The Springer Series on Challenges in Machine Learning

Sergio Escalera · Markus Weimer *Editors*

The NIPS ’17 Competition: Building Intelligent Systems

Springer

The Springer Series on Challenges in Machine Learning

The books of this innovative series collect papers written by successful competitions in machine learning. They also include analyses of the challenges, tutorial material, dataset descriptions, and pointers to data and software. Together with the websites of the challenge competitions, they offer a complete teaching toolkit and a valuable resource for engineers and scientists.

More information about this series at http://www.springer.com/series/15602

Sergio Escalera • Markus Weimer
Editors

The NIPS '17 Competition: Building Intelligent Systems

Springer

Editors
Sergio Escalera
Department Mathematics & Informatics
University of Barcelona
Barcelona, Spain

Markus Weimer
Microsoft (United States)
Redmond, Washington, USA

ISSN 2520-131X ISSN 2520-1328 (electronic)
The Springer Series on Challenges in Machine Learning
ISBN 978-3-319-94041-0 ISBN 978-3-319-94042-7 (eBook)
https://doi.org/10.1007/978-3-319-94042-7

Library of Congress Control Number: 2018955494

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Years after the first KDD cup that pioneered the idea of associating competitions with data science conferences in 1993, the NIPS conference has launched its competition program. The NIPS workshops however have been hosting competitions since 2001 when the "learning from unlabelled data competition" was first launched, followed by the NIPS 2003 "feature selection challenge" and many more. Since 2013, there is a yearly workshop for competition organizers, the "Challenges in Machine Learning" workshop. This all contributed to grow a community of challenge organizers and increasing more rigorous standards of evaluation.

For its first edition, the NIPS competition program has brought to the community a very exciting set of events covering a wide range of machine learning topics. Among 23 pier reviewed proposals, 5 were accepted:

- The Conversational Intelligence Challenge
- Classifying Clinically Actionable Genetic Mutations
- Learning to Run
- Human-Computer Question Answering Competition (Quiz Bowl)
- Adversarial Attacks and Defences

Evaluation was based on the quality of data, problem interest and impact, promoting the design of new models, and a proper schedule and managing procedure. The online competitions lasted between 2 and 6 months. The Quiz Bowl competition was also run live between a team of human champions and the winning artificial system.

The workshop also included a presentation of the AI XPRIZE (https://ai.xprize.org), a 4-year contest run by IBM to encourage entrepreneurship in AI, featuring milestone results in mental health and addiction monitoring, drug design, satellite imaging of crops, virtual tutors, decontamination, and other exciting topics. A presentation was also made by the organizers of the DeepArt competition (https://deepart.io/nips/), which featured art posters decorating the NIPS conference, made with Deep Learning technology.

This book gathers contributions of the organizers and top ranking participants. Having attended the competition workshop, I was particularly impressed. How much can be expected from a handful of researchers tackling a task for such a short time? A lot indeed.

Some competitions advanced more classical aspects of machine learning such as the competition on “Classifying Clinically Actionable Genetic Mutations” but others explored successfully completely new grounds. The “Adversarial Attacks and Defences” competition examined the problem of making learning systems robust against being confused by samples closely resembling training samples, but having an entirely different meaning. This problem is particularly important to avoid malicious attacks such as modifying traffic signs to cause road accidents by fooling the computer vision systems of autonomous vehicles.

In the Learning to Run competition, the organizers provided a rather elaborate human musculoskeletal model and a physics-based simulation environment. The goal was to teach the human avatar to run through an obstacle course as quickly as possible by controlling the muscles and joints (see a video of the winners https://www.youtube.com/watch?v=8xLghMb97T0).

Two competitions dealt with natural language processing tasks and approached as closely as to get to the state of the art in artificial intelligence. The “Conversational Intelligence Challenge” competition used the scenarios of chatbots, imitating the famous Turing test. The “Human-Computer Question Answering Competition” has a regular offline version, and then the winning system competed against human champions of the Quiz Bowl game, a game similar to Jeopardy, the quiz show one year ago by IBM’s Deep Blue computer. Impressively, a neural-network based system won the game against the human champions, see the YouTube video https://www.youtube.com/watch?v=0kgnEUDMeug, with considerably less human effort and compute power than Deep Blue.

But of course, the results are only as good as the data, and progress can really be made only over a period of time with the organization of recurrent events providing each year new fresh data of ever better quality. Efforts made by governments to open up data to the public will hopefully nicely complement research and competition programs in every domain of machine learning. We wish long-lasting success and impact to the future NIPS competition programs.

UPSud/INRIA, Univ. Paris-Saclay, France
and ChaLearn, USA
NIPS 2017 General Co-Chair

Prof. Isabelle Guyon

Editor Notes This book consists of the reports from the competitions of the inaugural NIPS competitions track at NIPS 2017 in Long Beach, CA.

Competitions have become a staple in the research calendar. They allow us to gage what works and what does not, force us to develop novel approaches to problems outside the usual, and provide us with new data sets to develop machine learning approaches for. But with all the competitions happening already, and some even professionally organized by companies devoted to the task, one might ask: What can a NIPS workshop track add that is not covered? What is its niche?

