Open AI 5

Artificial Intelligence for Games

Prof. Dr. Köthe

Report by

Steven Kollortz
3224192

steven.kollortz@gmx.de

Heidelberg University

Contents

Open AI 5

1. Dota 2

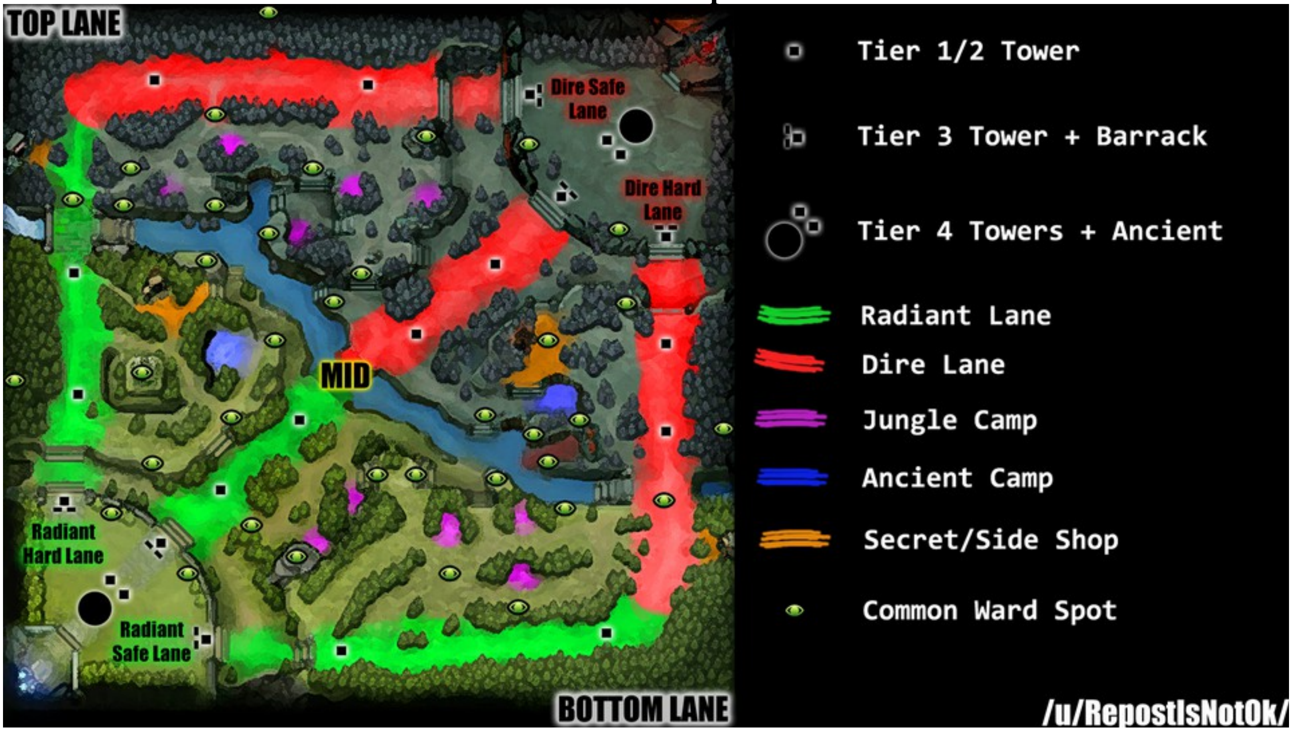
In Blizzard's game Warcraft 3 Dota was one of the most played custom maps and got developed by different people during the time. One of the developers was called Icefrog and got hired to work on Dota 2 by Steam.

In Dota two teams with five heros, each controlled by a player, face each other in order to destroy the enemies Ancient while a continous wave of non player controlled minions, also called creeps, spawn every 30 seconds in each of the three lanes attacking the opponent's creeps.
The Ancient is protected by two towers and each lane has additional three towers slowing down the process of reaching the Ancient. Each tower can only be attacked, when the lower tier tower has been destroyed which denies the possibility of bypassing everything. Once three towers of the same lane have been destroyed a barrack behind the tier 3 tower within the enemies base close to the Ancient is attackable. Destroying the barrack amplifies your own creeps on that lane, making it easier to push to the Ancient. When all three barracks have been destroyed the creeps get even stronger making it basically impossible to defend anymore.

