Deep Learning: An Introduction

Muhammad Abdul-Mageed
Canada Research Chair (NLP and ML)
Director, I Trust AI
The University of British Columbia
Twitter: @mageed

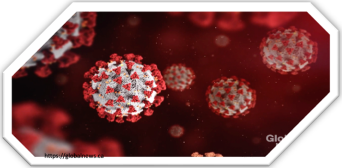
Lanzarote– (2022-10-26)

AI: Art, Nature, and Diversity



Pressing Problems

Misinformation



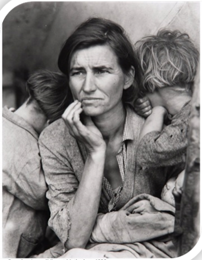
Conflict

Climate Change

The Challenge of **Social Inequality**

Wikimedia Commons

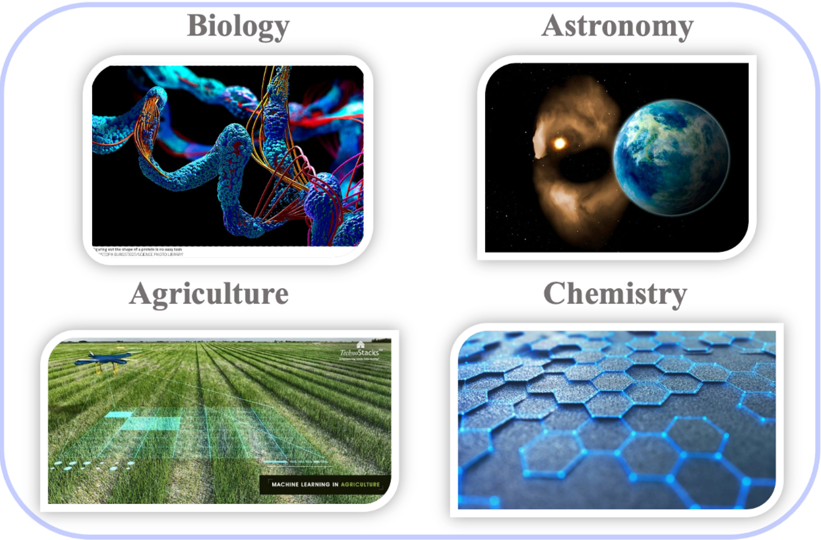
kkrid.com



Dorothea Lange 'Migrant Mother', ca. 1936

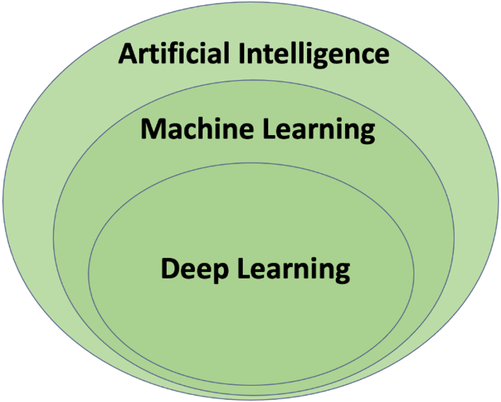
businesseconomics.in

Breakthroughs

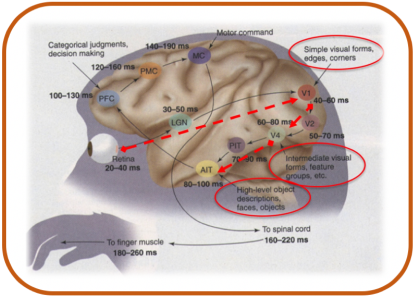