When we considered the task of chairing this workshop track, we asked ourselves exactly that question. We identified a couple of *hopes* for this new track:

- Academic rigor: This being the NIPS community, its workshop track was to be as academically rigorous as the best out there.
- Spotlight novel problems: We envisioned the NIPS community to have enough draw for novel problems and data sets to be proposed as competitions.
- Generate new benchmark data sets: One of our goals was to attract new and interesting data sets to be released for the competitions.

In order to attract truly novel and hard competitions, and to make best use of the NIPS audience, we added a twist to our call for proposals: In addition to the known format for data science competitions, we introduced a new format: Live competitions to be held science-fare style at NIPS itself.

In retrospect, the gamble with both the new track at NIPS and the novel format paid off: In response to our call, we received 23 proposals. Selecting the top five of those to be run as part of the track was a difficult process, as many more were exciting. Our selection process was aided by reviewers with experience in running and winning competitions. It focused on the goals outlined above as well as practical matters of running successful competitions.

Four competitions were run in the "traditional" mode data science competition style:

- "Classifying Clinically Actionable Genetic Mutations"
- "Learning to Run"
- "Adversarial Attacks and Defences"
- "The Conversational Intelligence Challenge"

One was live competition:

- "Human-Computer Question Answering Competition"

The results of the first four were presented in the workshop at NIPS, and the latter one was run at NIPS. During that workshop, we witnessed the first win of a computer against five humans in quiz bowl. This achievement is remarkable, as Quiz Bowl is arguably harder than Jeopardy, and the winning solution was achieved on a minuscule budget compared to IBM's landmark achievement in that game. Even more remarkable was the reaction in the audience: Instead of celebrating the win for our community, we immediately switched gears to discussing how to make next year's competitions *harder* for the computer.

Hence, the NIPS community deserves the competitions track. But more so, the competitions deserve to be exposed to this community. We are looking forward to next year's incarnation. But before we do, we would like to thank the NIPS Foundation, the organizing committee of NIPS 2017, and the organizers of the five successful competitions and the XPRIZE and DeepArt linked competitions for their support and tireless work. Special thanks to Prof. Isabelle Guyon, who bring to us the idea and the real possibility to incorporate competitions at NIPS. Without it, we would not have established a new stable in the NIPS schedule.

Sergio Escalera and Markus Weimer
NIPS Competitions Chairs 2017

Contents

Chapter 1
Introduction to NIPS 2017 Competition Track

Sergio Escalera, Markus Weimer, Mikhail Burtsev, Valentin Malykh, Varvara Logacheva, Ryan Lowe, Iulian Vlad Serban, Yoshua Bengio, Alexander Rudnicky, Alan W. Black, Shrimai Prabhumoye, Łukasz Kidziński, Sharada Prasanna Mohanty, Carmichael F. Ong, Jennifer L. Hicks, Sergey Levine, Marcel Salathé, Scott Delp, Iker Huerga, Alexander Grigorenko, Leifur Thorbergsson, Anasuya Das, Kyla Nemitz, Jenna Sandker, Stephen King, Alexander S. Ecker, Leon A. Gatys, Matthias Bethge, Jordan Boyd-Graber, Shi Feng, Pedro Rodriguez, Mohit Iyyer, He He, Hal Daumé III, Sean McGregor, Amir Banifatemi, Alexey Kurakin, Ian Goodfellow, and Samy Bengio

Abstract Competitions have become a popular tool in the data science community to solve hard problems, assess the state of the art and spur new research directions. Companies like Kaggle[1] and open source platforms like Codalab[2] connect people with data and a data science problem to those with the skills and means to solve it. Hence, the question arises: What, if anything, could NIPS add to this rich ecosystem?

[1] https://www.kaggle.com/

[2] http://codalab.org/

S. Escalera (✉)
Department Mathematics & Informatics, University of Barcelona, Barcelona, Spain
e-mail: sergio@maia.ub.es

M. Weimer
Microsoft (United States), Redmond, WA, USA
e-mail: Markus.Weimer@Microsoft.com

M. Burtsev · V. Malykh · V. Logacheva
Moscow Institute of Physics and Technology, Moscow, Russia
e-mail: burtcev.ms@mipt.ru; valentin.malykh@phystech.edu; logacheva.vk@mipt.ru

R. Lowe
McGill University, Montreal, QC, Canada
e-mail: ryan.lowe@cs.mcgill.ca

S. Escalera, M. Weimer (eds.), *The NIPS '17 Competition: Building Intelligent Systems*, The Springer Series on Challenges in Machine Learning,
https://doi.org/10.1007/978-3-319-94042-7_1

In 2017, we embarked to find out. We attracted 23 potential competitions, of which we selected five to be NIPS 2017 competitions. Our final selection features competitions advancing the state of the art in other sciences such as "Classifying Clinically Actionable Genetic Mutations" and "Learning to Run". Others, like "The Conversational Intelligence Challenge" and "Adversarial Attacks and Defences" generated new data sets that we expect to impact the progress in their respective communities for years to come. And "Human-Computer Question Answering Competition" showed us just how far we as a field have come in ability and efficiency since the break-through performance of Watson in Jeopardy. Two additional competitions, DeepArt and AI XPRIZE Milestions, were also associated to the NIPS 2017 competition track, whose results are also presented within this chapter.

