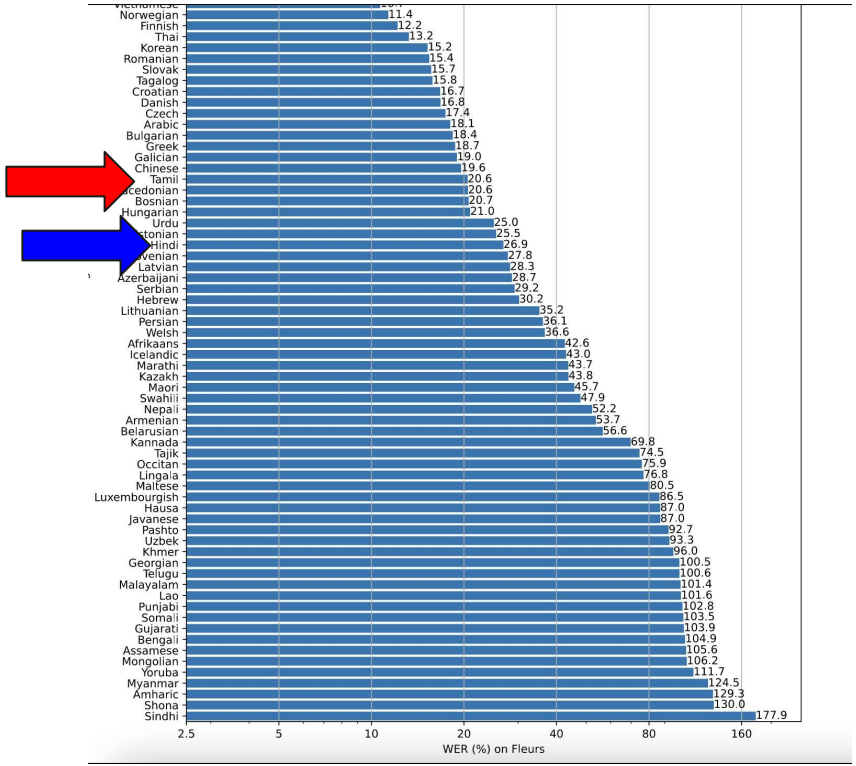
DISCLAIMER: Images belong to the respective copyright owners!

AI and Tamil Computing Opportunities
செயற்கையறிவு - தமிழ் மொழி சார்ந்த வாய்ப்புகள்

முத்து அண்ணாமலை, Ph.D. | ezhillang@gmail.com | @ezhillang

OpenAI Whisper - ASR model WER



WER of Tamil vs Hindi, (+Indian languages) on Whisper

1

ஆழக்கற்றல் - Deep Learning - அறிமுகம்

1.1. துறையின் வரலாறு, நூல்கள் (பொதுவெளி)
1.2. செயற்கை பின்னல்கள் - Neural Networks
1.3. பயன்பாடுகள்: இயல்மொழி பகுப்பாய்வு - NLP, காட்சிசார் புரிதல் - CV
1.4. சிக்கல்கள் - விளைவுகளை

Deep Learning - எந்திரவழி கற்றல்

இடது->வலது: டிம்னிட் ஜீப்ரு, அனிமா ஆனந்த்குமார், பெய்பெய் லி. பிறபல செயற்கையறிவு ஆய்வாளர்கள்.

> எந்திரவழி கற்றல் கணினி நிரல் எழுதாமல் தரவுகளின் வாயில் ஒரு செயலிகளை உருவாக்கலாம்

> தேவையானவை:

- அல்கொரிதங்கள் - கணினி செயற்கை பின்னல்களின் உருவாக்கம்
- தரவுகள் - செயற்கை பின்னல்களை உருவாக்கி பயில்வி
- கணினி பயில்விக்க வன்பொருள்கள்

2018 டூரிங் விருது பெற்றவர்கள்:
இடது->வலது: யான் லே கூன், ஜேப் ஹிண்டன்,
யாஷுவா பெஞ்சியோ

> முழுதான அறிமுகம் : பக்கம் 83-இல் இருந்து இந்த ஆவணத்
https://docs.google.com/presentation/d/1aITk-4FvHEsCDjyCK6AI3c6oKH
sp=sharing

- செயற்கை பின்னல்கள் வரலாறு, தோற்றம்
- மைல் கல்கள் - வளர்ச்சி
- சமிபத்திய வளர்ச்சி, மும்மூர்த்திகள்

செயற்கை பின்னல்கள் - Neural Networks

> மனித மூளையில் 8600 கோடி (86 பில்லியன்) நரம்புகள் உள்ளன ஆனால் 20 வாட் (இரவு மின்விளக்கு அளவில்) வேலை செய்கிறது!

