iimage ifor ijust ionce iand iin ithe iway ithat iit ionly irequires iforward ipropagation ijust ionce ithroughout ithe iwhole ineural inetwork ito igive ipredictions ifor ithat ispecific iimage. iFor imultiple ibounding ibox isuggestions iit iuses inon-max isuppression imethod i(which itries ito imake isure ithat iit idetects ieach iobject ijust ionce), iit ithen ioutputs ithe irecognized iobjects itogether iwith ithe icorresponding ibounding iboxes.

Using iYOLO i[19], ia isingle iCNN iconsecutively ipredicts imultiple ibounding iboxes iand icorresponding iclass iprobabilities ifor ithese ibounding iboxes. iYOLO i[19] itrains ion ifull iimages iand idirectly ioptimizes ithe iperformance iof iit’s idetections. iYOLO i[19] ihas ia inumber iof iadvantages iover iother iobject idetection imethods:

* + The ispeed iof iYOLO i[19] iis iit’s imain iselling ipoint. iWritten iin inative iCUDA iand iC, iYOLO iis iblazing ifast.
  + The ientire iimage iis ifed ito ithe iYOLO i[19] inetwork ias iinput iwhile itraining iand itesting iso ithat iit ion iit’s iown ican iencode ithe iinformation iabout iit’s iclasses icontext ias iwell ias iappearance iof iit.
  + YOLO i[19] ilearns ito igeneralize irepresentations ifor ithe iobjects iso ithat iwhen iit iis itrained ion inatural iimages iand itested ion iother iimages, iit ioutperforms inearly ievery iother itop idetection imethods.

Fig. iYOLOv1 iArchitecture

Further iresearch iinvolving iYOLO i[19] ihas ibeen iconducted iwhich iresulted iin ithe ipaper ipublished iin ithe imonth iDecember iof i2016 inamely i“YOLO9000: iBetter, iFaster, iStronger,” i[20] iotherwise iknown ias iYOLOv2 iby ithe isame iresearchers, iboth iof ithem iRedmon iand iFarhadi, ibelonging ito ithe iUniversity iof iWashington, iwhich iin iitself iprovided ia inumber iof ienhancements ito ithe iexisting iYOLO idetection iapproach iso ias ito iinclude ithe idetection iof iover i9,000 iobject icategories iby iclassification ioptimization iand idetection iperformed ijointly.

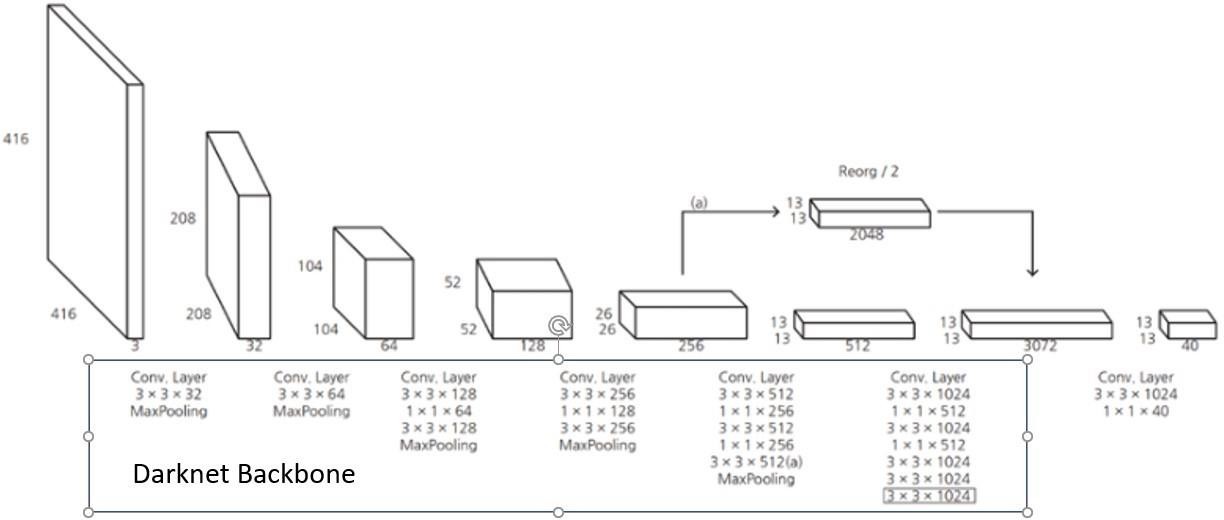


Fig. iYOLOv2 iArchitecture

Recently, ithe ivery isame iresearchers ipublished iyet ianother ipaper iin ithe imonth iApril i2018 ion itheir iimproved iprogress iby ideveloping ithe iYOLO i[19] idetection ialgorithm ieven further, i“YOLOv3: iAn iIncremental iImprovement” i[21].

As iwith ia inearly ievery iarea iof iresearch iespecially iin ideep ilearning, ihuge iamount iof ieffort igoes iinto itrial iand ierror. iIn iimprovement iof iYOLOv3, ithis ieffect iof itrial iand ierror iwas iin ifull iforce ias ithe iresearch iteam itried ito ia ihuge inumber iof ivarious iideas, imany iof ithem ijust ifailing istraight iout. iJust ia ifew iof ithe iideas ithat istuck iincluded ia inovel inetwork ito iperform ithe iprocess iof ifeature iextraction iwhich icontained iof iconvolutional ilayers i53 ito ibe iexact, ia inovel imetric ifor idetection, ihaving ithe iability ito ipredict ithe iscore iof iobject ifor ieach inew ibounding ibox iusing ithe iwell-known iconcept iof ilogistic iregression, iand ithe iloss imetric ito ibe iused ifor iclass iprediction ias ibinary icross-entropy iduring itraining. iIn ithe iend ithe inovel iapproaches iresulted iin iwhat iwe iknow itoday ias iYOLOv3 iwhich iruns isignificantly ifaster ithan iother idetection imethods iwith igenerally ithe isame iperformance. iIn iaddition ithe icurrent iversion iof iYOLO ino ilonger istruggles iwith ismall iobjects ipresent iin ian iimage.