I. V. Serban · Y. Bengio
University of Montreal, Montreal, QC, Canada
e-mail: iulian.vlad.serban@umontreal.ca; yoshua.bengio@umontreal.ca

A. Rudnicky · A. W. Black · S. Prabhumoye
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: air@cs.cmu.edu; awb@cs.cmu.edu; sprabhum@andrew.cmu.edu

S. P. Mohanty
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: sharada.mohanty@epfl.ch

Ł. Kidziński · C. F. Ong · S. Delp
Stanford University, Stanford, CA, USA
e-mail: lukasz.kidzinski@stanford.edu; ongcf@stanford.edu; delp@stanford.edu

J. L. Hicks
Department of Bioengineering, Stanford University, Stanford, CA, USA
e-mail: jenhicks@stanford.edu

S. Levine
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
e-mail: svlevine@eecs.berkeley.edu

M. Salathé
School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
e-mail: marcel.salathe@epfl.ch

I. Huerga · K. Nemitz
Director of Engineering and Applied Machine Learning, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: huergasi@mskcc.org; nemitzk@mskcc.org

A. Grigorenko
Lead Data Scientist, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: grigoreak@mskcc.org

All these competitions emphasize advancing the state of the art of Neural Information Processing Systems as opposed to solving a singular instance of a data science problem. And this focus is the answer to the question what NIPS can add to the rich tapestry of competitions out there. And as you will find in this and other chapters in this book, the advances made are substantial.

L. Thorbergsson · A. Das
Sr Data Scientist, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: thorberl@mskcc.org; dasa@mskcc.org

J. Sandker · S. King
Talent Community Manager, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: muchaj@mskcc.org; kings1@mskcc.org

L. A. Gatys
University of Tuebingen, Tuebingen, Germany
e-mail: leon.gatys@bethgelab.org

J. Boyd-Graber
Computer Science, iSchool UMIACS, Language Science, University of Maryland, College Park, MD, USA
e-mail: jbg@umiacs.umd.edu

S. Feng · P. Rodriguez
Computer Science, University of Maryland, College Park, MD, USA
e-mail: shifeng@cs.umd.edu; pedro@snowgeek.org

S. McGregor
Technical Lead, IBM Watson AI XPRIZE, XPRIZE Foundation, Culver City, CA, USA

Member of Technical Staff, Syntiant Corporation, Irvine, CA, USA
e-mail: NIPSCompetitionBook@seanbmcgregor.com

A. Banifatemi
Artificial Intelligence Lead and IBM Watson AI XPRIZE Lead, XPRIZE Foundation,
Culver City, CA, USA
e-mail: amir.banifatemi@xprize.org

I. Goodfellow · S. Bengio
Google Brain, Mountain View, CA, USA

A. S. Ecker · M. Bethge
University of Tübingen, Germany

M. Iyyer
UMass Amherst, Amherst, USA

H. He
Stanford University, California, USA

H. Daumé III
University of Maryland, Maryland, USA

A. Kurakin
Google, San Francisco, Bay Area, USA

1.1 The Conversational Intelligence Challenge

Recent advances in the area of natural language processing driven by deep neural networks have sparked a renewed interest for dialogue systems in the research community. In addition to the growing real-world applications, the capacity to converse is closely related to the overall goal of AI. The Conversational Intelligence Challenge had a goal to unify the community around the challenging task of building systems capable of intelligent conversations. Teams were expected to submit dialogue systems able to carry out intelligent and natural conversations about snippets of Wikipedia articles with humans. At the evaluation stage of the competition participants, as well as volunteers, were randomly matched with a bot or a human to chat and score answers of a peer. The competition had two major outcomes: (1) an assessment of state-of-the-art dialogue systems quality compared to human, and (2) an open-source dataset collected from evaluated dialogues.

1.1.1 Task

The goal of competing bots was to maximize an average score of dialogs rated by human evaluators. The evaluation was performed through a blind cross testing of bots and other human users in a series of chat sessions. Members of participating teams, as well as volunteers, were asked to log into an anonymous chat system and communicate with randomly chosen bot or another human user. No information about identity of the peer was provided. Both peers received the text of a snippet from a Wikipedia article. Discussion of the article proceeded till one of the peers ended dialog. Then human user was asked to score the quality of every response and the dialog as a whole.

1.1.2 Running the Competition

The Conversation Intelligence Challenge was split in four stages. Starting from the beginning of April of 2017 participants submitted applications consisting of a proposal describing details of scientific approach and statement of work as well as a system architecture and relevant technical information. After review of applications teams were invited to submit working solutions for the qualification round till the middle of July, 2017. During the last week of July these solutions were evaluated by participants of the summer school-hackathon DeepHack Turing[3] and volunteers.

This evaluation process generated the dataset of rated human-to-bot and human-to-human dialogs. Dataset of rated dialogs was open sourced and participating teams were able to tune their solutions on these data. Two weeks before the NIPS

[3] http://turing.tilda.ws/

conference final versions of bots were run in the test system and final evaluation round was started and lasted till the day before the Competition track session at NIPS.

Competing teams were required to provide their solutions in the form of executable source code supporting a common interface (API). These solutions were run in isolated virtual environments (containers) and were not be able to access any external services or the Internet to prevent cheating. The master bot created by organizers facilitated communication between human evaluators and the competitors' solutions. It was implemented for popular messenger services[4] and allows to connect a participant to a randomly selected solution or peer and log the evaluation process.