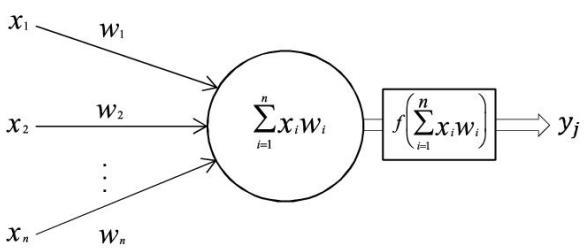
> function approximation

> gradient descent algorithm to train ANN

> DNN = {ANN} பல அடுக்குகள் கொண்ட செயற்கை பின்ன

https://en.wikipedia.org/wiki/Artificial_neural_network

(a)



$$h_\theta(x) = \sum_{j=0}^{n} \theta_j x_j$$

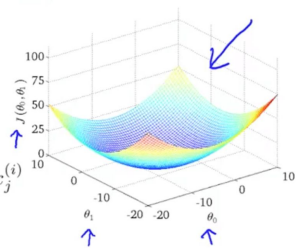
$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

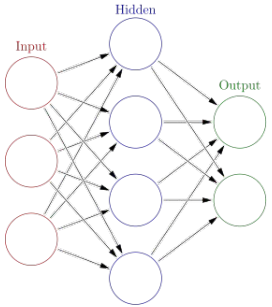
Repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

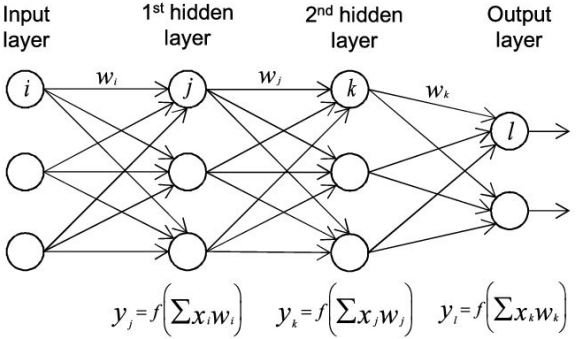
(for every $j = 0, \ldots, n$)

}





(b)



Regression and Classification

> எந்திரவழி கற்றல் கணினி கற்றலில் இரண்டு வகையான செயல்பாடுகள் (supervised learning)

> Regression

- ஒரு ஊரில் உள்ள வீட்டின் மதிப்பீடுகளை கொண்டு ஒரு செயற்கை மாதிரியை தயார் செய்வது
- அதன் வாயில் ஒரு புதிய, பயிற்சி செய்யாத, வீட்டின் மதிப்பை கணக்கிடுவது

> Classification - பகுப்பாய்வு

- ஒரு சொற்குவியலின் படி ஒரு மொழிமாதிரியை தயார் செய்வது
- ஒரு புதிய சொல் என்ன வகையானது ? பெயர்ச்சொல், வினைச்சொல் ?

இயல் மொழி புரிதல் - Natural Language Processing

> NLP Understanding Tasks

> Classification

> Topic Modeling

> Sentiment Analysis

> Large Language Models (BERT, Transformers, GPT, Megatron, SQUAD, StableDiffusion, Whisper, etc.)

> Collaborative Filtering

> Recommendation Engines

செயற்கை அறிவு, பின்னல்களின் ஆபத்து 💀 🔴

- AI + Climate Change - உலக வெப்பமயமாதலுக்கு செயற்கை பின்னல்களின் பயிற்ச்சிக்கான மின் தேவைகள் பங்களிக்கின்றன!
  - மனித மூளையில் 8600 கோடி (86 பில்லியன்) நரம்புகள் உள்ளன ஆனால் 20 வாட் (இரவு மின்விளக்கு அளவில்) வேலை செய்கிறது!
  - சமிபத்திய வன்பொருள்கள் (GPU) போன்றவை 300 வாட் அளவில் ஒரு கணிப்பு செய்கிறது; பயிற்சி என்பது பல் கிலோவாட் மின் சக்தி தேவைப்படுகிறது
- சர்வாதிகார கண்காணிப்பு
  - சமிபத்தில் சீனா போன்ற கம்யூனிச நாடுகளில் சிறுபான்மியனரின் உரிமைகள் கண்காணிப்புக்கு உள்ளாக்கப்பட்டது
- Bias in AI
  - தரவுகளில் உள்ள தப்புகள் செயற்கைஅறிவினால் உள்வாங்கி சிறுபான்மையினரை தவறாக பாவிக்கிறது
  - E.g. credit score, விளம்பரங்களில் ஓரவாஞ்சனை (Latanya Sweeny, Harvard), முகம் அறிதல் [கருப்பினத்தவர்களின் முகங்களை கண்டறிய மறுப்பது] ...
- “Coded Bias” - சமுகத்தில் உள்ள ஒடுக்குமுறைகளை செயற்கையறிவில் வரையறுப்பது சரியா?
- குப்பம்மா - உளிவீரன் அப்படின்னு பெயர்வெச்சா கடன் அட்டை கிடைக்காமல் போகவும் ராகுல், ப்ரியா என்று பெயர் வைத்தால் கிடைப்பதற்கும் உள்ள வித்தியாசம் தான் "Coded Bias" - எனில் செயற்கைஅறிவு உங்களுக்கு இது கிடைக்குமா