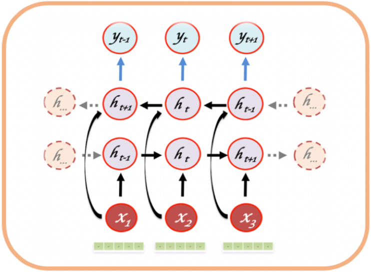
AI & Archives

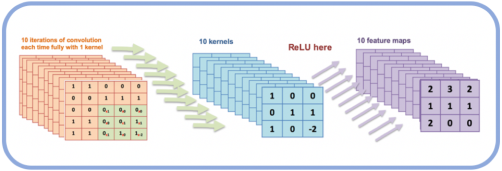
Artificial Intelligence

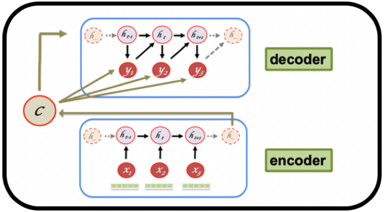


Deep Learning

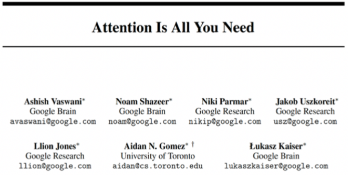






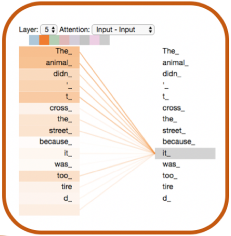


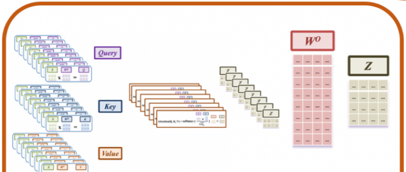
The Transformer



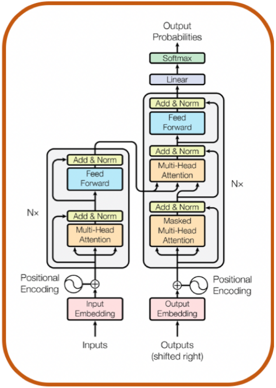
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com





$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

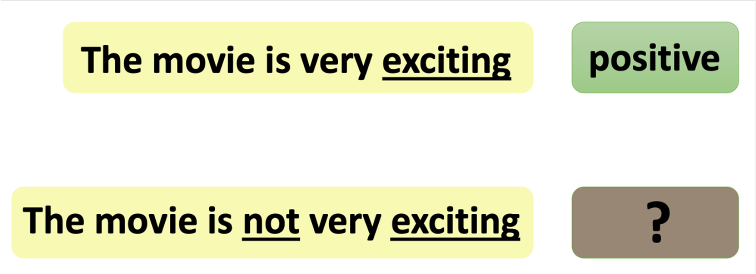


Introductory Example

The movie is very <u>exciting</u> | positive

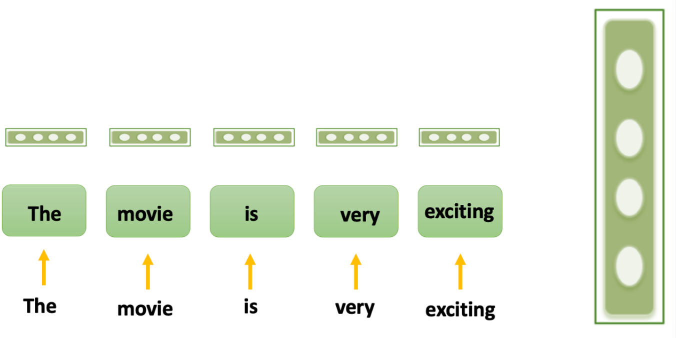
The movie is very <u>boring</u> | negative

