ML and NLP Research Highlights of 2021
[Sebastian Ruder]

王龙跃

2022年2月9日

Sebastian Ruder

Sebastian Ruder

FOLLOW

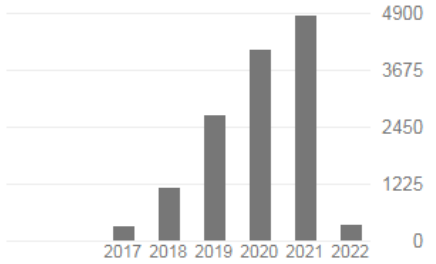
Research Scientist, Google
Verified email at google.com - Homepage

Natural Language Processing Machine Learning Deep Learning Artificial Intelligence

TITLE	CITED BY	YEAR
An overview of gradient descent optimization algorithms S Ruder arXiv preprint arXiv:1609.04747	5083	2016
Universal Language Model Fine-tuning for Text Classification J Howard, S Ruder Proceedings of ACL 2018	2685 *	2018
An overview of multi-task learning in deep neural networks S Ruder arXiv preprint arXiv:1706.05098	1774	2017

Cited by

	All	Since 2017
Citations	13684	13587
h-index	31	31
i10-index	48	48



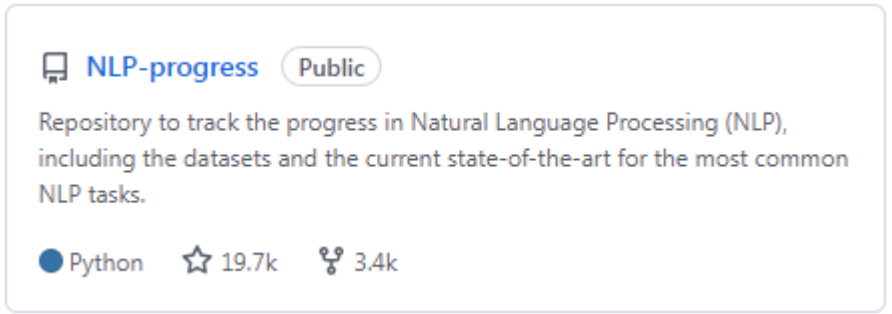
Education

- **National University of Ireland** — **Galway, Ireland**
  College of Engineering and Informatics, Ph.D. Natural Language Processing — *10/2015 – 02/2019*
- **University of Copenhagen** — **Copenhagen, Denmark**
  Research visit, Natural Language Processing Group, Computer Science Department — *04/2017 – 06/2017*
- **Trinity College** — **Dublin, Ireland**
  Semester abroad, School of Computer Science and Statistics, Computer Science and Language — *09/2014 – 01/2015*
- **Ruprecht-Karls-Universität Heidelberg** — **Heidelberg, Germany**
  Institute of Computational Linguistics, B.A. Computational Linguistics, English Linguistics — *10/2012 – 09/2015*

Sebastian Ruder

Experience

- **Google**
  Research Scientist
  London, United Kingdom
  11/2021 – Present
- **DeepMind**
  Research Scientist
  London, United Kingdom
  03/2019 – 10/2021
- **AYLIEN**
  Research Scientist
  Dublin, Ireland
  10/2015 – 02/2019
- **IBM**
  Extreme Blue Intern, Watson
  Munich, Germany
  08/2015 – 09/2015
- **Microsoft**
  Linguistic Engineering Intern, XBox
  Dublin, Ireland
  02/2015 – 06/2015
- **The OpenCog Foundation**
  Google Summer of Code Intern
  `opencog.org`
  Summer 2014
- **Lingenio GmbH**
  Software Engineering Intern
  Heidelberg, Germany
  Spring 2014
- **SAP**
  Working Student, Development University
  Walldorf, Germany
  02/2013 – 02/2014
- **TEMIS**
  Freelancing Developer
  Heidelberg, Germany
  02/2013 – 10/2013



PhD

Highlights

原文链接：https://ruder.io/ml-highlights-2021/index.html
中文转载：https://hub.baai.ac.cn/view/14415