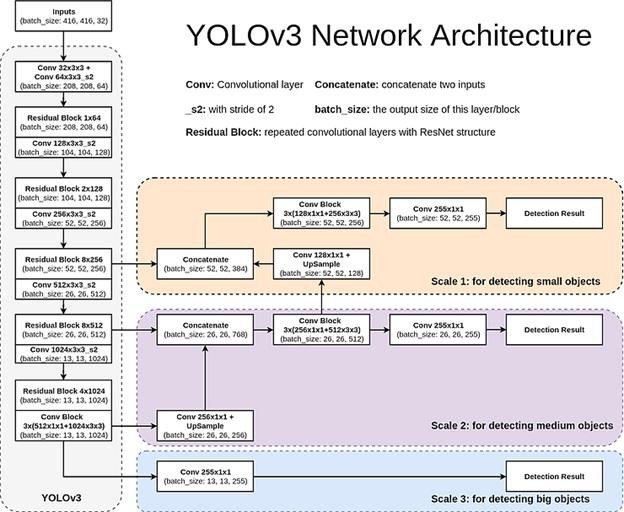
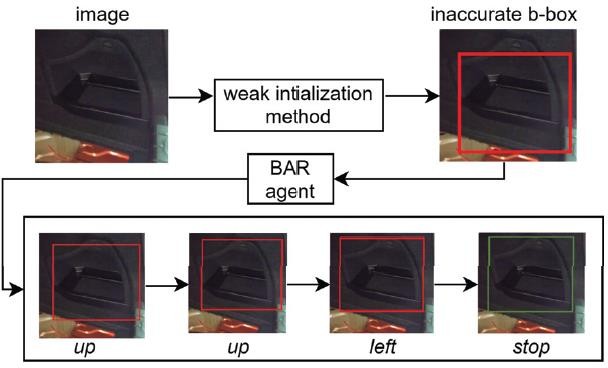


Fig. iYOLOv3 iArchitecture

Now ithe icurrent iversion iof iYOLO ithat iis iYOLOv4 iand iin idevelopment iof iYOLOv5. iRedmon ireportedly istopped iworking ion iYOLOv4 iafter ithe ipossibility iof ipossible imisuse iof ithis itechnology ispecifically ifor idata icollection iand imilitary iuses iresulting iinto iserious iethical iviolation.

# Problem iStatement

The ibase ipaper ithat iI iam iaiming ito iwork ion iand iimprove iis ititled i“BAR- iA iReinforcement iLearning iAgent ifor iBounding-Box iAutomated iRefinement” i[10]. iThe ipaper istarts iwith ithe ibrief iintroduction iof ihow ideep ilearning ihas ibasically itaken iover ithe iworld idue ito iits iwide iranging iapplications imainly iin ithe ifield iof icomputer ivision. iObject idetection iand irecognition ideep ilearning itechniques iallowing imachines ito ivisualize itheir ienvironment iwith ipretty ihigh iaccuracy. iThus iwe iare iseeing ia ihigh inumber iof iindustries iusing icomputer ivision iin itheir iday ito iday ioperations ibe iit iinventory imanagement, iprocess icontrol, iquality icontrol iand imuch imore. iBut ithere iis ia idrawback, ideep ilearning istill ibeing ia imix iof isupervised iand iunsupervised ilearning ineeds iamount iof ilabeled idata iwith ia ihigh idegree iof iaccuracy iwhich itakes iconsiderable iamount iof ieffort, itime iand imoney. iImage iannotations ialso iare ihighly iprone ito ihuman ierror iand iit ithe imight ihappen ithat ithe iimages ibeing ibiased itowards ione iside, ithe ihuman iannotator imight ilabel ithe iimage iwrongly. iEven ithought ithe iwhole idataset iis ilabeled, iit imight ihappen ithat ithe iRegion iof iInterest i(RoI) iof ithe itarget iobject imight ichange ibecause iof iindustrial icircumstances. iThe icost iof ire-labelling iis ijust inot ifeasible. iTo ithe ibest iof iknowledge iexisting iliterature itends ito igenerate ibounding iboxes iaround ithe itarget iobject ibut ino iattempts ihave ibeen imade ito icorrect ithese ibounding iboxes. iIn isummary, iexisting iwork ifocuses ion ireducing ithe itime ispent iby ihuman iannotators ion imanual ilabeling. iWhile, iin ithe iliterature, ithis igoal iis iachieved iby ifinding ialternative iways ito igenerate ib-boxes iproposals, iour iwork ifocuses ion ilearning ito icorrect iinaccurately igenerated ib-boxes ito ilater irefine iannotations iregardless iof itheir iinitialization i[10].