Introductory Example *Contd.*

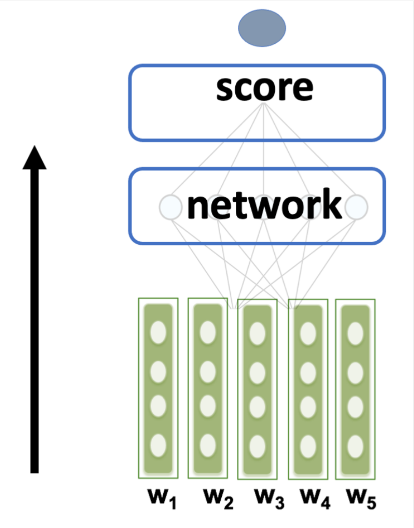


Vectorization

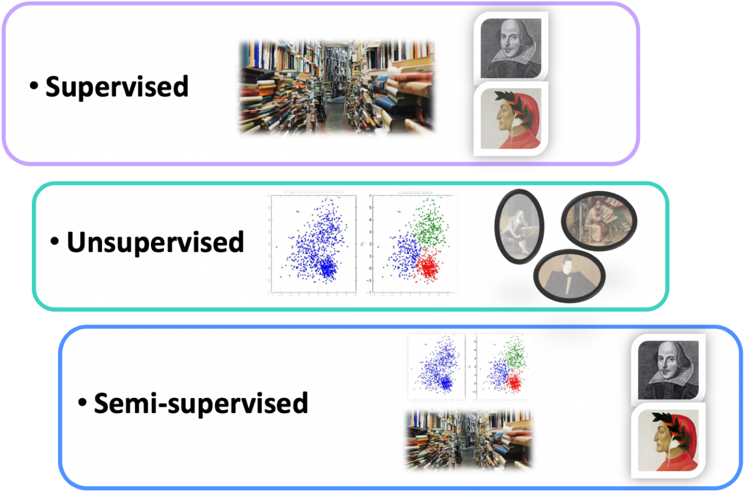
Vectors for Text Classification



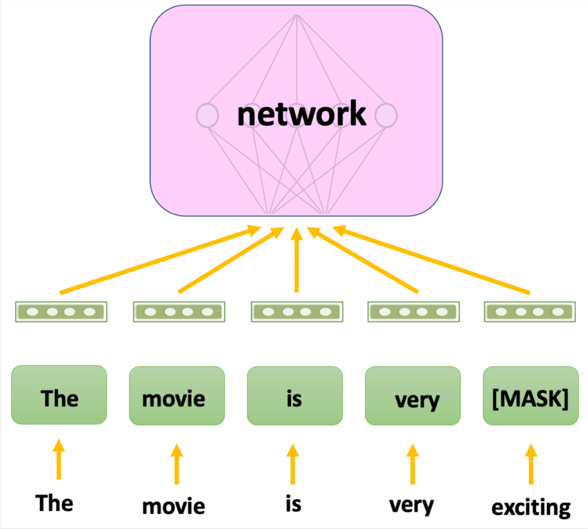
The Network



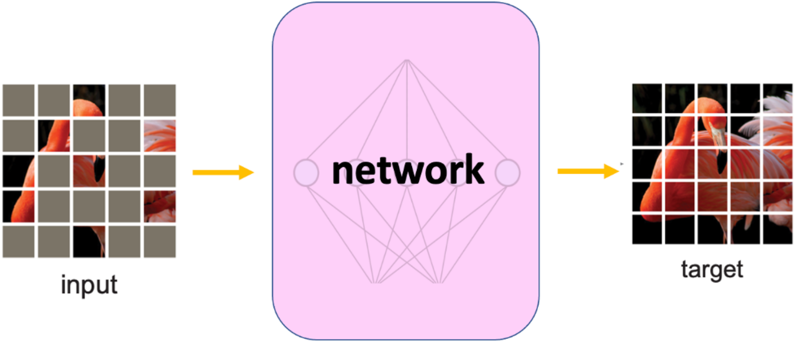
Supervision



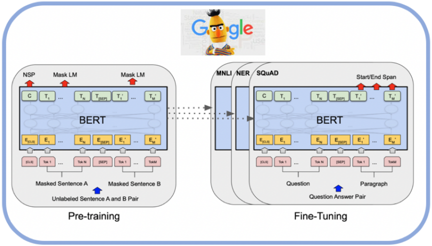
Self-Supervised (Text)

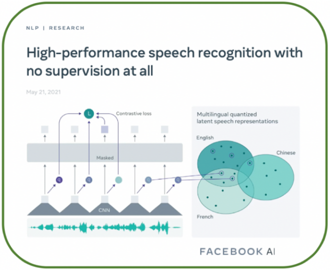


Self-Supervised (Image)