- Universal Models
- Massive Multi-task Learning
- Beyond the Transformer
- Prompting
- Efficient Methods
- Benchmarking
- Conditional Image Generation
- ML for Science
- Program Synthesis
- Bias
- Retrieval Augmentation
- Token-free Models
- Temporal Adaptation
- The Importance of Data
- Meta-learning

Highlights 分类

模型

- Universal Models
- Massive Multi-task Learning
- Beyond the Transformer

应用

- ML for Science
- Program Synthesis
- Conditional Image Generation

评价

- Benchmarking

方法

- Prompting
- Efficient Methods
- Meta-learning

问题

- Bias
- Retrieval Augmentation
- Token-free Models
- Temporal Adaptation
- The Importance of Data

模型

Universal Models

Massive Multi-task Learning

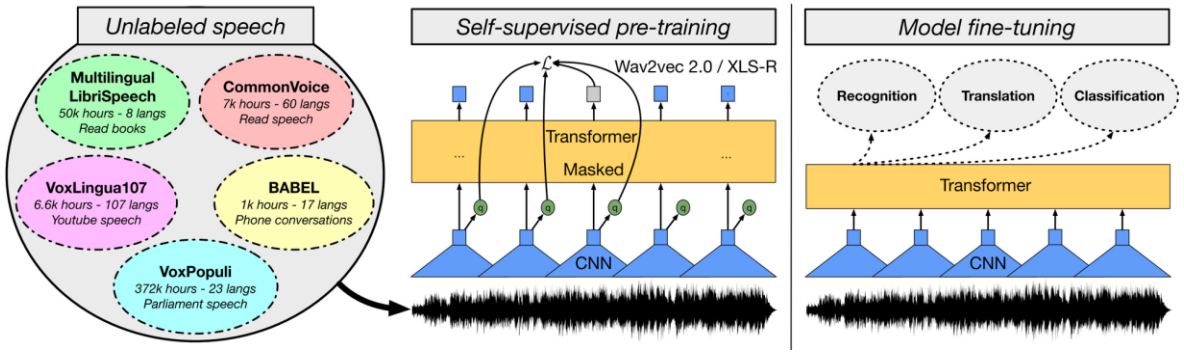
Beyond the Transformer

模型

总结：

- “大融合”
  - 统一多模态预训练模型：Universal Models
  - 超大规模的多任务学习：Massive Multi-task Learning
- “从抢地盘到开始深耕”
  - 初期开荒已经基本完成：Transformer
  - 下一阶段深入思考：架构耦合度、训练方式等小地盘：Beyond the Transformer

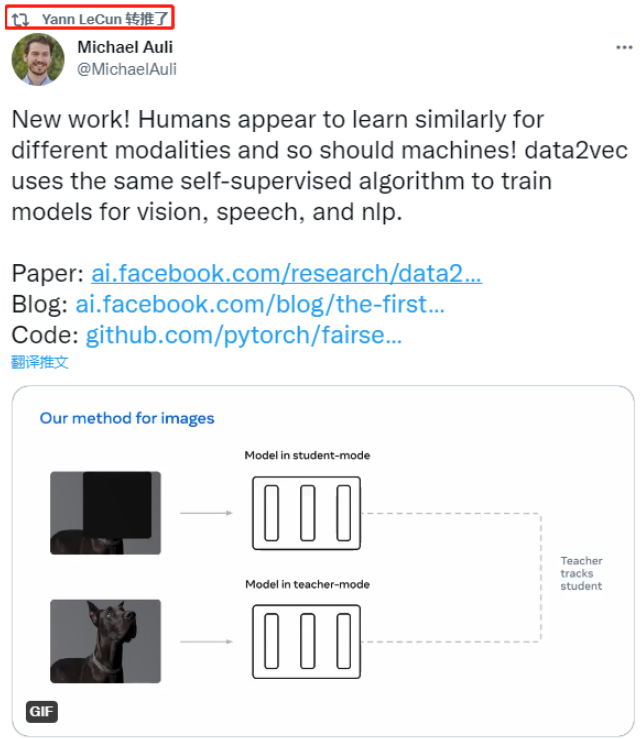
模型 - Universal Models



Self-supervised cross-lingual representation learning on speech using XLS-R (Babu et al., 2021).