Killing enemy creeps awards experience to friendly heros that are close, which is used to level up and unlock new abilities up to level 25. Dealing the killing blow, called last hitting, to an enemy creep also awards gold to the player which is used to buy items that have different useful effects, some items are one time use only like wards, which can be placed on the map and provide vision to that area while the ward lasts, and others provide on use effects or passive increase of stats like damage or health. Last hitting an own minion, called a deny, denies the enemy heros a portion of the experience and keeps them from earning gold.

On the map are neutral camps consisting of different amounts of neutral creeps that do not move across the map which can be killed to earn additional experience and gold.
Each hero usually has 4 unique abilities of which most have to be actively used with the heros resource mana and some are passive effects. One of the abilities, called ultimate, is usually way stronger than the other abilities but also has a longer cooldown after it is used. It unlocks at level 6 for most heros.

Dota 2 Map



2. *Why choosing Dota*

When thinking about artificial intelligence and how it can be used to solve real world problems and where it is used already you usually end up with a whole variety of tasks with different difficulties but most of them are tasks for single agents only. Playing dota on a high level is a difficult task even for humans and requires lots of time spent improving, not just as an individual, but also as a team.
The AI has to find the best solution in a high dimensional partially observed space with continuous actions that pays off the most on the long run while interacting with team mates and opponents.
The information available to a human player are represented by roughly 20000 variables, of which most are floating points with decisions made 7.5 times a second. This complexity lead to the use of a neural network since every move of the roughly 80000 per game, assuming an average length of 45 minutes, has to be calculated in real time.
Dota was chosen because it was the most viewed game on the streaming platform Twitch that supported Linux.

3. Open AI's dota timeline

November 2016	development started with the 1v1 bot
May 2017	1.5k mmr player beat the bot
Early June	bot beat the 1.5k mmr player
Late June	bot beat 3k mmr player
July	bot beat 7.5k mmr player
August 7th	bot beat Blitz (6.2k former pro player) 3-0 bot beat Pajkatt (8.5k pro player) 2-1 bot beat CC&C (8.9k pro player) 3-0
August 9th	bot beat Arteezy (10k pro player) 10-0
August 10th	bot beat Sumail (best 1v1 player worldwide) 6-0, but Sumail won against the August 9th version 2-1
August 11th	bot beat Dendi (former world champion) 2-0 bot had a 60% winrate against the version of August 10th
September 7th	the first player beat the bot with normal gameplay
Late 2017	Open AI five started development
June 2018	AI 5 plays on 4-6k mmr with strongly restricted rules
Early August	AI 5 plays on 6-7k mmr with less restricted rules (18 heros only)
Late August	AI 5 loses against pro teams (7-8k mmr) at The International 8 (TI8)
October to February 2019	AI 5 wins against different pro teams 2-0
April	AI 5 wins against OG, the winner of TI 8 and also the winner of TI 9 later that year OpenAI five arena opens where everyone can play against the AI. It ends up with a score of 7215-42