1.1.3 Outcomes

Major goals of the competition were establishing a new non-goal-driven but still topic-oriented task for dialogue, probing the current level of the conversational intelligence for this task and collecting dataset of evaluated dialogs.

Ten teams applied for the challenge and six of them were able to submit working solutions. Final score of the dialogue quality for the best bot was 2.746 compared to 3.8 for human.[5] We found that human-to-human dialogs were longer and humans used shorter utterances. Higher length of dialogs possibly indicates higher engagement of peers. It was also found that human performance in dialogue at both utterance and dialogue levels is generally rated high, but not exclusively high, which suggests that either human utterances or scores (or both) are not always reliable.

As a result of data collecting effort 4.750 dialogues were recorded in total. Among them there are 2.640 human-to-bot and 359 human-to-human *long* dialogues where each participant produced at least three utterances. The dataset is available in the competition repository[6] and as a task in ParlAI framework.[7]

Participation in this type of challenges requires significant engineering effort. To make the entrance in the field easier the source code of participated solutions was published in the repository of the competition.[8] A well-documented baseline solution for the future competition will also be available.

Better promotion of the competition in academy and industry is needed to get more participating teams and volunteers for evaluation. Another measure to increase

[4] http://m.me/convai.io or https://t.me/convaibot

[5] Possible scores were from one to five with former corresponding to the bad and the latter to the excellent dialogue quality.

[6] http://convai.io/data/

[7] http://parl.ai/

[8] https://github.com/DeepPavlov/convai/tree/master/2017/solutions

engagement of human evaluators might be to change the task from discussion of incidental text snippet to the discussion focused on the topics that are more interesting to the user.

The first Conversational Intelligence Challenge was a successful attempt to test the ground for a large scale dialogue system competition with evaluation by human volunteers. Results of the competition demonstrate that current state of conversational artificial intelligence allows to support dialogue with a human on a given topic but with quality significantly lower compared to human. Closing this gap will not only bring a major progress in solving fundamental problems of artificial intelligence research but also open possibilities for a wide range of industrial applications. We are looking forward to continue exploration of possible solutions to the problem of making machines talk like humans with the next Conversational Intelligence Challenges.

1.2 Classifying Clinically Actionable Genetic Mutations

The increase in genetic sequencing capabilities combined with the decrease in cost have been instrumental for the adoption of cancer genetic testing in the clinical practice. Genetic testing may detect changes that are clearly pathogenic, clearly neutral or variants of unclear clinical significance. Such variants present a considerable challenge to the diagnostic laboratory and the treating clinician in terms of interpretation and clear presentation of the implications of the results to the patient. There does not appear to be a consistent approach to interpreting and reporting the clinical significance of variants either among genes or among laboratories. The potential for confusion among clinicians and patients is considerable and misinterpretation may lead to inappropriate clinical consequences. Currently this clinical interpretation of genetic variants is being done manually.

This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic variant based on evidence from the clinical literature. MSK pioneered the creation of OncoKB, a knowledge base where evidence for these genetic variants is being collected, and manually curated. It takes a molecular pathologist around 3 h to curate a single variant. To date more than 88 million genetic variants have been discovered in the Human Genome by the 1,000 Genomes project.[9] Therefore this task is completely unfeasible via the current manual processes.

The scope of this competition was to develop a classification model that can compete with a human curator in some of the tasks described. This would have a considerable high impact on the health care and cancer domains.

[9] http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html

1.2.1 Task

This is a classification task with the main goal of using the evidence from the literature to classify the genetic variants in one of the *Oncogenicity* and *Mutation Effect* classes.

There are four **Oncogenicity** classes, *Likely Oncogenic*, *Oncogenic*, *Likely Neutral* and *Inconclusive*. There are nine **Mutation Effect** classes, *Likely Gain-Function*, *Loss-of-function*, *Likely Loss-of-function*, *Likely Neutral*, *Inconclusive*, *Neutral*, *Gain-of-function*, *Switch-of-function*, *Likely Switch-of-function*.

When the curator decides to investigate a genetic variant, she currently has to manually carry out two tasks. First, she has to manually search the medical literature to identify abstracts that can provide evidence for the interpretation of the genetic variant of study. Second, she needs to read and interpret all these abstracts to ultimately classify the genetic variant in one of the *Oncogenicity* and *Mutation Effect* classes. This second task is the most time consuming, and the goal of this competition. In a real-world scenario our curators would still make manual searches when a new genetic variant needs to be studied. But getting into this competition we could envision a situation where after identifying abstracts containing potential evidence from the literature, the human experts would pass them as input to a model that classifies them into their corresponding *Oncogenicity* and *Mutation Effect* classes.

1.2.2 Data

The data for this competition was made available in the public domain via the OncoKB Data Access[10] page. It could be accessed via REST APIs, or simply downloaded in two different versions - Actionable Variants[11] or All Variants.[12]

The Table 1.1 below shows a detailed description of a manually annotated genetic variant in the Actionable Variants data set. The first column *Gene* refers

Table 1.1 Detailed description of a genetic variant in the data set

A sample annotation

Gene	Alteration	Oncogenicity	Mutation effect	PMIDs for mutation effect	Abstracts for mutation effect
ERBB2	L869R	Likely oncogenic	Likely gain-of-function	21531810, 26305917, 16397024	Hyman et al.