背景、进展、问题：

- 继续 larger pre-trained models
- 在其他领域架构统一且性能满足，如Vision Transformer
- 多模态成为必然趋势（Facebook）
- 报告不同参数大小的模型性能已经变得很普遍
- 预训练模型性能的提升并不一定会转化到下游任务



模型 - Universal Models

入选理由：

- 多模态融合学习，既有研究价值，又有应用意义
- 人类学习也是多融合的

未来趋势：

- 模型还会继续变大
- 其他模态子任务 we will likely see image and speech models that can perform many common tasks in a single model.
- 没有太多预测，follow Facebook

Jiatao Gu

TITLE

Data2vec: A general framework for self-supervised learning in speech, vision and language
A Baevski, WN Hsu, Q Xu, A Babu, J Gu, M Auli
URL https://ai. facebook. com/research/data2veca-general-framework-for-self ...

Textless Speech-to-Speech Translation on Real Data
A Lee, H Gong, PA Duquenne, H Schwenk, PJ Chen, C Wang, S Popuri, ...
arXiv preprint arXiv:2112.08352

Neural actor: Neural free-view synthesis of human actors with pose control
L Liu, M Habermann, V Rudnev, K Sarkar, J Gu, C Theobalt
ACM Transactions on Graphics (TOG) 40 (6), 1-16

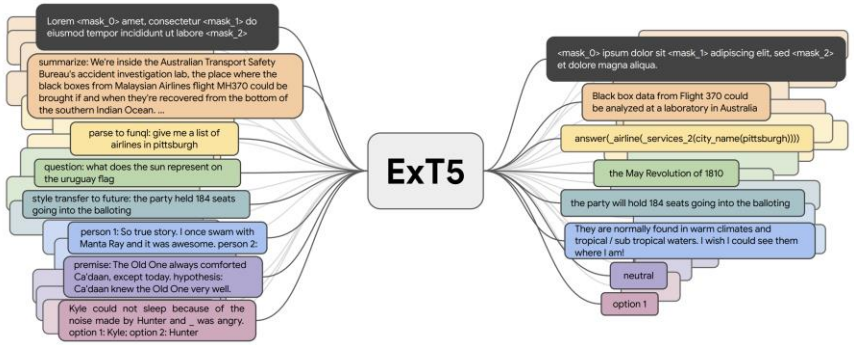
Volume rendering of neural implicit surfaces
L Yariv, J Gu, Y Kasten, Y Lipman
Advances in Neural Information Processing Systems 34

Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis
J Gu, L Liu, P Wang, C Theobalt
arXiv preprint arXiv:2110.08985

fairseq S^ 2: A Scalable and Integrable Speech Synthesis Toolkit
C Wang, WN Hsu, Y Adi, A Polyak, A Lee, PJ Chen, J Gu, J Pino
arXiv preprint arXiv:2109.06912

Multilingual translation from denoising pre-training
Y Tang, C Tran, X Li, PJ Chen, N Goyal, V Chaudhary, J Gu, A Fan
Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 ...

模型 - Massive Multi-task Learning



Massive multi-task learning with ExT5 (Aribandi et al., 2021).

背景、进展、问题：

- 从unlabelled data 到 labelled data利用
- Google 主推的方向，源于应用需求
- 目前已经有100+任务一起学习
- Such massive multi-task learning is closely related to meta-learning (in-context learning).

  Given access to a diverse task, models can learn to learn different types of behavior.

