Building, growing and sustaining ML communities

Martin Andrews
Red Dragon AI
Singapore
martin@reddragon.ai

Sam Witteveen
Red Dragon AI
Singapore
sam@reddragon.ai

Abstract

While there are multiple research-based groups for the ML community around the world, the adoption of these skills by a broader base of developers will require new communities that reach beyond researchers to flourish at a large scale.

The Singapore TensorFlow & Deep Learning community is a group of over 3,000 people from different backgrounds and levels that is pushing the adoption of ML in South-East Asia, via monthly in-person meetings, guest talks, and special events.

In the proposed short talk, we will present some of the challenges, lessons learned and solutions found to building machine learning communities at scale.

1 Communities for ML adoption

Building a community of a large size is a difficult undertaking, requiring both hard work while maintaining a delicate touch, since communities by their nature are challenging to steer, and are often prone to collapse or fragmentation.

Software developers and computer science students have very little knowledge of ML practices or their uses in the real world. Typically coming from a background of being able to dictate a program's logic, these developers often find the machine learning approach to be confusing and intimidating.

Our group has a `TensorFlow` [1] bias, however we also regularly have talks that cover `PyTorch` [2], `numpy` [3], `pandas` [4], and `opencv` [5] - with a mix of topics, from beginner to advanced.

1.1 MLOSS Libraries & Frameworks also need evangelists

The majority of people in the internal community of contributor to an OSS project are focused on writing code and adding features. However, this alone is unlikely to cause the library or framework to become widely adopted : Evangelists are needed to show the greater developer community what they stand to gain from the use of the OSS project. These evangelists need to be able to create relevant content that not only matches the level of understanding of the wider developer base, but also matches their real-world use cases.

Projects that fail to bridge this gap will find it hard to gather the early wide-spread adoption that they should otherwise deserve.

1.2 What Libraries & Frameworks can do to widen their appeal

If an MLOSS project wishes to gain wide user-acceptance, then the project needs to give clarity to the broader developer community in a variety of ways:

- **Differentiation** : being clear what the project is and what it is not. Many MLOSS projects stray into trying to do everything and end up doing nothing very well.

Preprint. Work in progress.

- **Examples** : having clear documentation with examples of uses for different levels of technical proficiency, as well as uses in several application areas.
- **Ecosystem** : explaining how the project fits into the overall ecosystem. This is often left for developers to find out for themselves.

2 Importance of in-person communities

As might be expected in any open source software project "working code wins". However, particularly in the ML arena, in-person meetings are particularly important:

- **Social aspect** : Software development is often a solitary activity, and being able to mingle with like-minded people can be an essential element of validation to keep one's spirits up.
- **Best practices and "know how"** : Seeing how others are tackling problems (through demos, ideally), and understanding what actually works in practice is invaluable.
- **Discovering new MLOSS projects** : Knowing that there is a local early-adopter of a project is often a key step to trying new technology oneself.

2.1 How projects can facilitate community up-take

There are several steps that can greatly aid community organisers in promoting an MLOSS project:

- **Relevant examples** : Often, MLOSS projects are lead by gifted engineers for whom delivering the shortest 'MNIST example' is the best way to demonstrate the project's cleverness. However, if that example is completely abstracted (to simplify the code), its teaching value is likely `nil` - because it will be extremely difficult for a newcomer to unpack what is actually required to use the tool in more realistic situations.
- **Aspirational goals / road maps** : These can be very useful to illustrate where the project is heading - even if the code doesn't exist yet (which is *relatable* in-person).
- **Shout-outs** : Simply linking to the community organisers' event in the `README.md` can (a) give the event credibility in the eyes of the audience, and (b) give the project itself valuable social proof to newcomers on the internet.

3 Lessons learned from a successful community

Reflecting on what has worked (and what hasn't) over the last 18+ months in the Singapore TensorFlow & Deep Learning MeetUp, we have identified several key factors, outlined below.