[10] http://oncokb.org/#/dataAccess

[11] http://oncokb.org/api/v1/utils/allActionableVariants.txt

[12] http://oncokb.org/api/v1/utils/allAnnotatedVariants.txt

to the gene that is being annotated. *Alteration* represent the aminoacid change for that specific mutation. *Oncogenicity* denotes whether or not this specific mutation has been identified as oncogenic, or cancer-causing, in the literature. *Mutation Effect* represent the effect of this mutation in downstream molecular pathways. *PMIDs for Mutation Effect* represents the Pubmed abstracts that the human curator had to read to be able to classify the *Oncogenicity* and *Mutation Effect* of the specific variant. Pubmed abstracts are publicly available via the National Library of Medicine's REST API. Finally *Abstracts for Mutation Effect* provides links to specific abstracts from the medical literature that might have been made available in selected conferences such as the American Society of Clinical Oncology (ASCO). These are also abstracts that the human curator will have to manually analyze in order to classify this genetic variation in one of the *Oncogenicity* and *Mutation Effect* classes.

1.2.3 Running the Competition

This competition was particularly challenging to run due to the size of the data set, with less than 10,000 observations. This particular problem is very common in the medical domain where obtaining manually labeled samples is extremely costly and very often machine learning practitioners need to come up with creative ways to counter for this limitation.

In our case, we decided to run this competition in two stages. During the first stage (June 26th, 2017 to September 25th, 2017) participants would have access to the OncoKB samples available in the public domain. And a short second stage (September 26th, 2017 - October 2nd, 2017) where we made available a holdout dataset with 1,000 samples. Finally, participants were evaluated only against the holdout dataset made available during stage two of the competition.

In terms of logistics, we used the Kaggle platform to run this competition. In our case this worked particularly well since we were able to leverage Kaggle's community to encourage users to participate in our competition. One factor to emphasize for future organizers of this type of competitions is that selecting the right platform to run your competition is one of the critical decisions to make.

1.2.4 Outcomes

Our main goal getting into this competition was twofold. First, we wanted to introduce the Machine Learning Community with real world challenges in health care that could potentially be solved via Machine Learning. Second, we wanted to leverage this community to find out a solution for a very particular problem that we at MSKCC have, classifying clinically actionable mutation.

We can now definitely say that we achieve these two goals. On the one hand we had more than 1,500 participants taking part and submitting at least one solution to our competition. This definitely proves that the competition was a success in terms of raising awareness within the community. On the other hand, the best scenario possible for MSKCC was that at least one of the participants would come up with an innovative solution to our particular problem that we could implement and deploy into our production clinical pipeline. Thanks to this competition we found not just one, but two. The solutions from the Cornell University and Uber Technologies teams are currently being evaluated by our clinicians for their integration within our clinical workflow. Therefore, we also clearly achieved our second goal of finding a solution that would have a clear clinical impact. These two solutions will be described in detail in two separate chapters in this volume.

1.3 Learning to Run

Synthesizing physiologically accurate movement is a major challenge at the intersection of orthopedics, biomechanics, and neuroscience. An accurate model of the interplay of bones, muscles, and nerves could potentially allow to predict variability in movement patterns under interventions (e.g. a surgery) or new conditions (e.g. an assistive device or prosthetics).

In this challenge, participants were tasked to build controllers for a neuromusculoskeletal system without any experimental data, i.e. solely through exploration of simulated physics. The role of a controller was to observe sensory information and actuate muscles in order to make the model move forward as quickly as possible, while avoiding obstacles. Over the course of 4 months 442 participants submitted 2154 controllers. The competition has proven not only that the task is approachable despite high dimensionality of the solution space, but also that the movement patterns generated through reinforcement learning resemble human gait patterns.

1.3.1 Task

Given a neuromusculoskeletal model, i.e. a set of bones connected by joints and muscles attached to the bones, participants were tasked to make the model move as far as possible within 10 seconds of simulation time. To control the models, they were sending signals actuating muscles causing the model to move according to predefined dynamics. Decisions were taken on a discretized time-grid of 1000 equidistributed time-points.

Actuation signals were defined as vectors $a_t = [0, 1]^{18}$ corresponding to excitation of 18 muscles (0 – no excitation, 1 – full excitation). The simulation of dynamics was performed in OpenSim (Delp et al. 2007) – a physics engine dedicated to musculoskeletal simulations.

For the purpose of this section, the simulation engine can be seen as a function M from the space of states and muscle actuations to the space of states. Let S_t be a state of the system at time t, then:

$$S_{t+1} = M(S_t, a_t).$$

Participants did not observe the entire state, but only a function $O(S_{t+1})$ which included positions and velocities of the center of mass, bodies, and joints. Simulations were terminated either when the time finished (i.e. after 1000 time-steps) or when the vertical position of the pelvis fell below 0.65 meter, what was interpreted as a fall.

The objective of the competition was to build a controller synthesizing the fastest movement, without falling and without extensive use of ligaments. We quantified this objective through a reward function. At each step of simulation, agent receives a reward $r(t)$ defined as

$$r(t) = d_x(t) - \lambda\sqrt{L(t)},$$

where $d_x(t)$ is the change of position of pelvis in this time-step, $L(t)$ is the sum of squared forces generated by ligaments at time t and $\lambda = 10^{-7}$ is a scaling factor. Let T be the termination time-step. The total score of the agent is the cumulative reward till T, i.e. $R = \sum_{t=1}^{T} r(t)$.