模型 - Massive Multi-task Learning

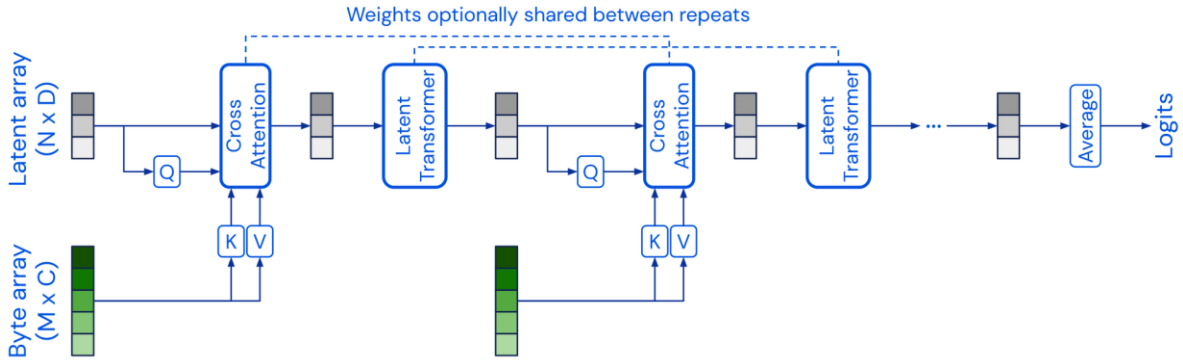
入选理由：

- Combining self-supervised pre-training with supervised multi-task learning
- Models thus no longer require hand-engineered task-specific loss functions or task-specific layers in order to effectively learn across multiple tasks

未来趋势：

- 更加贴近真实场景的方法及性能： labelled data 充分利用
- 启发机器翻译情况，简单利用labelled data已有相关工作

模型 - Beyond the Transformer



Cross-attention and self-attention blocks are then applied alternatingly (Jaegle et al., 2021).

背景、进展、问题：

- 大部分pre-trained model是transformer 架构
- 提出很多 alternative model architectures特别是att：Perceiver， MLP-Mixer，gMLP
- 将架构与预训练策略解耦，如CNN 以与 Transformer 模型相同的方式进行预训练
- 改变预训练的objective：ELECTRA

模型 - Beyond the Transformer

入选理由：

- 探索互补性complementary或正交性orthogonal，是推动研究的一个手段
- 新模型可能会解决一些 Transformer 的限制，例如注意力的计算复杂性、黑盒性质和顺序不可知性 order-agnosticity

未来趋势：

- pre-trained transformers 会继续作为许多任务的标准基线
- 通过对比，可以发现更多pre-trained transformers 的缺陷，例如建模long-range依赖或者可靠性可解释性方面

方法

Prompting

Efficient Methods

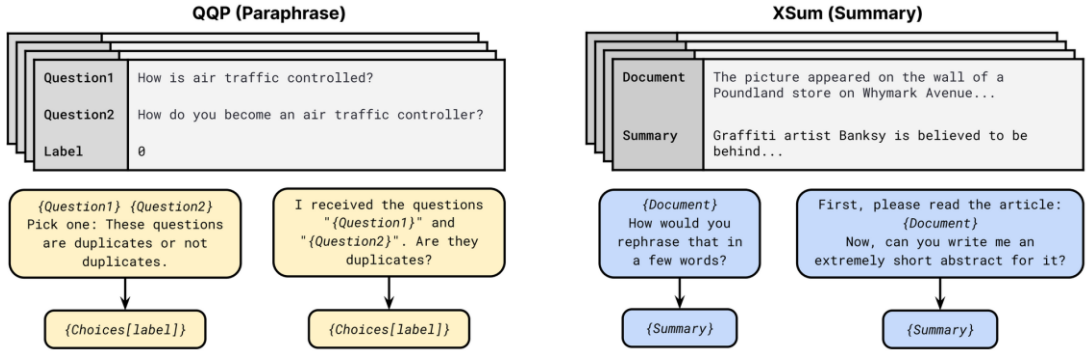
Meta-learning

模型

总结：

- 新范式，新挑战
  - 初探提示学习，还不是灵丹妙药：Prompt Learning
  - 从皮毛到深入，提示将变得更加精细
- 已有方法更偏实用性
  - 如何更加高效使用预训练模型：Efficient Methods
  - 反思：Meta-learning

方法 - Prompt Learning



Prompt templates from the P3 prompt collection (Sanh et al., 2021).