4. Hardware

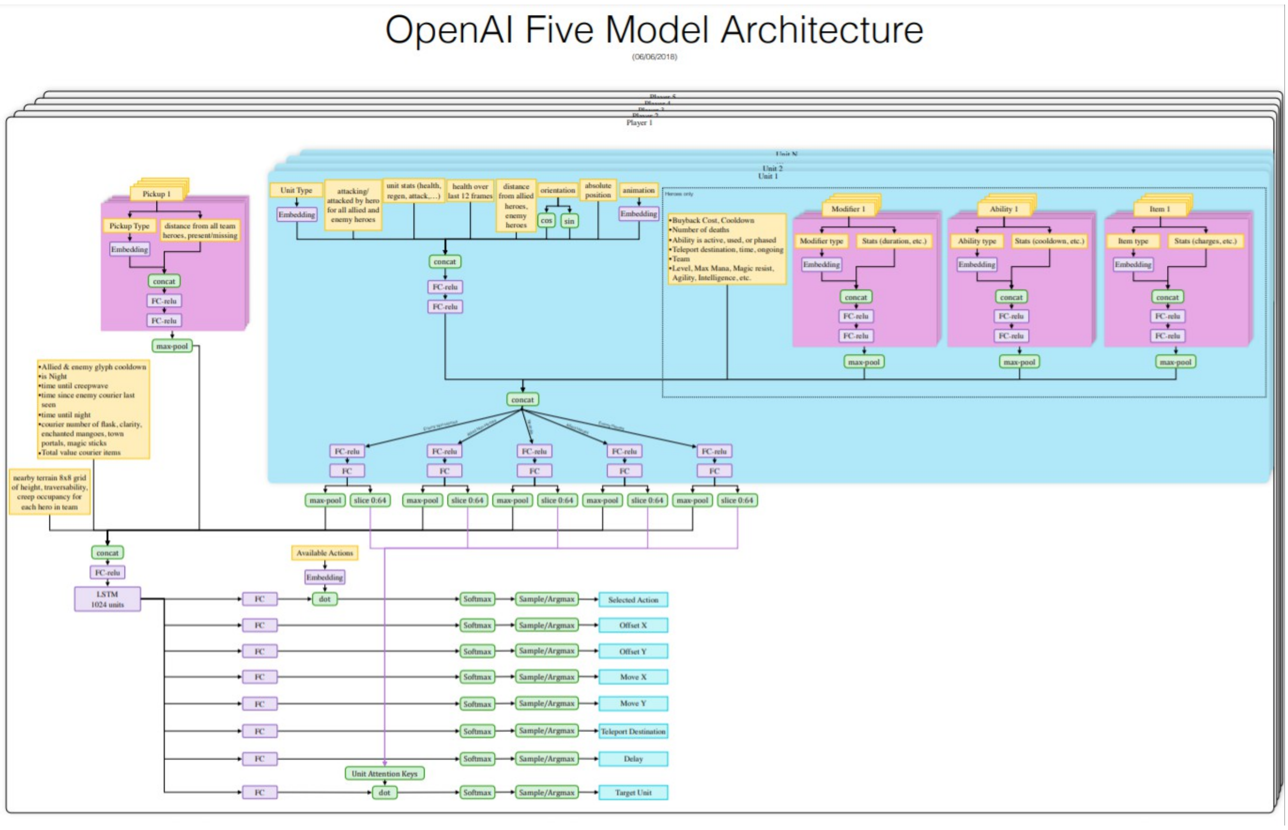
The training phase is very expensive and requires a lot of calculation power because of the input size of available data and the big possible output size.

CPUs	128000 cores on Googles Cloud Platform
GPUs	256 P100
Experience	180 years per day per hero
Observation size	36.8 kB