In order to enforce building robust controllers, we introduced two types of obstacles, randomly chosen for each simulation. First, we had variable strength of the psoas muscles simulating an injury. Second, we placed spherical obstacles along the path, enforcing adaptation of the steps. Information about both kind of obstacles was included in the observation vector.

1.3.2 Running the Competition

The competition was running in two stages: the open stage (4 months) and the play-off stage (2 weeks). In the open stage, participants were interacting with the grading server iteratively, at every time-step. Thanks to an elementary API, this allowed for very simple on-boarding of participants. In the play-off stage, participants were asked to prepare a docker container with their solution. This allowed for testing the solution in exactly the same environments and for reproducibility of the actual controllers, which was crucial for a post-hoc analysis of the results.

For running the challenge we needed a customized platform. First, our challenge did not rely on any data, so it did not fit classical data science settings, typical to platforms like Kaggle. Second, both stages of the challenge required customized solutions: the first one requires direct interaction with the grader, while the second one requires a docker-based infrastructure. These circumstances directed us towards open platforms and we decided to host the challenge on the crowdAI,[13] while leveraging OpenAI gym (Brockman et al. 2016) infrastructure for grading. Implementations of all components of our challenge, i.e. the simulation engine, the grading server, the docker-based grading system as well as the entire crowdAI platform are all open source.

1.3.3 Outcomes

The objective of the challenge was to answer two questions: (1) Are the modern reinforcement learning techniques capable of solving high-dimensional non-linear continuous control problems? (2) are the movement patterns emerging from reinforcement learning physiologically relevant? Due to the very large space of solutions and no theoretical guarantees on finding global solutions with reinforcement learning, we cannot expect to definitely answer these questions through a challenge. Instead, we rather perceive them as exploration of potential new directions of research in computational biomechanics.

Top solutions submitted to the challenge partly answer both questions. First, the winning solution was running at around 4.5 m/s, equivalent to fast human jogging. The running gait pattern is very complex and the fact that it emerge under very weak assumptions imposed on the controller is most remarkable. Second, we observed weak similarities in angular joint kinematics between the top solutions and experimental data on running. We discuss them in detail in the "Learning to run" chapter (Kidziński et al. 2018).

1.4 Human-Computer Question Answering Competition

Question answering is a core problem in natural language processing: given a question, provide the entity that it is asking about. When top humans compete in this task, they answer questions incrementally; i.e., players can interrupt the questions to show they know the subject better than their slower competitors. This formalism is called *quizbowl* and was the subject of the NIPS 2015 best demonstration.

In this year's iteration, competitors could submit their own system to compete in a quiz bowl competition between computers and humans. Entrants created systems that receive questions one word at a time and decide when to answer. This then provided a framework for the system to compete against a top human team of quizbowl players in a final game.

[13] http://crowdai.org/

1.4.1 Data

We created a server to accept answers to a set of questions and provided quiz bowl question data to train and validate systems. This data also comes with preprocessed text versions of the Wikipedia pages associated with each answer in the training set. We encourage the use of external data in addition to what we have provided.

1.4.1.1 Test Set

The test set had possible answers from any Wikipedia page. However, many of the answers will likely be in the train set (the same things get asked about again and again). Around 80% of test questions are about answers in the train set. The test questions were written by quiz bowl writers based on the standard high school distribution.

1.4.2 Competition

The competition had two phases: a machine competition to select a top computer team and a human-machine phase to pit the top computer entry against a strong team of trivia experts.

1.4.2.1 Machine Evaluation

We evaluated systems (and humans) in pairwise competition. The system that gives a correct answer first (i.e., after requesting the fewest number of words) gets 10 or 15 points (15 points are available for early, correct buzzes). A system that gives an incorrect answer first will lose 5 points. There is no penalty for guessing at the end of a question. The system with the higher overall score wins.

Participants interacted with a server architecture that replicates the process of playing a quiz bowl game. Systems get each word in the question incrementally and can decide to answer (or not) after every word. We break ties randomly when systems are evaluated against each other.

1.4.2.2 Human-Machine Evaluation

The top computer team faced off against six strong trivia players from the Los Angeles area and from the NIPS community. The questions came from the same pool of questions used in the computer competition. The system OUSIA decisively won the competition, 475–200 (Fig. 1.1).

Fig. 1.1 Our human-computer competition at NIPS 2017. The top computer submission faced off against a team of top trivia players from the LA area

1.4.3 Outcomes

In the competition overview chapter, we describe how this competition is not the final word for comparing question answering abilities of humans and machines. In many ways, this competition is tilted in favor of the machines, and we can improve the competitiveness of the competition through adversarial writing of questions, forcing machines to interpret speech, changing the difficulty of questions, and focusing on questions whose answers are less well-represented in Wikipedia.

1.5 Adversarial Attacks and Defenses

Most existing machine learning classifiers are highly vulnerable to adversarial examples. An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it.

Adversarial examples pose security concerns because they could be used to perform an attack on machine learning systems, even if the adversary has no access to the underlying model.

The purpose of adversarial attacks and defenses competition was to increase awareness of the problem and stimulate more researchers to explore potential solutions. In this competition participants were invited to submit methods which craft adversarial examples (attacks) and classifiers which are robust to adversarial examples (defenses). Attack methods were ranked based on how many times they fool defenses and defense methods were ranked based on their accuracy on adversarial examples.