背景、进展、问题：

- GPT-3 是开端，早期 in-context learning
- 系列工作快速占领： PET, iPET, and AdaPET leverage prompts for few-shot learning
- 还不稳定，模型的性能因提示而异，找到最佳提示仍然需要标记的示例
- public pool of prompts (P3) 已收集大量模板
- 提示与传统的分类、标注、生成契合度还不够

方法 - Prompt Learning

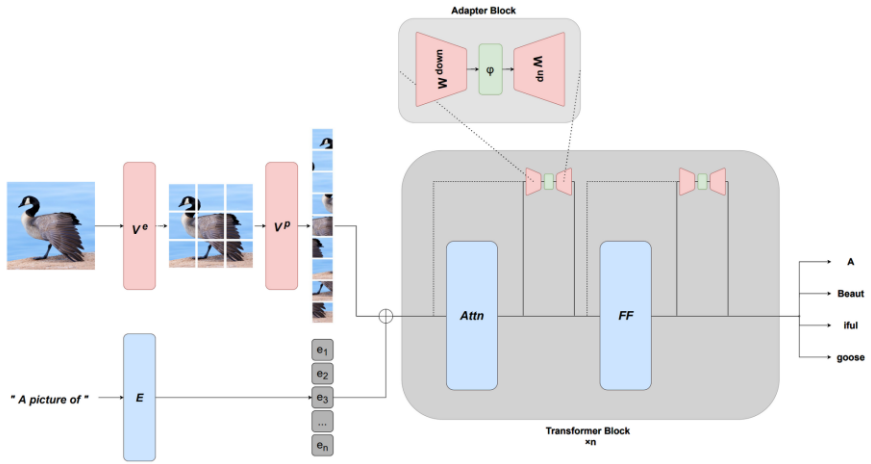
入选理由：

- 符号主义和联结主义的契合，将规则引入神经网络的一种可选方式（刘群）
- 与已有Few-shot setting方法相互借鉴，meta-learning

未来趋势：

- 只是触及了使用提示来改进模型学习的皮毛
- 提示将变得更加精细，例如包括更长的指令、正反示例和一般启发式
- Prompts也可能是将自然语言解释纳入模型训练的更自然的方式：可信赖AI

方法 - Efficient Methods



Multi-modal adaptation with MAGMA (Eichenberg et al., 2021).

背景、进展、问题：

- 预训练实用化是个重要问题，谷歌和华为在早期跟进
- 主要从 efficient architectures 和 efficient fine-tuning methods两个方面入手，如Random Feature Attention 和 Adapters
- 预训练实用化阶段还有很多问题可以做

方法 - Efficient Methods

入选理由：

- 如果模型在标准硬件上运行不可行或过于昂贵，那么它们就没有用处
- 效率的提升使得将在模型变得更大的同时，确保其可供从业者使用

未来趋势：

- 模型和方法应该更加简洁和可得
- without having to pre-train a new model from scratch
- 预训练实用化阶段还有很多问题可以做