2

Opportunities
வாய்ப்புகள்

Abdul Majed Raja, Muthiah Annamalai, “Tools for constructing AI/ML solutions in Tamil,” Tamil Internet Conference, 2022.

2.1 Generative AI - DALL-E images

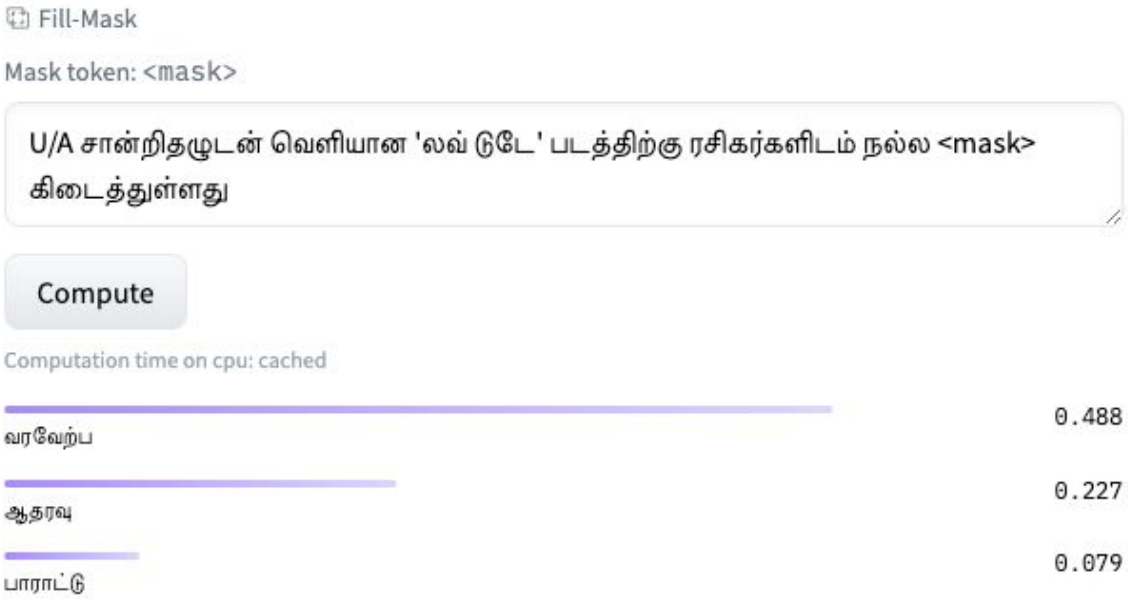
2.2 OpenAI Whisper - ASR in multiple-languages

புளி = உளி
பதை = பறை

ஓ காவிரியால் நீர் மடிக்கி அம்பரமாய் அணை எடுத்தான். நீர் சத்தம் கேட்டதுமே நெல் பூத்து நிக்கும் புளி சத்தம் கேட்டதுமே கல் பூத்து நிக்கும் பதை சத்தம் கேட்டதுமே வில் பூத்து நிக்கும் சொழத்தின் பெருமை கூற சொல் பூத்து நிக்கும்

நன்றி : மலைக்கண்ணன், சாமா
டெக்னாலஜீஸ்

2.3 Large Language Models - multiple applications



3. Tooling Requirements

Tamil datasets are available than Tamil tools – arunthamizh அருந்தமிழ்.

However the accessibility of fully-trained models and capability of providing pre-trained models are much harder and still require domain expertise in hardware and software. P

Inadequate to scale the breadth of Tamil computing needs in AI world among:

1. NLP – Text Classification, Recommendation, Spell Checking, Correction tasks
2. TTS – speech synthesis tasks
3. ASR – speech recognition

Ultimately the tooling provides capability to quickly compose AI services based on open-source tools and existing compute environment to host services and devices in Tamil space.