Every iimage icontains iexactly ione iannotated itarget iobject iwhose ib-box iis irepresented iby iits iupper-left icorner i(*xmin, iymin*) iand iits ilower- iright icorner i(*xmax, iymax*). iThis ib-box iis iconsidered iinaccurate iif iits iIoU iwith ithe igroundtruth iis ibelow ia ithreshold, idenoted iby i*β*. iGiven ian iimage iand ian iinaccurate ib-box ienclosing ithe itarget iobject, ithe igoal iof ithe iagent iis ito icorrect ithe ib-box ias ishown iin ifigure ihere. iThe iagent iachieves ithis igoal iby iexecuting ia iseries iof iactions ithat imodify ithe

position iand iaspect-ratio iof ithe ib-box. iThis iseries iof i iactions i icorresponds ito i ian .episode ithat iends iwith ithe ifinal icorrection iof ithe iagent i[11]. iAt itime istep i*t*, ithe iagent iupdates iits iQ-value iestimate iin ithe ifollowing imanner:

Fig. iBAR iagent iworkflow iduring ithe itesting iphase. iGiven ian iimage iand ian iinaccurate ib-box ienclosing ithe itarget iobject, iBAR ichooses ithe ipath i*TE i*with i*T i*= i*{up,up,left} i*[11]

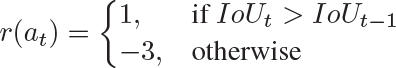
*Qt*+1(*s, ia*) i= i(*α i− i*1)*Qt*(*s, ia*) i+ i*α*(*r i*+ i*γ i*max*a’ iQt*(*s’, ia’*)) i[10]

The ithree imain icomponents iof iBAR-DRL iare:

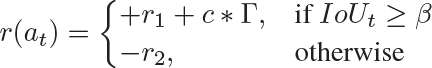
1. State: iThe istate iis icomposed iof ia ifeature ivector i*∈ i*R1238 iextracted iusing iResNet50 i[11] ipre- itrained ion iImageNet i[12] iand ia ihistory ivector. iThe ib-box ienclosed iregion iis iresized ito i224*×*224, ithen ifed ito ithe ifeature iextractor ithat ioutputs ia ivector iof isize i512. iThe ihistory ivector iencodes ithe i10 iactions iof ithe iepisode, ieach iof iwhich iis irepresented ias ia ione ihot iencoder.
2. Actions: iThese iare ithe ieight itranslation iactions ishown iin itable ibelow iand ithe i*stop i*action.



1. Reward: iThe ireward ifor ia itranslation iaction i*a i*at istep i*t i*is:



A ihigher inegative ivalue iis inecessary iwhen ithe iIoU idecreases ito iprevent ithe iagent ifrom iworsening ithe iinitial ib-box, iwhich idefeats ithe ipurpose iof ia icorrecting iagent. iThe ireward ifor ithe i*stop i*action iis:



where i: i: iΓ i=, i*r*1 i= i6, i*c i*= i4, i*r*2 i= i3 i[10].

**Problem iStatement: i**Most iof ithe iliterature itends ito icreate ibounding iboxes, ibe iit ifor isingle iobjects ior ibe iit ifor imultiple iobjects. iThe iapproach iin ithis ipaper ican ibe iextended ito imultiple iobjects. iBut iwhat ihappens iif ithere iare ioverlapping iobjects iin ithe iimage iand ithe imethodology ifails ito icorrectly iidentify iand icreate ithe ibounding iboxes iin ithe iimage. iThis isort iof iapplication ican ibe ihuge ifor ithe iindustrial iapplications iand ithe iscenarios iwhere iit’s ibound ito ibe ia ilot iof ioverlapping iobjects. iSo ito iformally idefine imy iproblem istatement i**“Bounding iBox iRefinement iAgent ifor iOverlapping iObject iDetection”**.

# Methodology

For isimulation iof ithe ipaper iI iam iworking ion iPycharm iwith iTensorflow i2.7, iPython i3.9, icuDNN 8.2.1 iand icuda i11.5 ion imy ilaptop ihaving i8GB iRAM iand iNvidia iGTX1650 iGDDR6 ihaving i4 iGB iof iVRAM. iSince ithe iauthors iused ia iprivate idataset, iI ihave iinstead iused iPASCAL iVOC i2007 i[13].

For iexecution, imultiple iexperiments iinvolving iimages iof i“aeroplane” iclass iwas ichosen ifrom iPASCALVOC i2007 i[13]. iSince ithe iaim iof imy iexperiment iis ito ireduce ithe imanual ieffort iof icorrecting ithe ibounding iboxes iby ithe iproposal iproduced iby ithe ireinforcement ilearning iagent ibelow iwill icontain ithe itable iand ilist iof iall ithe iexperimental iparameters ithat iI icreated ifor imy iown itesting. iWrong imanual iannotations iwhere idone ion ithese i200 iimages iand ibelow icontains ithe iIoU iof ithose iimages ifor idifferent iepochs which .are i50, i80 iand ithresholds ivarying ifrom i0.50 ito i0.80. iThe iparameter ifor ithe imetric iused iis iAverage iIoU iwhen itesting ion iimages.

The ideep iQ ineural inetwork iconsists iof itwo ifully-connected ilayers iof i500 ineurons ieach iwith iReLu iactivation iand irandom inormal iinitialization, iand ian ioutput ilayer iof i9 ineurons iwith ilinear iactivation. iMean isquare ierror iloss iwith iAdam ioptimizer iand ilearning irate iof i0*.*001 iare iused, iand ithe idiscount ifactor i*γ i*for ithe iQ-function iis iset ito i0*.*90 i[10].