1.5.1 Task and Evaluation Metrics

Adversarial attacks and defenses competition had 3 tracks and participants were invited to submit a solution in one or several tracks:

- **Non-targeted adversarial attack.** In this track participants were invited to submit a method which performs non-targeted adversarial attack, i.e. given an input image generate adversarial image which potentially be misclassified by a defense.
- **Targeted adversarial attack.** In this track participants were invited to submit a method which performs targeted adversarial attack, i.e. given an input image and a target class generate adversarial image which potentially be classified as a given target class by a defense.
- **Defense against adversarial attacks.** In this track participants were invited to submit an image classifier which is robust to adversarial examples.

During evaluation all attack methods were run on the provided dataset to craft adversarial images, then these adversarial images were fed into all defenses and classification labels were computed.

An attack got 1 point each time it was able to fool a defense on a single image. If attack was unable to fool a defense or was unable to generate adversarial image then it got 0 points for that image. A defense got 1 point for each correctly classified image and 0 points for incorrect classification or failure to produce classification label. Points for each submission were added together and then normalized (using common normalization constant for all submissions), such that the final scores of all submissions were in the range [0, 1], where 1 means success on all images and 0 means failure on all images.

1.5.2 Dataset

Dataset of source images which were fed to attacks was composed of ImageNet-compatible images. We constructed this dataset by collecting images available online under CC-BY license, automatically cropping and classifying these images with help of the state-of-the art ImageNet classifier, then manually verifying labels and discarding images with invalid labels.

We prepared two datasets. DEV dataset contained 1000 images and was provided for development of the solutions as well as for evaluation of development round. FINAL dataset contained 5000, was kept secret and was used for final evaluations of all solutions.

1.5.3 Running the Competition

Competition was announced in May 2017, launched in the beginning of July 2017 and finished on October 1st, 2017. Competition was run in multiple rounds. There were three development rounds (on August 1, 2017, on September 1, 2017 and September 15, 2017) followed by the final round with submission deadline on October 1st, 2017.

Development rounds were optional and their main purpose was to help participants to try and test their solution. Final round was used to compute final scores of submissions and determine winners.

We partened with Kaggle,[14] which hosted competition web-site, forum, leaderboard and was used to upload submissions. During the evaluation of each round, we disabled submission uploads and took all already uploaded submissions from Kaggle and run them on our customly build infrastructure on Google Cloud[15] platform. Then results were published online and submission upload was re-enabled.

1.5.4 Outcomes

Main goals of the competition were to increase awareness of the adversarial examples and stimulate researchers to propose novel approaches to the problem.

Competition definitely increased awareness of the problem. Article «AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks»[16] was published in MIT Technology review about the competition. And in the end we got 91 non-targered attack submissions, 65 targeted attack submission and 107 defense submissions participating in the final round.

There were good results and interesting approaches among the submissions. Best non-targeted attack achieved 78% success rate against all defenses on all images. Best targeted attack achieved 40% success rate, which is quite impressive because targeted black box attacks are generally hard. Top defense submission got 95% accuracy on all adversarial images produces by all attacks. This indicates that it may eventually be possible to be robust to adversarial examples at least in the black box situation (i.e. when attacker is unaware of the exact defense).

[14] www.kaggle.com

[15] www.cloud.google.com

[16] https://www.technologyreview.com/s/608288

Tools, competition datasets and several baseline method were published online[17] as a part of development toolset. Additionally most of the participants released their submissions under open source licences[18] as was required by competition rules.

1.6 IBM Watson AI XPRIZE Milestones

The IBM Watson AI XPRIZE is a 4 year competition awarding a $5 million prize purse to teams improving the world with artificial intelligence (AI). Teams competing for the prize are permitted to propose the grand challenge they will solve with AI and judges select advancing teams in each year of the competition. Teams advance on the basis of technical and logistical achievement and the importance for humanity of the team's presumed solution.

The competition began in 2017 with 148 teams competing to solve problems in sustainability, robotics, artificial general intelligence, healthcare, education, and a variety of other grand challenge problem domains (see Table 1.2). The first judgment round winnowed the field to 59 teams. Of these 59 advancing teams, 10 teams were nominated for Milestone Awards and two teams won Milestone Awards.

Table 1.2 High level problem domain descriptions for teams competing for the IBM Watson AI XPRIZE. The rows are ordered from domains with the highest advancement rate (top) to the lowest advancement rate (bottom). Figure 1.2 gives additional details on advancement rates

Problem domain	Team count	Example problem area
Humanizing AI	7	Moral and ethical norming
Emergency Management	5	Planning disaster response logistics
Health	13	Drug efficacy prediction
Life Wellbeing	21	Augmenting the visually impaired
Environment	8	Automated recycling
Education/Human Learning	17	Intelligent tutoring system
Civil Society	11	Online filter bubbles
Health Diagnostics	12	Radiography image segmentation
Robotics	5	Robotic surgery
Knowledge Modeling	7	Automated research assistant
Civil Infrastructure	9	Earthquake resilience testing
Business	19	Optimizing social investment
Artificial General Intelligence	8	* (all of them)
Brain Modeling and Neural Networks	6	Cognition emulation

[17] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition

[18] Links to code of non-targeted attacks: https://www.kaggle.com/c/6864/discussion/40420, targeted attacks: https://www.kaggle.com/c/6866/discussion/40421, defenses: https://www.kaggle.com/c/6867/discussion/40422

1.6.1 Running the Competition

Teams submitted competition plans to the XPRIZE Foundation in the first quarter of 2017. Following a survey of team problem areas, the XPRIZE Foundation recruited a panel of 34 judges with core competencies in either artificial intelligence research (e.g., natural language processing, robotics, etc.) or team problem areas (e.g., cancer diagnosis, civil society, etc.). Teams submitted four page First Annual Reports in September. Judges bid on these reports within the EasyChair conference management system and an assignment algorithm generated a proposed set of reviewers. XPRIZE staff then adjusted EasyChair report assignments to ensure that every team would be reviewed by at least one AI researcher.