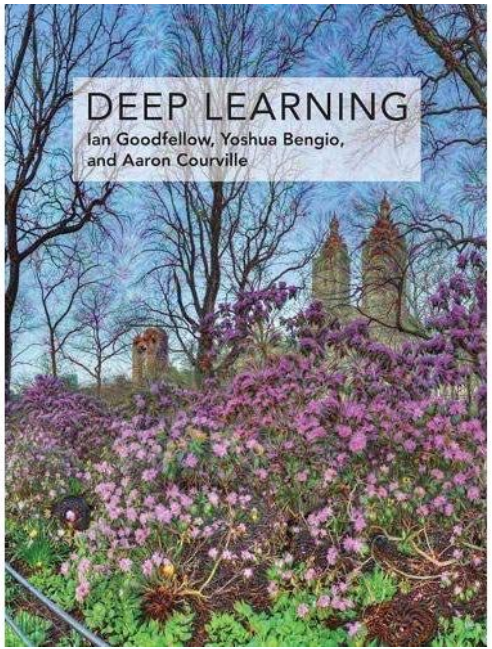
3.1 Proposal - fill tooling gaps

1. Develop a open-source toolbox for pre-training and task training specialization
2. Identify good components to base effort
3. Contribute engineering effort, testing, and validation
    1. R&D – DataScience, Infra, AI framework
    2. Engineering Validation – DataScience, Tamil language expertise
    3. Engineering – packaging, documentation, distribution
    4. Project management
4. Library to be liberally licensed MIT/BSD
5. Open-Source license for developed models
6. Find hardware resources for AI model pre-training etc.
7. Managed by a steering committee / nominated BDFL
8. Scope – decade time frame
9. Financial support for such a wide effort

3.2 Call to Action

- Let's build a pytorch-lightning like API for Tamil tasks across NLP, TTS, ASR via AI.
- Consortium / co-operative ownership model
- Open-Source foundations
- Convert datasets into functions

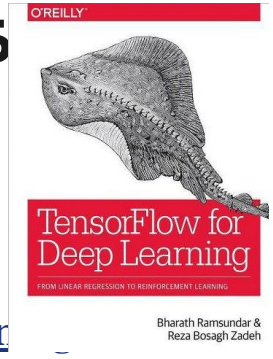
Deep Learning - நூல்கள் - பொது வளங்க

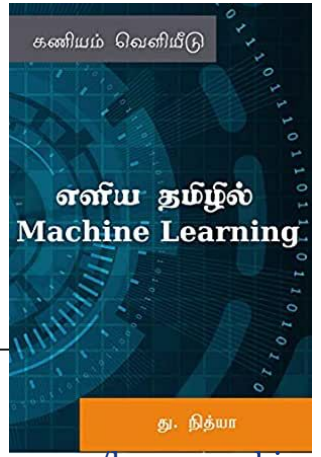


https://www.deeplearningbook.org

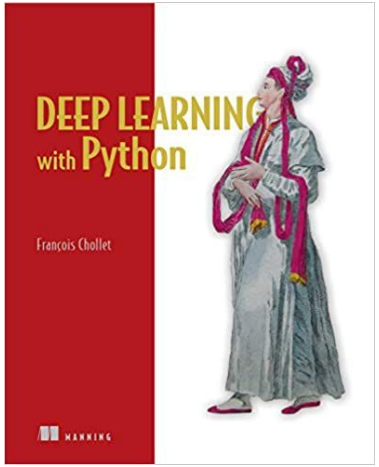
Neural Networks and Deep Learning - M. Nielsen, 2015

http://neuralnetworksanddeeplearn





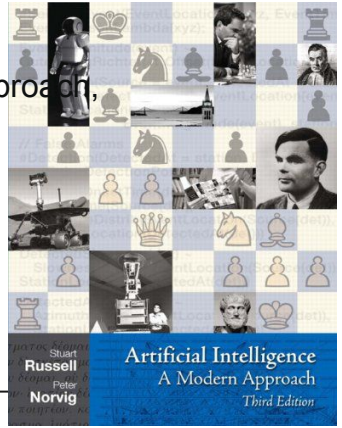
https://www.manning.com/books/deep-learning-with-python



http://www.kaniyam.com/learn-machine-learning-in-tamil

Artificial Intelligence: A Modern Approach, 2nd ed.

Peter Norving, and Stewart Russel



தமிழ் முன்னுரை - https://ezhillang.blog/2019/04/25/deep-learning-

நன்றி!

தொடர்பு:

அஞ்சல்: ezhillang@gmail.com

வலைப்பூ: https://ezhillang.wordpress.com

கீச்சுகள்: @ezhillang

நிரலகம்: http://github.com/Ezhil-Language-Foundation/

திட்டங்கள்: http://tamilpesu.us