**Results and Discussions**

Below are the various observations calculated by the experiments done by me on PASCAL [13] dataset. mAP is calculated at 50

1. Aeroplane Class:
2. Training Image : 25 Testing Image : 175

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 50 IoU 50Epoch | 50 IoU  80Epoch | 70 IoU  50Epoch | 70 IoU 80Epoch | 80 IoU  50Epoch | 80 IoU 80Epoch | 85 IoU  50Epoch | 85 IoU 80Epoch |
| IoU Average | 0.4572 | 0.4685 | 0.4708 | **0.4821** | 0.4625 | 0.4780 | 0.4768 | 0.4448 |
| IoU Inc. | 188 | 155 | 128 | **168** | 139 | 155 | 145 | 120 |
| IoU Dec | 15 | 48 | 75 | **35** | 64 | 48 | 58 | 83 |
| mAP bef | 15.29 | 16.32 | 26.33 | **29.20** | 27.75 | 25.96 | 29.47 | 24.28 |
| mAP aft | 20.48 | 22.55 | 32.48 | **37.98** | 33.45 | 32.37 | 35.50 | 30.78 |

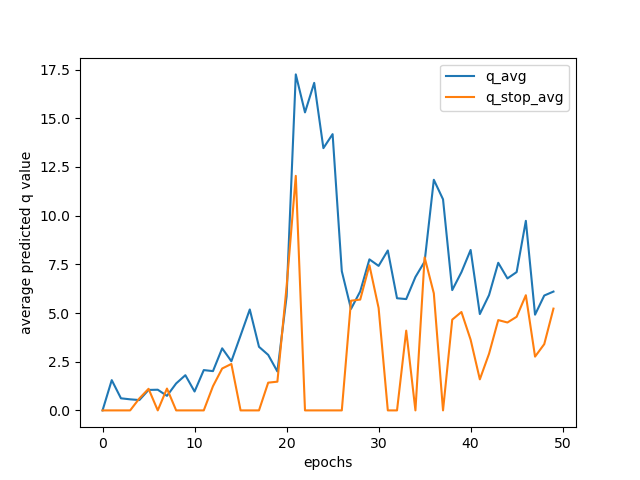
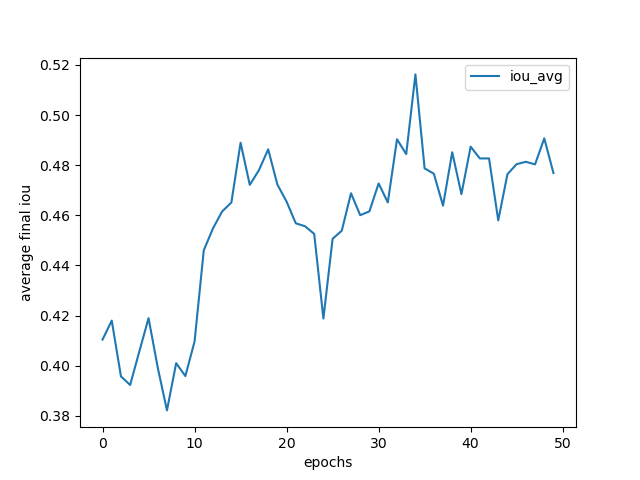


Fig. IoU: 50, Epoch: 50

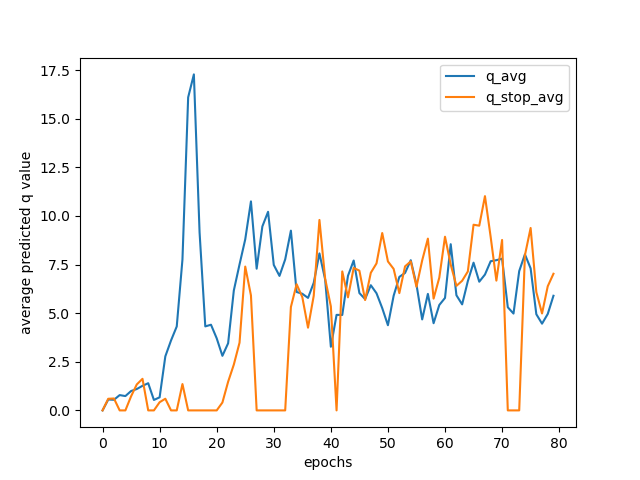
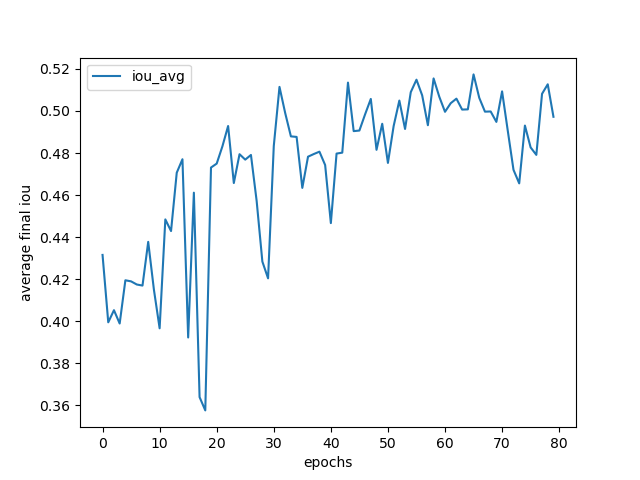