Judge reviews separated teams into Milestone Nominees, advancing, and rejected groups on the basis of their *overall rating, importance for humanity, existing solution status, progress indicators,* and *technological capacity for solving the problem*. The 10 teams with the highest average overall rating were nominated for Milestone Awards.

The judges then each reviewed two additional teams from the Milestone nominee list and labeled one of the teams as having the better First Annual Report. After ranking the teams from the pairwise comparisons, the top two teams were awarded a total of $15,000 during the NIPS Competition Track.

1.6.2 Outcomes

The characteristics of advancing, rejected, and awarded teams highlight the problem domains with the greatest challenges and opportunities for improving the world with artificial intelligence. Our NIPS Competition Track chapter surveys the problem domains and technologies of the IBM Watson AI XPRIZE, details the prize judgement process executed to date, and treats the advancement decisions of judges as opportunity indicators for the “AI for Good” movement (see Fig. 1.2). The results show where AI researchers may fruitfully direct their efforts to address problems that are simultaneously important for humanity, technically challenging, and feasible to solve within 4 year timelines.

1.7 Neural Art Challenge

Since its introduction in 2015, Neural Style Transfer (Gatys et al. 2016) has had a big impact in a number of areas. It not only produces beautiful artistic pictures, which attracted world-wide media attention. But it also introduced novel perceptual loss functions to measure image similarity, which was particularly useful for fields such as image processing and image synthesis.

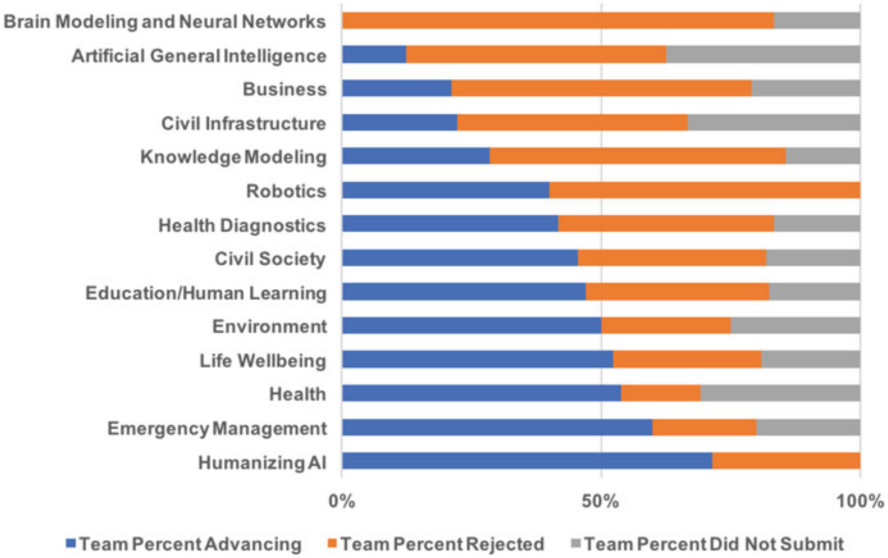

Fig. 1.2 Stacked percent bar charts for the 148 teams. The first series (blue) represents percentage of the teams within the problem domain advancing. Similarly, the orange and gray bars represent teams that were judged and rejected and did not submit a report, respectively

Neural style transfer requires two ingredients: a photograph that defines the content and a painting that defines the style. The algorithm then combines the two and renders the content of the photograph in the style (or texture) of the painting. The rendering is done via a so-called pre-image search. We iteratively update the content image by gradient descent until it minimizes the sum of two loss functions: a content loss and a style loss. Both losses are computed in the feature space obtained by passing the images into a deep convolutional neural network (VGG-19 (Simonyan and Zisserman 2015)) trained on large-scale image recognition (ImageNet (Russakovsky et al. 2015)). The content loss tries to match the activations of the content image and the rendering in a high-level convolutional layer (conv4_2). These high-level layers are relatively invariant to low-level features like color or small local perturbations, which allows for some flexibility in changing the style while maintaining the important shapes (content) of the image. The style loss tries to match spatial summary statistics (correlations of feature maps) between the style image and the rendering in a number of layers. Matching summary statistics enforces that the texture features from the style image are transferred onto the rendering, but does not constrain their spatial arrangement. The resulting image looks like the content of the photograph has been swapped into the painting.

The goal of the Neural Art Challenge was not to advance science in any way, but instead to demonstrate the breadth of artistic effects that can be achieved with this simple image synthesis procedure. Our main goal was to engage the NIPS community in a fun project to decorate the conference center with neural art.