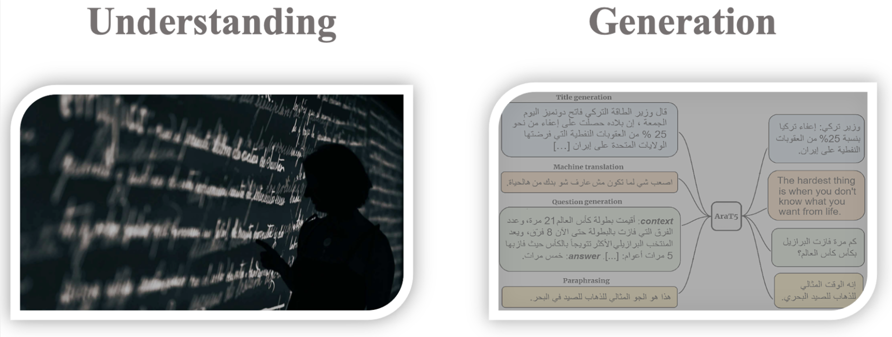


SSL Empowering Models

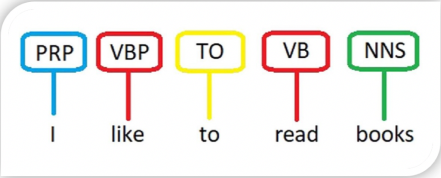




Natural Language Processing



Part of Speech Tagging



CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative

```
>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

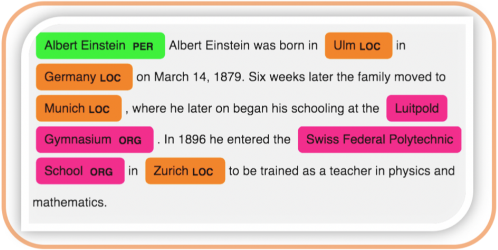
POS Tagging Tutorial

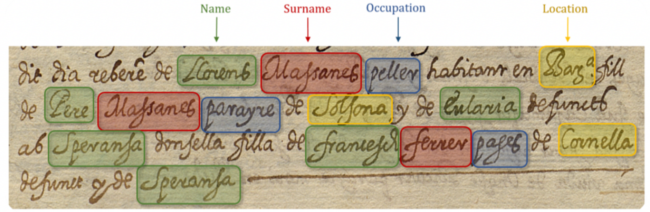
Part of Speech (POS) Tagging

	Category	Descriptions	Link
1	POS Tagging	POS with spaCy	notebook
2	POS Tagging	Train BiLSTM with PyTorch from Scratch	notebook
3	POS Tagging	Finetune with BERT from Scratch	notebook

Figure: [Link]

Named Entity Recognition





(**Esposalles database**, a marriage license book conserved at the Archives of the Cathedral of Barcelona)
Source: https://rrc.cvc.uab.es.

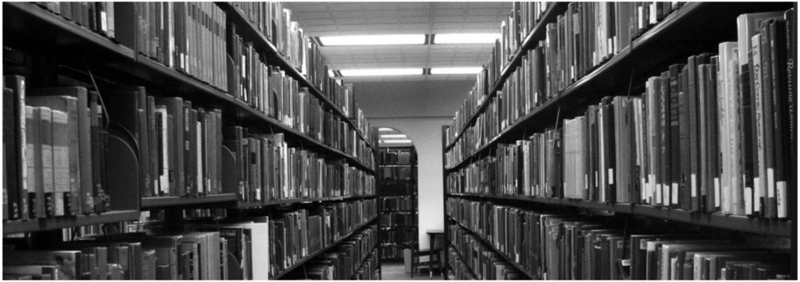
Named Entity Recognition Tutorials

Named Entity Recognition (NER)

	Category	Descriptions	Link
1	Named Entity Recognition	Introduction to NER and out-of-box solution with Spacy	notebook
2	Named Entity Recognition	Train BiLSTM with PyTorch from Scratch	notebook
3	Named Entity Recognition	Fine-tune BERT with Huggingface	notebook

Figure: [Link]

Topic Modeling

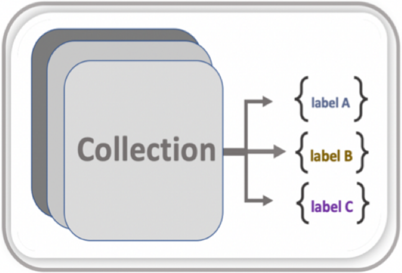


Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

(Slide credit: David Blei)

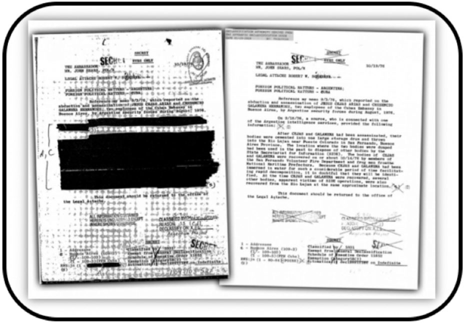
Text Classification



(1) Just got chased through my house with a bowl of tuna fish. ing. **[Disgust]**

(2) I love waiting 2 hours to see 2 min. Of a loved family members part in a dance show #sarcasm **[Sarcastic]**

(3) USER Awww CUPCAKES SUCK IT UP. SHE LOST GET OVER IT **[Offensive]**



Text Classification

Text Classification

Text classification aims to assign a given text to one or more categories. We can find a wide range of real-world applications of text classification, such as spam filtering and sentiment analysis. In this section, two tutorials are included. We discuss what text classification is and solve a classification task in the first tutorial. The second tutorial address a classification task using a Transformer-based deep learning model.

	Category	Descriptions	Link
1	Text Classification	Intro and Classical Machine Learning	notebook
2	Text Classification	Deep Learning (BERT)	notebook

Figure: [Link]

Machine Translation

Facebook's AI Just Set A New Record In Translation And Why It Matters