Open AI 5

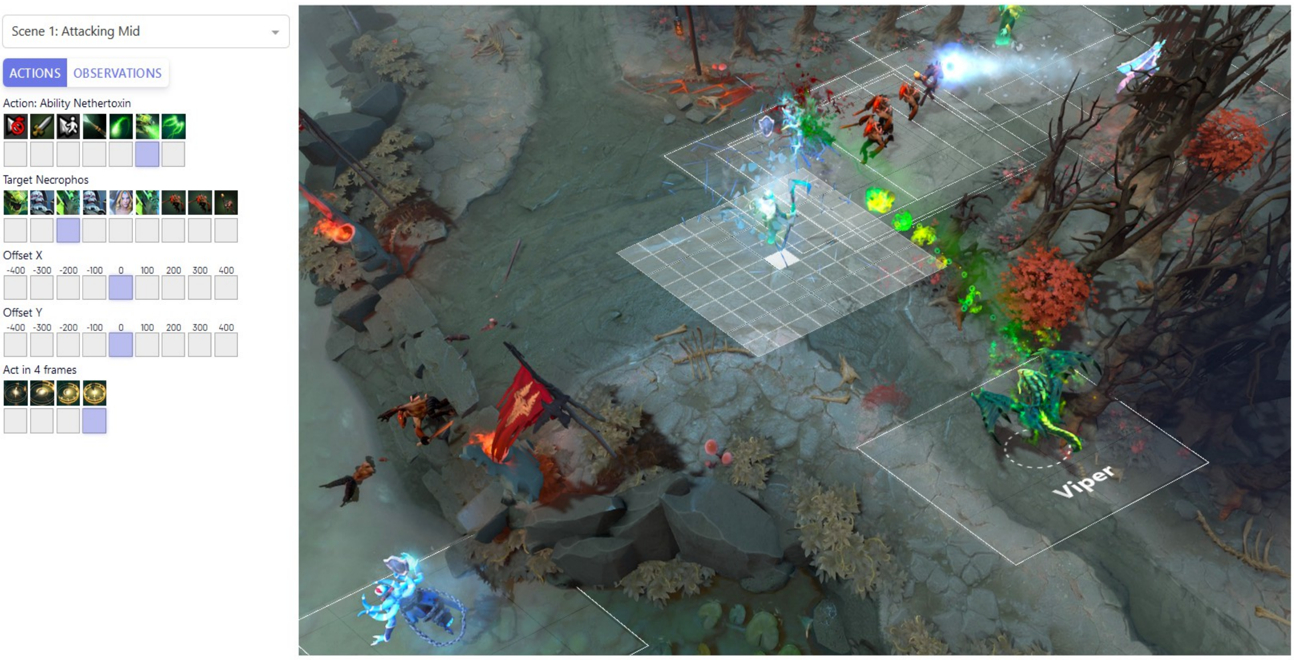
Observations per second	7.5
Batch size	1,048,576
Batches per minute	60

5. *Open AI Model*



The picture above shows the model used for each of the AI's. As input it takes all the information available for every unit, hero and pickups (some form of power ups) and global information about the map like spawn timers or local map information of allied heros.
The information includes the unit type (heros, minions, neutrals), distances to other heros, health and its development over a short time intervall, attack, defense and movement speed values, position, orientation and if it is attacking or being attacked and the animation. The animation is important because each ability or standard attacks have a short animation time that can also be canceled, in order to make the opponent react to something that is not happening or simply to guess the exact impact time of attacks. For heros it collects additional information like the cooldown of abilities, mana, modifiers like buffs or debuffs and their durations and the items with the remaining cooldowns and charges.

These information then get concatenated and feed into fully connected ReLU layers and max pooled at the end. The results of the max pool layers for the units get concatenated and split into 5 groups for allied heros, enemy heros, neutral units, allied non heros and enemy non heros. Each group is again fed into fully connected ReLU layers and then max pooled. The results of the layers handling the units and the layers for the information about the map get concatenated into a fully connected ReLU layer and then forwarded into a 1024 units big long short term memory unit. Feeding the result of the LSTM unit, the available action space and the information about the units into softmax and argmax layers produce an output of probabilities for the best action, where to move, teleport destination, delay until executed, the best target unit and it's distance.



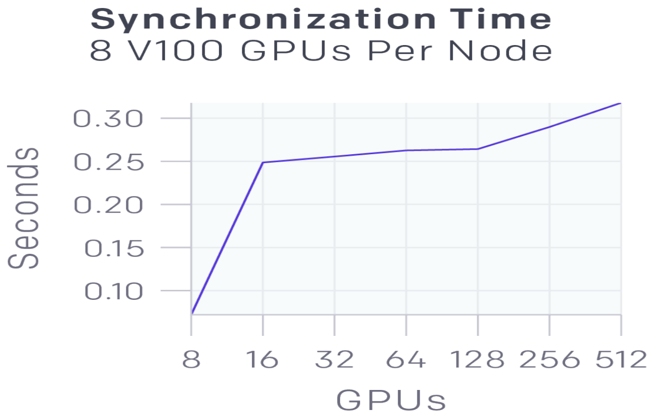
This picture shows what the AI considers for just one of its abilities, the ability selected affects a whole area. The squares around the units are the locations where the hero could cast the ability on, in order to hit the unit.

Using a big network comes with the problem of very slow training speed, by either using an expensive optimization or converging very slowly, thus OpenAI used a different algorithm than the other state of the art optimizations.