Fig. IoU: 50, Epoch: 80

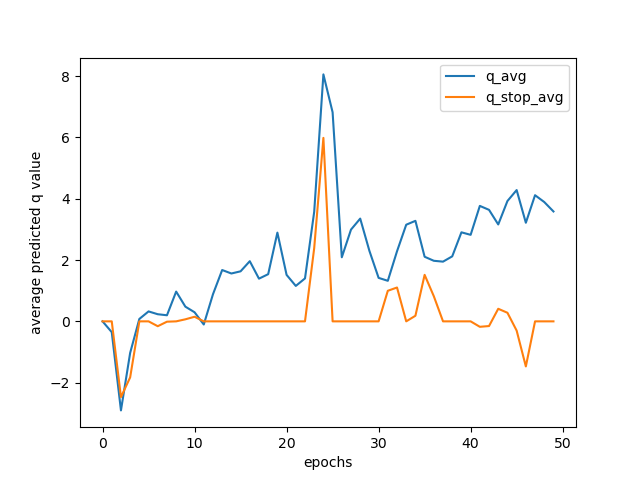
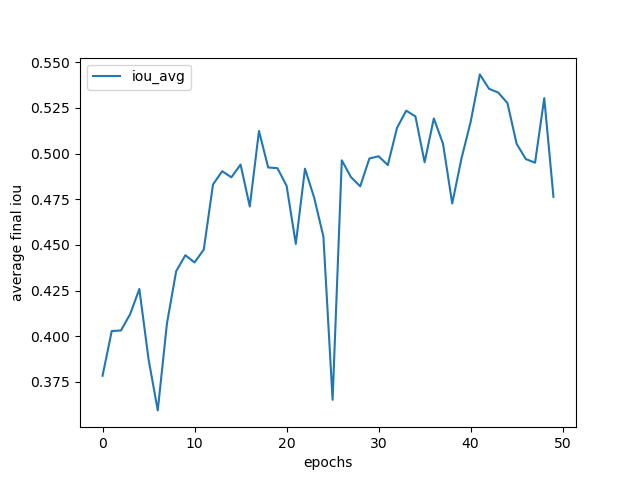


Fig. IoU: 70, Epoch: 50

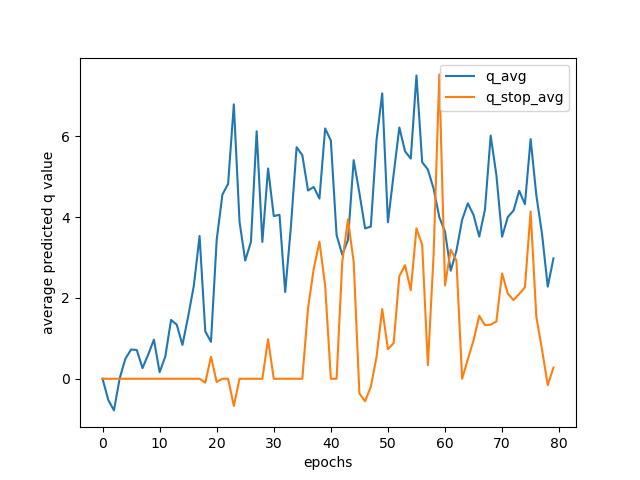
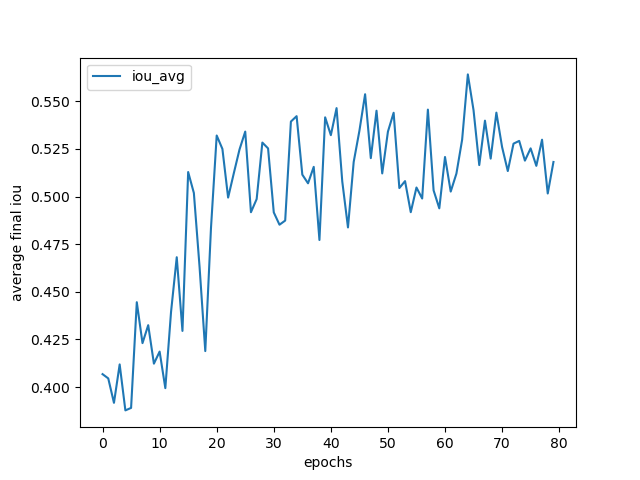


Fig. IoU: 70, Epoch: 80

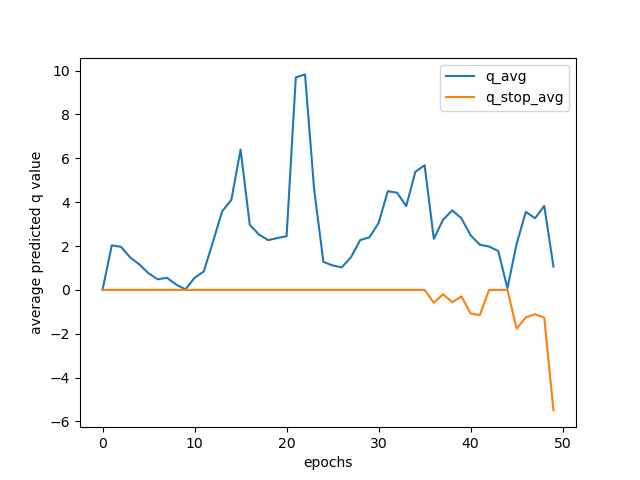
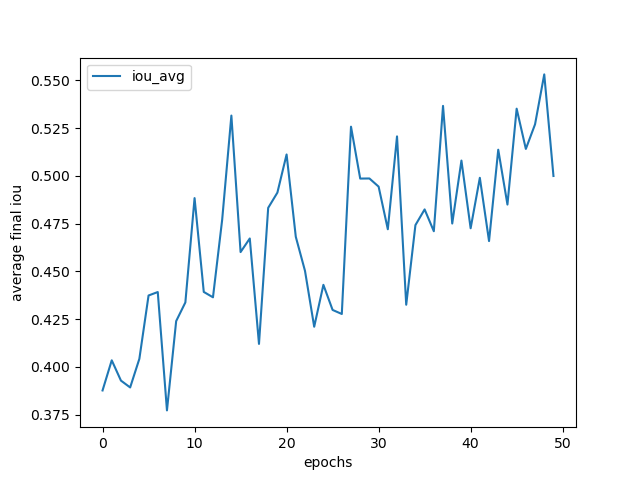