Hieroglyphics are seen on the sarcophagus once belonging to the judge and prime minister Gemenefherbek of Sais at the Egyptian Museum in Turin, Italy, Tuesday, February 21, 2006. Photographer: Adam Berry/Bloomberg News

The Atlantic

The Shallowness of Google Translate

The program uses state-of-the-art AI techniques, but simple tests show that it's a long way from real understanding.

DOUGLAS HOFSTADTER | JAN 30, 2018 | TECHNOLOGY

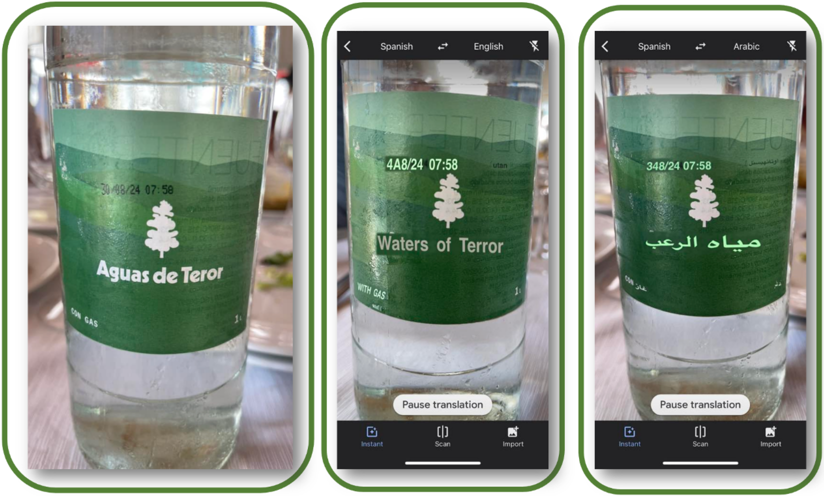
Google

Translate

English | Arabic | French | Translate

يا له من يوم جميل في فانكوفر!	What a lovely day in Vancouver!
صباحنا قشطة!	Morning cream!
الولد ده لسا ضارب كشري.	The boy is a lion.
الولد ده لسا ضارب كشري..	The boy is a loser.

Machine Translation Issues



Low-Resource NMT

IndT5: A Text-to-Text Transformer for 10 Indigenous Languages

El Moatez Billah Nagoudi[1], Wei-Rui Chen[1], Muhammad Abdul-Mageed[1], Hasan Cavusoglu[2]
[1] Natural Language Processing Lab,
[1,2] The University of British Columbia
[1] {moatez.nagoudi,weirui.chen,muhammad.mageed}@ubc.ca, [2] cavusoglu@sauder.ubc.ca

Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation

Michael Przystupa
University of British Columbia
<first>.<last>@gmail.com

Muhammad Abdul-Mageed
University of British Columbia
muhammad.mageed@ubc.ca

Translating the Unseen? Yorùbá→English MT in Low-Resource, Morphologically-Unmarked Settings

Ife Adebara[1,2] Muhammad Abdul-Mageed[1,2] Miikka Silfverberg[1]
Department of Linguistics[1]
Natural Language Processing Lab[2]
The University of British Columbia
{ife.adebara,muhammad.mageed,miikka.silfverberg}@ubc.ca

Code-Switching in NMT

Investigating Code-Mixed Modern Standard Arabic-Egyptian to English Machine Translation

El Moatez Billah Nagoudi   AbdelRahim Elmadany   Muhammad Abdul-Mageed
Natural Language Processing Lab
The University of British Columbia
{moatez.nagoudi,a.elmadany,muhammad.mageed}@ubc.ca

Hard problem

(1) MSAEA	أنا عايز شغل جامد يا جدعان
Human	I want hard work, guys.
Google	I want a rigid job, Jadaan.

Results

Model	Setting	Blue
S2ST	Zero Shot EA	21.34
	Fine-tuned DA	22.51
	Zero Shot EA (true-cased)	23.68
	Fine-tuned DA (true-cased)	**25.72**
mT5	Fine-tuned DA	16.41
	Fine-tuned DA (true-cased)	18.80
mBART	Fine-tuned DA	17.17
	Fine-tuned DA (true-cased)	19.79

System output

Source:	مش عارفين نتأكد و مش عارفين البنات فين.
S2ST	we don't know for sure and the girls don't know finn .
mT5	we can't make sure and we don't know where the girls are