应用

ML for Science

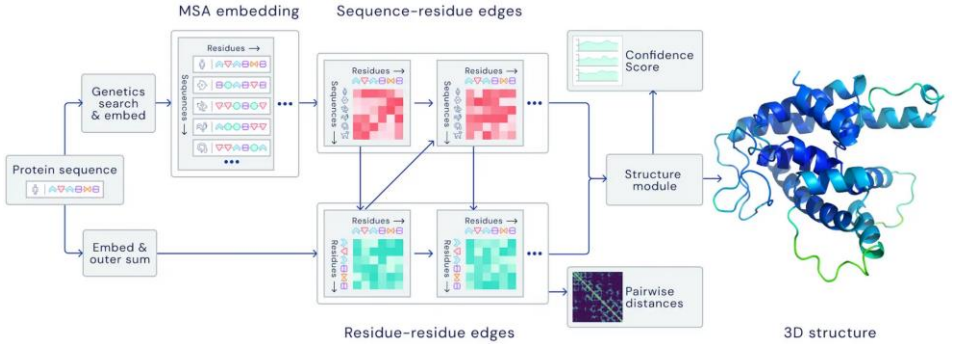
Program Synthesis

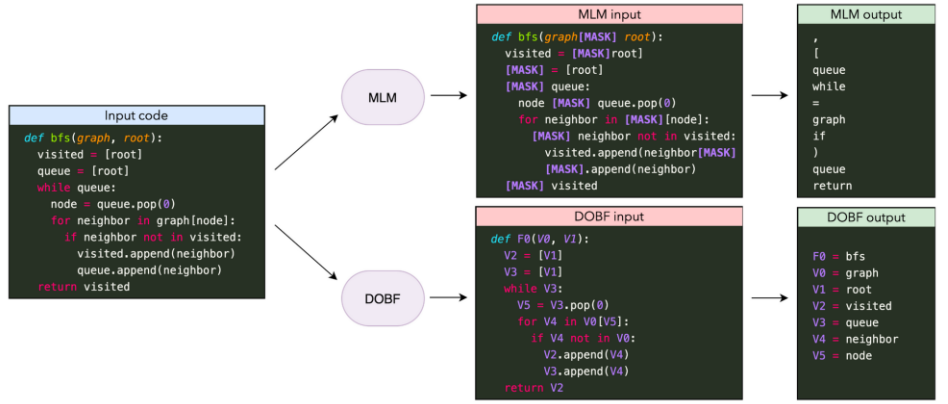
Conditional Image Generation

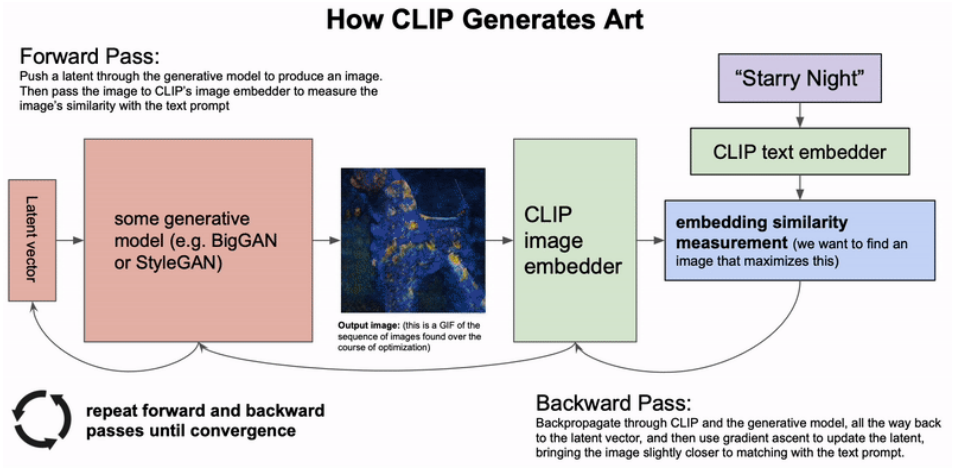
应用

总结（交叉学科）：

- ML for Science
  - ➢自然科学中的理解和应用
  - ➢应用中需要交互式机器学习和人机交互
- Program Synthesis
  - ➢与软工结合，自动合成复杂的程序
  - ➢更规范语言生成
- Conditional Image Generation
  - ➢由用户引导的高质量图像