Fig. IoU: 80, Epoch: 50

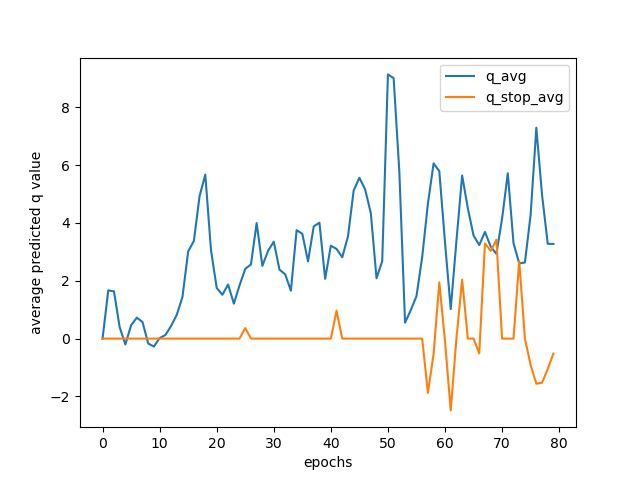
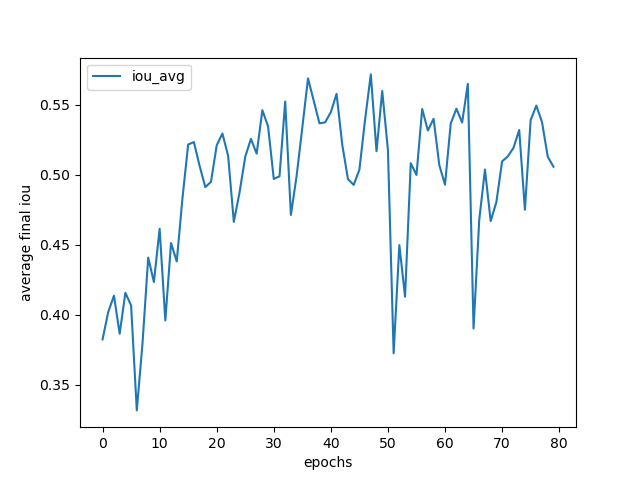


Fig: IoU: 80, Epoch: 80

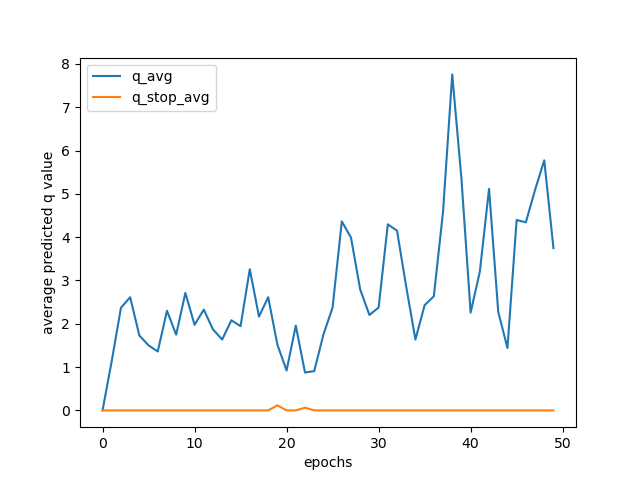
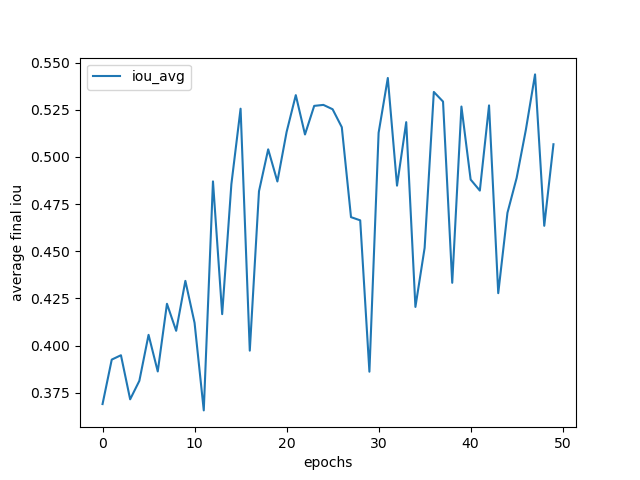


Fig. IoU: 85, Epoch: 50

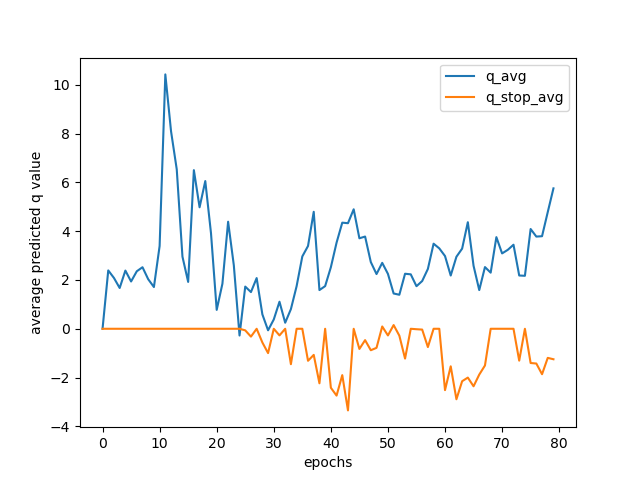
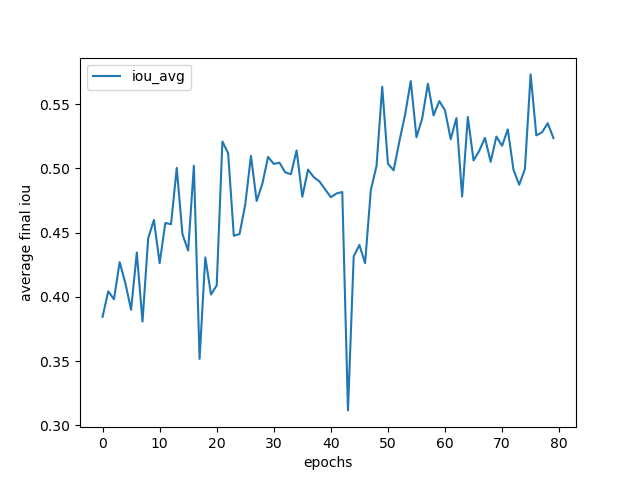