mBART	we don't know where to make sure and we don't know where the girls are

Code-Switched in NMT *Contd.*

Exploring Text-to-Text Transformers for English to Hinglish Machine Translation with Synthetic Code-Mixing

Ganesh Jawahar[1,2] **El Moatez Billah Nagoudi**[1]
Muhammad Abdul-Mageed[1,2] **Laks V.S. Lakshmanan**[2]
Natural Language Processing Lab[1]
Department of Computer Science[2]
The University of British Columbia
ganeshjwhr@gmail.com, {moatez.nagoudi, muhammad.mageed}@ubc.ca, laks@cs.ubc.ca

Hinglish to English translation (Dhar et al. (2018), Srivastava and Singh (2020))

Hinglish: Hi there! Chat ke liye ready ho? → **English**: Hi there! Ready to chat?

English to Hinglish translation (our task)

English: Maybe it's to teach kids to challenge themselves → **Hinglish**: maybe kida ko teach karna unka challenge ho saktha hein

Code-Switched in NMT *Contd.*

1st substitution

it was that good (English)

it was that achi (Hinglish)

2nd substitution

it was that good (English)

ye was that achi (Hinglish)

English (Gold): And they grow apart. She is the protector of the Moors forest.
Hinglish (Prediction): Aur wo apart grow karte hai. Wo Moors forest ka (ki) protector hai.

English (Gold): I watched it at least twice.. it was that good. I love female superheros
Hinglish (Prediction): Maine ise kam se kam ek (do) baar dekha hai. Ye itni achi thi. Mujhe female superheros pasand hai.

English (Gold): I loved the story & how true they made it in how teen girls act but I honestly don't know why I didn't rate it highly as all the right ingredients were there. I cannot believe it was 2004 it was released though, 14 years ago!
Hinglish (Prediction): mujhe story bahut pasand aaya aur teen girls ka act kaise hota lekin main honestly nahi janta kyon ki main ise highly rate nahi kar raha tha kyunki sahi ingredients wahan they. mujhe yakin nahi hota ki 2004 mein release huyi thi, 14 saal pehle!

cs method	BLEU
baseline (mBART model)	11.00
LinCE leaderboard (only best results)	
LTRC Team	12.22
IITP-MT Team	10.09
CMMTOne Team	2.58
Romanization	
OPUS	12.38
Paraphrasing	
Para	12.1
Backtranslation	
Backward model	11.47
Social media	
PHINC	11.9
Equivalence constraint theory	
ECT (100K)	12.45
CMDR (ours)	
CMDR-unigram (roman)	12.25
CMDR-bigram (native)	12.63
CMDR-bigram (roman)	12.08
CMDR-trigram (native)	**12.67**
CMDR-trigram (roman)	12.05
Method Combinations	
CMDR-unigram (roman) + PHINC	11.58
ECT (100K) + CMDR-trigram (native)	12.27