6. *Proximal Policy Optimization*

PPO is a class of reinforcement learning algorithms by OpenAI. The problems that come with policy gradient methods like using a too small or too big stepsize result in either slow progress or dropping performance. In general it takes a huge amount of samples to train for simple tasks. Considering the complexity of playing Dota it is clear that better methods have to be used and better ones exist. But approaches like Sample Efficient Actor-Critic with Experience Replay (ACER) are quite complicated and require additional adjustment while not performing much better than PPO. Trust Region Policy Optimization (TRPO) does not fit perfectly for the task as well leading to the decision to use PPO which utilizes a balance of complexity to implement, sample efficiency and is easier to tune. It uses multiple epochs of stochastic gradient descent to perform a policy update while clipping the maximum difference between the old and the new policy. This leads to rather quick convergence while stopping big changes that could ruin the overall performance.

Because the training was run on multiple GPUs, the results had to be synchronized. This was done using NCCL2, the NVIDIA Collective Communications Library. It takes around 0.3 seconds to synchronize the 58MB of the model's parameters across 512 GPUs.



7. Learning

The AI is trained using inverse reinforcement learning which ends up with better performance than using supervised learning but being very cost intensive. The data for training is generated during self play, while playing 80% against the present version and 20% against past versions to lower the chances of playing in a very exploitable style that only works against the current version.
When Open AI trained specific things they noticed for example that when placing a ward, it would often drop the ward in the direction it was about to go while the movement direction should not be the only factor for deciding where to place a ward. This made them use 'surgery' tools to split one action head into two and mapping the old parameters as initialization to the new copy which then is dedicated to being trained on wards in this example.

Open AI 5

8. Reward function

Open AI used a very low discount factor, halving further rewards only every 10 minutes. In an experiment, where they just awarded for winning or losing for the 1v1 bot, the AI trained a magnitude slower but was still able to reach decent performance (trueskill of 70 compared to the 90 their best 1v1 bot achieved). Below are some of the weights used for the reward function.

Individual	Weight	Awarded for
Experience	0.002	Per unit of experience.
Gold	0.006	Per unit of gold gained[1].
Mana	0.75	Mana (fraction of total).
Hero Health	2.0	Gaining (or losing) health[2].
Last Hit	0.16	Last Hitting an enemy creep[3].
Deny	0.2	Last Hitting an allied creep[3].
Kill	-0.6	Killing an enemy hero[3].
Death	-1.0	Dying.

1: Buying items costs gold, but spending gold does not lower the reward.
2: Health scales quadratic from 0 to 100%, punishing lower healths even more.
3: A kill awards lots of experience and gold, making the total reward for a kill positive.

The buildings reward for losing health as well, but below are just the weights for the bonus of a kill.

Buildings	Weight
Shrine	0.75
Tower (T1)	0.75
Tower (T2)	1.0
Tower (T3)	1.5
Tower (T4)	0.75

Buildings	Weight
Barracks	2.0
Ancient[4]	2.5

Extra Team	Weight	Awarded for
Mega Creeps	4.0	Killing last enemy barracks.
Win	2.5[4]	Winning the game.

4: A win basically means that the Ancient got the destroyed, awarding 5 for the lost health, 2.5 for the destruction and a bonus 2.5 for the win.

9. References

http://i.imgur.com/iqE4fxr.png
https://openai.com/blog/ai-and-compute/
https://liquipedia.net/dota2/The_International/2018/Main_Event
https://openai.com/blog/how-to-train-your-openai-five/
https://arena.openai.com/#/results
https://openai.com/blog/openai-five/
https://openai.com/blog/more-on-dota-2/
https://openai.com/five/
https://d4mucfpksywv.cloudfront.net/research-covers/openai-five/network-architecture.pdf
https://openai.com/blog/openai-baselines-ppo/#ppo
https://medium.com/@jonathan_hui/rl-proximal-policy-optimization-ppo-explained-77f014ec3f12
https://gist.github.com/dfarhi/66ec9d760ae0c49a5c492c9fae93984a
https://arxiv.org/pdf/1611.01224.pdf
https://arxiv.org/pdf/1707.06347.pdf