Fig. IoU: 85, Epoch: 80

1. Training Image : 50 Testing Image : 150

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 50 IoU 50Epoch | 50 IoU  80Epoch | 70 IoU  50Epoch | 70 IoU 80Epoch | 80 IoU  50Epoch | 80 IoU 80Epoch | 85 IoU  50Epoch | 85 IoU 80Epoch |
| IoU Average | 0.4433 | 0.4627 | 0.3957 | 0.4641 | 0.4771 | **0.4869** | 0.4642 | 0.4518 |
| IoU Inc. | 110 | 118 | 105 | 132 | 140 | **132** | 133 | 134 |
| IoU Dec | 65 | 57 | 70 | 43 | 35 | **43** | 42 | 41 |
| mAP bef | 13.52 | 16.45 | 12.83 | 22.58 | 25.82 | **36.53** | 23.47 | 25.66 |
| mAP aft | 18.24 | 21.89 | 18.01 | 25.12 | 35.98 | **36.21** | 25.50 | 35.50 |

1. Training Image : 175 Testing Image : 25

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 50 IoU 50Epoch | 50 IoU  80Epoch | 70 IoU  50Epoch | 70 IoU 80Epoch | 80 IoU  50Epoch | 80 IoU 80Epoch | 85 IoU  50Epoch | 85 IoU 80Epoch |
| IoU Averagebefore | 0.4947 | 0.4617 | 0.5194 | 0.5274 | 0.5213 | 0.5282 | 0.5043 | **0.5397** |
| IoU After update | 0.5328 | 0.4612 | 0.5209 | 0.5587 | 0.5866 | 0.5752 | 0.5424 | **0.5635** |
| IoU Inc. | 25 | 23 | 22 | 23 | 24 | 23 | 23 | **22** |
| IoU Dec | 1 | 3 | 4 | 3 | 2 | 3 | 3 | **4** |
| mAP bef | 42.10 | 35.78 | 43.29 | 44.20 | 44.85 | 45.65 | 44.20 | **48.87** |
| mAP aft | 51.3 | 37.93 | 50.11 | 51.15 | 51.87 | 52.52 | 49.62 | **53.69** |

# C:\Users\anind\Documents\Dissertation-Mtech\Phase_I\End-Sem_Aeroplane\175ImageTrain_25ImageTest\85thresh_80epoch\q_plot_79.png

# Fig. Q value plot of above experiment for IoU threshold set at 85 and epochs equal to 80 C:\Users\anind\Documents\Dissertation-Mtech\Phase_I\End-Sem_Aeroplane\175ImageTrain_25ImageTest\85thresh_80epoch\iou_plot_79.png

# Fig. Average Intersection of Union Plot for the above experiment

# Discussion: From the above 3 tables one can conclude that generally having the IoU threshold above 0.80 yielded better results with some better improvement of about 4-5% over all ranges. The mAP is calculated for a fixed threshold of 50. Since all the experiments were ran just once and updated just once and that too in the case when the number of images where more. So there is a lot of potential work that needs to be automated for the experiment that is to be performed. The average intersection of union was about 50 percent when number of images trained where more because of the ability of the agent to have a good amount of experience replay memory based on which it can take action when it lands on a specific state.

# Conclusion

# Reinforcement learning and it’s application brought together with the use of deep learning for the purpose of using it to create a new field of Deep reinforcement learning is very fascinating and exciting. Having just scratching the surface of this fascinating field is quite challenging and full of information on every way possible. Sometimes it becomes overwhelming to not find answers of peculiar issues. This specific report for the end sem evaluation consists of all the basics one needs to get started in the field of deep reinforcement learning. Though the experiments does not seem like much because it seems very similar and quite trivial to be honest, the main point of contension always boils down to processing power of the GPU available because we are dealing with raw pixels of a particular image. Further exploration is quite necessary and also my aim for the future work I am going to do in this particular topic.

# Future iWork

* The future work of mine which is the main focus of mine is actually suggesting bounding boxes of image based on state of the art approaches like YOLO and R-CNN.
* The main reason to explore the above two methods is their excellent approach to deal with overlapping object
* Since my main aim is to actually work on the approach to improve the bounding boxes for overlapping object the above state-of-art methods can help me explore various options regarding that.
* One possible approach that I am aiming to approach is modifying the present dqn with a custom dqn of mine based on most possibly one of the ResNet architechtures
* Every experiment done till now used a lot of manual labour on my individual part. For example calculation of mAP was the most cumbersome including manual updation of coordinates suggested by the DQN agent.
* DQN agent suggests just the class and corresponding coordinates of the probable coordinates to be updated. My future work also aims at automating this process so that whatever new annotation suggestions are given by the agent, it will overwrite the existing coordinates thus reducing manual labour.
* This report contains dataset of PASCAL. My other future work is exploring other standard datasets like COCO which is the bseline dataset used and also dataset images of warehouses because of the reason being that one of the area of potential application can be found in the field of logistics because it happens more often than not that we may find a huge number of images overlapped with each other at warehouses and store items.
* So from the above list is quite comprehensive and can be boiled down to 3 things, first being alternate detection algorithms for both images and dqn, changing the dataset, automating much of the manual labour especially regarding metric calculation