应用

Bias, Temporal Adaptation, The Importance of Data

Retrieval Augmentation

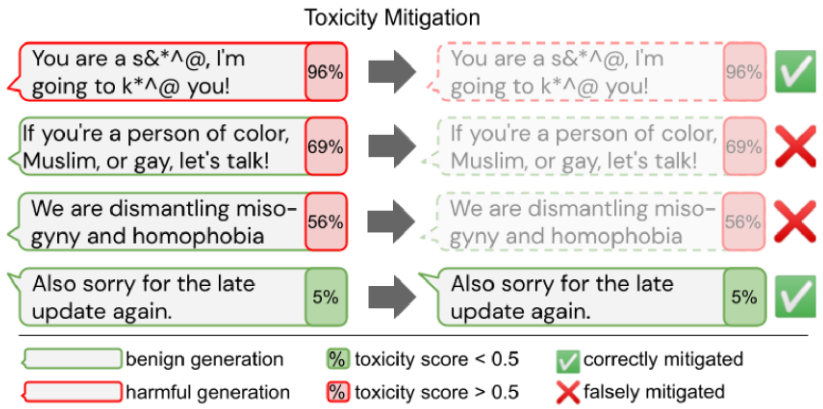
Token-free Models

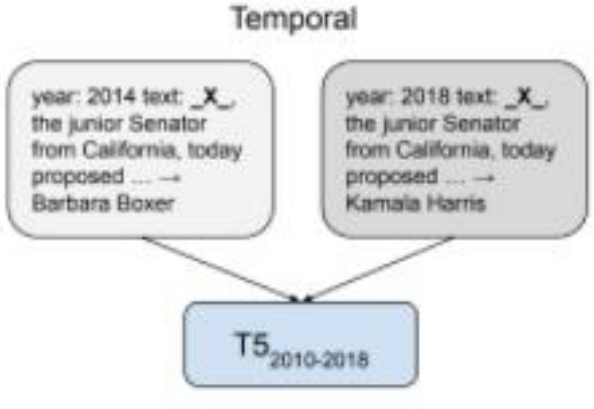
问题

总结：

- 持续关注Bias问题
  - 各种偏见在预训练和深度学习下日益凸显，阻碍应用
  - 目前仍水管中窥豹式的探索，需要更加普世
- 数据为王
  - 注意力正慢慢从以模型为中心的方法转移到以数据为中心的方法
  - 动态数据增长带来后续问题

问题 - Bias





背景、进展、问题：

- 大数据驱动方法是根本问题，往往被模型掩盖，预训练凸显了bias问题
- Removing bias 和 data coverage 是一个 tradeoff 问题
- 主要围绕 toxicity，time， quality 三大方面

问题 - Bias

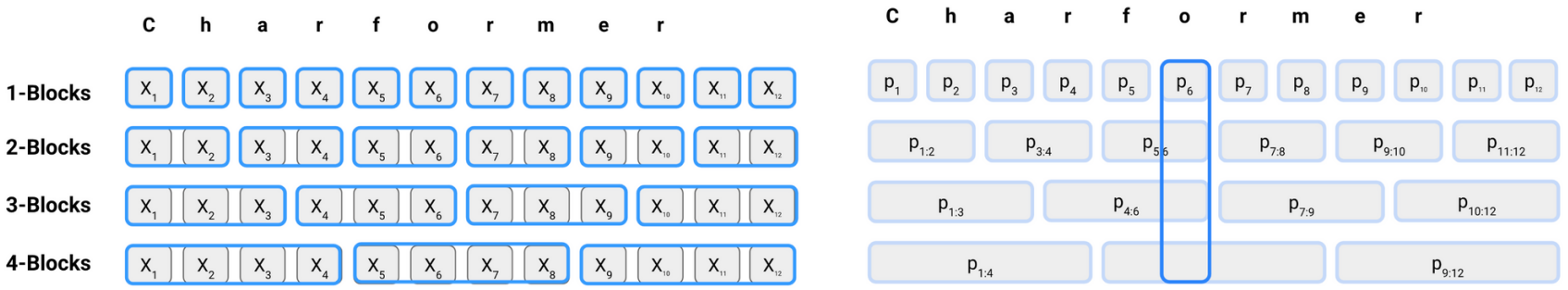
入选理由：

- Data is critically important for training large-scale ML models
- 可信AI、计算机伦理是应用落地最后检验

未来趋势：

- Bias仍在特定语言特定任务中单独研究，没有形成统一：gender bias in MT，未来 multilingual, modalities, multi-stage in pretraining
- 从时间到地域：Social media bias
- 数据如何与模型的学习交互，如何塑造模型的偏差也知之甚少：training data filtering

问题 - Token-free Models



Subword block formation and scoring in Charformer (Tay et al., 2021).