3.1 Holding events

- **Consistency** : Holding the MeetUp on a monthly basis *without fail* is an important trust factor for the group. This, in turn, assures members (and prospective new members) that the event will be worth attending, even without a content description.
- **Content mix** : Enforcing a mix of content at each event has also been an important factor. Our stated mantra is "Talks for people starting out; Something from the bleeding-edge; and Lightning Talks". This requires a steady supply of speakers, many of which need to be prepared to talk in simple terms to the beginners in the audience.
- **Entertainment value** : We also recognise that each audience member has a choice about how to spend their free time. We have found that the easiest way to educate people is also to entertain them. [1] [2]

[1] An example 'topical' talk was Satay-not-Satay, which is a nice local variation on the Hotdog-not-Hotdog theme, and was easy to re-purposed into a discussion of transfer learning; models on mobile; data issues with the training set; etc. Similarly, pix2code [6] is an attention-grabbing demonstration for developers.

[2] In addition, we find that most developers find mistakes made by ML models to be very entertaining.

3.2 Audience Mix

Understanding the community's make-up is important to holding successful events:

- **Experience levels** : We are constantly surprised how much of our audience considers themselves to be beginners (~40%). However, to build a sustainable community, it is essential that the new members see that there is a path to gaining understanding, and quickly 'upgrading' to being able to follow (for instance) explanations and demonstrations of recent academic papers. There are also advanced members (~20%) that will gladly come to events just to be 'involved' - but they (to some extent) do not actually need special content.
- **Diversity** : We have found that by making sure the audience sees 'themselves' reflected in the speaker make-up, the diversity of the group is self-sustaining. Moreover, because of our monthly multi-talk format, we have a huge incentive for getting more people involved, and try to be as welcoming as possible.
- **Willingness to participate** : A simple technique for encouraging people to speak is though requiring lightning talks to have a hard limit on both time and number of slides. Often, speakers will discover that it is easy to be enthusiastic for 5 minutes, and find that they would prefer to have more time 'next time'. Our lightning talk rubric emphasises enthusiasm and having a 'can do' attitude over achieving perfection before presenting.

3.3 Speakers

Finding speakers is often difficult[3], but the following ideas have proven persuasive:

- **Deadlines create action** : An underappreciated aspect of speaking is that having a deadline is a huge motivator to make progress on a personal project.
- **Thematic events** : Centering an event around a theme (like `NLP` or `Explainability`) may also prompt speakers to volunteer, since they understand that a similar opportunity might not arise for some time in the future.
- **Social standing and résumé value** : The "1:9:90" meme (~ creators:contributors:lurkers [7]) can be used to convince audience members to 'step up'.

Some speakers have a ulterior corporate motivation for speaking (eg: recruiting). Our rule that speakers "must show code" has been effective in eliminating low-value marketing pitches.

4 Conclusions

This short paper has covered some of the themes in our proposed short talk about community building. We are advocates for the value of in-person communication, and our vibrant group of over 3,000 members demonstrates that there is substantial appetite for MLOSS from the developer community.

5 Resources

The Singapore TensorFlow & Deep Learning community pages can be found at:

`https://www.meetup.com/TensorFlow-and-Deep-Learning-Singapore/`

Source code for the Deep Learning Workshop - including a repository of code examples:

`https://github.com/mdda/deep-learning-workshop`

Presentations and videos of talks to be found in each MeetUp's comments section, and at:

`http://blog.mdda.net/tags.html#Presentation-ref`

[3] The two main group organisers are on always on stand-by to pick up any remaining vacant slots.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[3] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[4] Wes McKinney. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, pages 1–9, 2011.

[5] Gary Bradski and Adrian Kaehler. Opencv. *Dr. Dobb's journal of software tools*, 3, 2000.

[6] Tony Beltramelli. pix2code: Generating code from a graphical user interface screenshot. *CoRR*, abs/1705.07962, 2017. URL `http://arxiv.org/abs/1705.07962`.

[7] Trevor van Mierlo. The 1% rule in four digital health social networks: An observational study. *J Med Internet Res*, 16(2):e33, Feb 2014. ISSN 14388871. doi: 10.2196/jmir.2966. URL `http://www.jmir.org/2014/2/e33/`.