Open-Source NMT

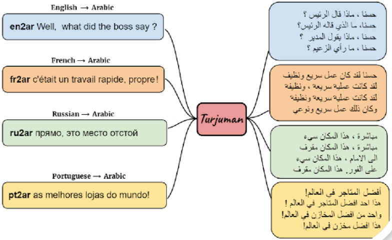
TURJUMAN:
A Public Toolkit for Neural Arabic Machine Translation

El Moatez Billah Nagoudi*   AbdelRahim Elmadany*   Muhammad Abdul-Mageed*
Deep Learning & Natural Language Processing Group
The University of British Columbia
{moatez.nagoudi,a.elmadany,muhammad.mageed}@ubc.ca

Turjuman

Turjuman is a neural machine translation toolkit. It translates from 20 languages into Modern Standard Arabic (MSA). Turjuman is described in this paper: **"TURJUMAN: A Public Toolkit for Neural Arabic Machine Translation".**

Turjuman exploits our **AraT5 model**. This endows Turjuman with a powerful ability to decode into Arabic. The toolkit offers the possibility of employing a number of diverse decoding methods, making it suited for acquiring paraphrases for the MSA translations as an added value.



Turjuman Awarded Best Paper

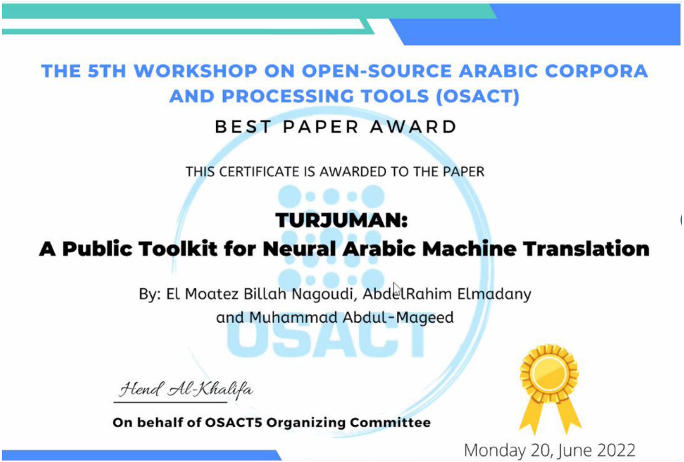


Figure: [Demo]

MT Tutorial

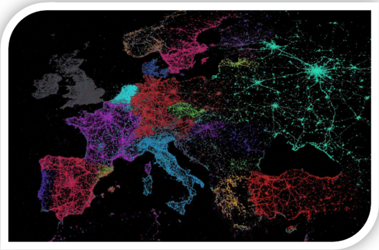
Machine Translation (MT)

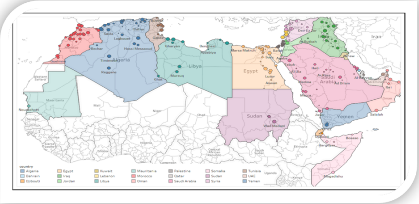
Machine Translation aims to learn a automatic system to translate a given text from a language to another language. This section includes a tutorial of neural-based machine translation. We introduce a important architecture in machine translation: sequence to sequence network, in which two recurrent neural networks work together to transform one sequence (e.g., sentence) to another.

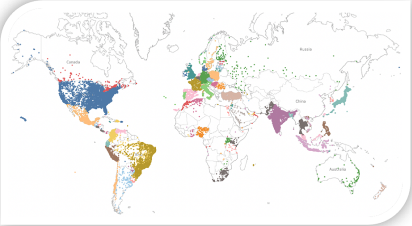
	Category	Descriptions	Link
1	Machine Translation	Seq2seq	notebook

Figure: [Link]

Multilinguality







Afrocentric NLP

Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go

Ife Adebara
Deep Learning and
Natural Language Processing Group
The University of British Columbia
ife.adebara@ubc.ca

Muhammad Abdul-Mageed
Deep Learning and
Natural Language Processing Group
The University of British Columbia
muhammad.mageed@ubc.ca

Abstract

Aligning with ACL 2022 special Theme on "Language Diversity: from Low Resource to Endangered Languages", we discuss the major linguistic and sociopolitical challenges facing development of NLP technologies for African languages. Situating African languages in a typological framework, we discuss how the particulars of these languages can be harnessed. To facilitate future research, we also highlight current efforts, communities, venues, datasets, and tools. Our main objective is to motivate and advocate for an Afrocentric approach to technology develop-