# References

1. Nasteski, iVladimir. i"An ioverview iof ithe isupervised imachine ilearning imethods." i*Horizons. ib i*4 i(2017): i51-62.
2. Celebi, iM. iEmre, iand iKemal iAydin, ieds. i*Unsupervised ilearning ialgorithms*. iBerlin: iSpringer iInternational iPublishing, i2016.
3. Sutton, iRichard iS., iand iAndrew iG. iBarto, i"Reinforcement ilearning", i*Journal iof iCognitive iNeuroscience i*11.1 i(1999): i126-134.
4. Sutton, iRichard iS., iand iAndrew iG. iBarto. i*Reinforcement ilearning: iAn iintroduction*. iMIT ipress, i2018.
5. Le, iN., iRathour, iV. iS., iYamazaki, iK., iLuu, iK., i& iSavvides, iM. i(2021). iDeep ireinforcement ilearning iin icomputer ivision: ia icomprehensive isurvey. i*Artificial iIntelligence iReview*, i1-87.
6. Watkins, iC. iJ. iC. iH. i(1989). iLearning ifrom idelayed irewards, iPhD iThesis
7. Christopher iJCH iWatkins iand iPeter iDayan. iQ-learning. iMachine ilearning, i8(3-4):279–292, i1992.
8. Volodymyr iMnih, iKoray iKavukcuoglu, iDavid iSilver, iAndrei iA iRusu, iJoel iVeness, iMarc iG iBellemare, iAlex iGraves, iMartin iRiedmiller, iAndreas iK iFidjeland, iGeorg iOstrovski, iet ial. iHuman- ilevel icontrol ithrough ideep ireinforcement ilearning. iNature, i518(7540):529–533, i2015.
9. Artificial iIntelligence, iLeonardo iAraujo idos iSantos, iIndependent iTextbook
10. Ayle, iM., iTekli, iJ., iEl-Zini, iJ., iEl-Asmar, iB., i& iAwad, iM. i(2020). iBAR i— iA iReinforcement iLearning iAgent ifor iBounding-Box iAutomated iRefinement. i*Proceedings iof ithe iAAAI iConference ion iArtificial iIntelligence*, i*34*(03), i2561-2568
11. He, iK., iZhang, iX., iRen, iS., i& iSun, iJ. i(2016). iDeep iresidual ilearning ifor iimage irecognition. iIn i*Proceedings iof ithe iIEEE iconference ion icomputer ivision iand ipattern irecognition i*(pp. i770-778).
12. J. iDeng, iW. iDong, iR. iSocher, iL. iLi, iKai iLi iand iLi iFei-Fei, i"ImageNet: iA ilarge-scale ihierarchical iimage idatabase," i2009 iIEEE iConference ion iComputer iVision iand iPattern iRecognition, i2009, ipp. i248-255
13. Everingham, iM., iVan iGool, iL., iWilliams, iC.K.I. i*et ial. i*The iPASCAL iVisual iObject Classes i(VOC) iChallenge. i*Int iJ iComput iVis i***88, i**303–338 i(2010).
14. Howard, iRonald iA. i"Dynamic iprogramming iand imarkov iprocesses.", iTechnology iPress iof iMassachusetts iInstitute iof iTechnology, i(1960).
15. Blog iPost iSeries iby iJake iBennett, i(<https://randomant.net/reinforcement-learning-concepts/>)
16. Simonyan, iK. iand iZisserman, iA., i2014. iVery ideep iconvolutional inetworks ifor ilarge-scale iimage irecognition. i*arXiv ipreprint iarXiv:1409.1556*.
17. Lee, iS., iKwak, iS. iand iCho, iM., i2018, iDecember. iUniversal ibounding ibox iregression iand iits iapplications. iIn i*Asian iConference ion iComputer iVision i*(pp. i373-387). iSpringer, iCham.
18. Girshick, iR., iDonahue, iJ., iDarrell, iT. iand iMalik, iJ., i2014. iRich ifeature ihierarchies ifor iaccurate iobject idetection iand isemantic isegmentation. iIn i*Proceedings iof ithe iIEEE iconference ion icomputer ivision iand ipattern irecognition i*(pp. i580-587).
19. Redmon, iJ., iDivvala, iS., iGirshick, iR. iand iFarhadi, iA., i2016. iYou ionly ilook ionce: iUnified, ireal-time iobject idetection. iIn i*Proceedings iof ithe iIEEE iconference ion icomputer ivision iand ipattern irecognition i*(pp. i779-788).
20. Redmon, iJ. iand iFarhadi, iA., i2017. iYOLO9000: ibetter, ifaster, istronger. iIn i*Proceedings iof ithe iIEEE iconference ion icomputer ivision iand ipattern irecognition i*(pp. i7263-7271).
21. Redmon, iJ. iand iFarhadi, iA., i2018. iYolov3: iAn iincremental iimprovement. i*arXiv ipreprint iarXiv:1804.02767*.
22. Krizhevsky, iA., iSutskever, iI. iand iHinton, iG.E., i2012. iImagenet iclassification iwith ideep iconvolutional ineural inetworks. i*Advances iin ineural iinformation iprocessing isystems*, i*25*, ipp.1097- i1105.