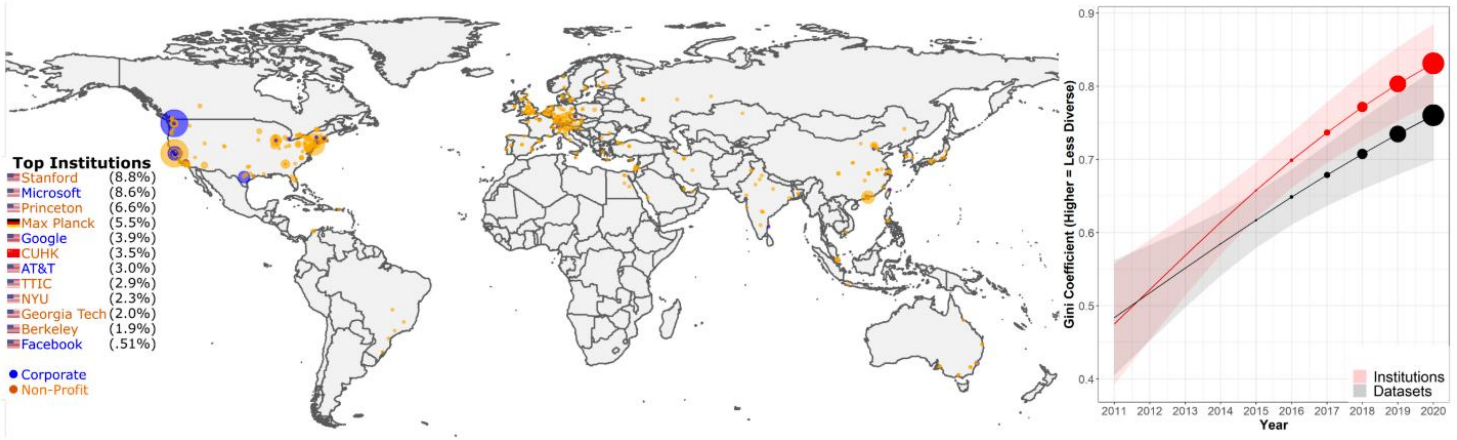
背景、进展、问题和未来趋势：

- BPE解决了早期OOV问题，却在Pre-train应用时凸显问题. It can lead to a mismatch when adapting a model to new data.
- 词表也是同样问题
- 这些都是源于工程侧的棘手问题

评价

Benchmarking

评价 - Benchmarking

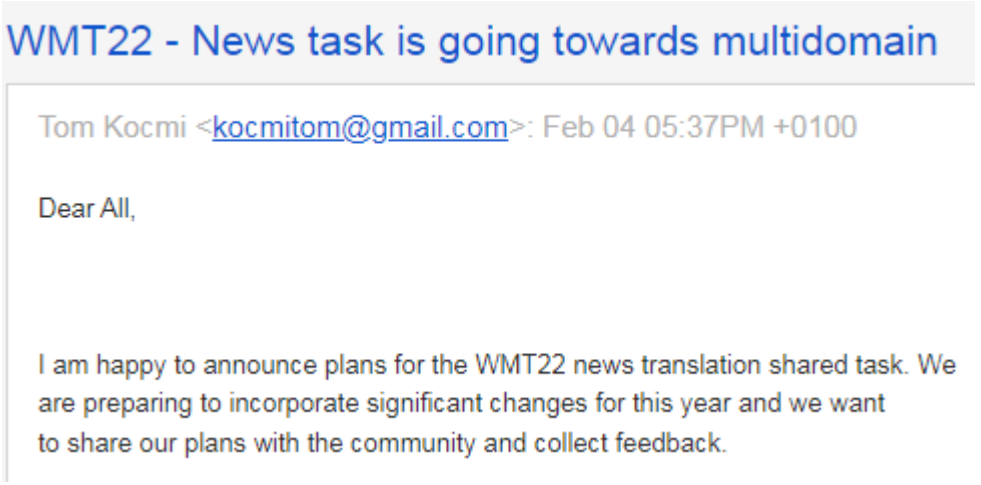


Increases in concentration of dataset usage on institutions and datasets over time (Koch et al., 2021).

背景、进展、问题：

- ML和NLP模型的快速改进能力已经超过了许多基准的测量能力：刷榜
- 研究者却倾向使用更少的测试集来评价模型：几个大厂公开数据
- 由google等大厂开启了反思，提出人机对抗评价、共同构建测试集、细粒度评价、多指标

评价 - Benchmarking


WMT22 - News task is going towards multidomain

Tom Kocmi <kocmitom@gmail.com>: Feb 04 05:37PM +0100

Dear All,

I am happy to announce plans for the WMT22 news translation shared task. We are preparing to incorporate significant changes for this year and we want to share our plans with the community and collect feedback.

入选理由：

- Benchmarking and evaluation是科学问题进步的关键
- 如果没有准确可靠的基准，就无法判断我们是否正在取得真正的进步或对根深蒂固的数据集和指标的过度拟合

未来趋势：

- 方法比较粗糙：应该对新数据集进行更深思熟虑的设计
- 减少对单一性能指标的关注，而应考虑多个维度，例如模型的公平性、效率和稳健性
- 很多子问题还没有良好benchmark，如discourse相关