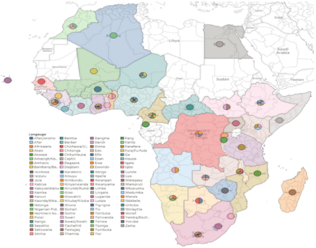
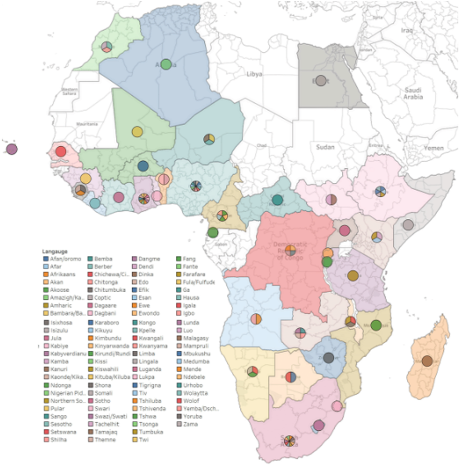


Figure 1: African languages discussed in this paper. A high quality version is in Figure F.1 (Appendix).

Afrocentric NLP

~94 African Languages

- Niger-Congo
- Afro Asiatic
- Nilo-Saharan
- Creole
- Indo-European
- Austronesian



The State & Fate of African Languages

- **left-behinds** - probably impossible to build resources for them
- **scraping-bys** - **no** labelled datasets
- **hopefuls** – **few** labeled datasets, researchers, and language support communities
- **rising-stars** - strong web presence but **insufficient** labeled data collection

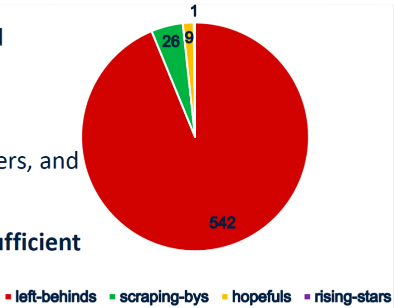
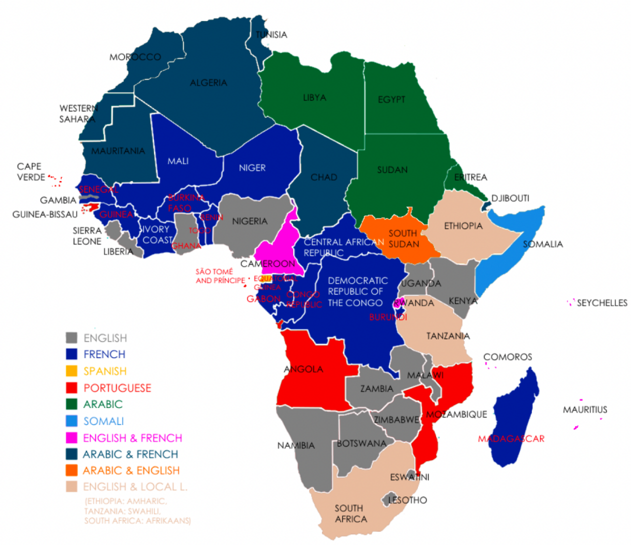


Figure: [Joshi, et al., 2020]

Language Policy (e.g., Main Business Languages)



African Language Identification

AfroLID: A Neural Language Identification Tool for African Languages

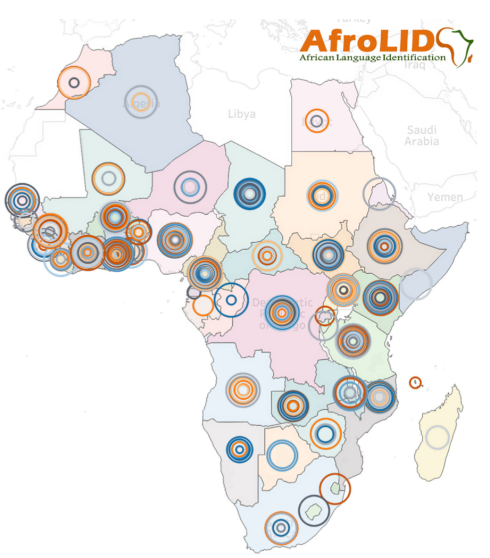
Ife Adebara* AbdelRahim Elmadany* Muhammad Abdul-Mageed Alcides Alcoba Inciarte
Deep Learning & Natural Language Processing Group
The University of British Columbia
{ife.adebara@,a.elmadany@,muhammad.mageed@,alcobaaj@mail.}ubc.ca

Abstract

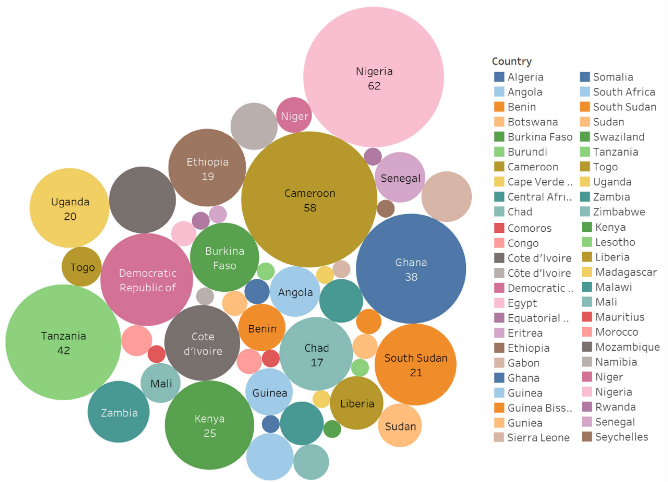
Language identification (LID) is a crucial precursor for NLP, especially for mining web data. Problematically, most of the world's 7000+ languages today are not covered by LID technologies. We address this pressing issue for Africa by introducing AfroLID, a neural LID toolkit for 517 African languages and varieties. AfroLID exploits a multi-domain web dataset manually curated from across 14 language families utilizing five orthographic systems. When evaluated on our blind Test set, AfroLID achieves 95.89 F_1-score. We also compare AfroLID to five existing LID tools that each cover a small number of African languages, finding it to outperform them on most languages. We further show the utility of AfroLID in the wild by testing it on the acutely under-served Twitter domain. Finally, we offer a number of controlled case studies and perform a linguistically-motivated error analysis that allow us to both showcase AfroLID's powerful capabilities and limitations.[1]

Figure 1: All 50 African countries in our data, with our 517 languages/language varieties in colored circles overlayed within respective countries. More details are in Appendix E.

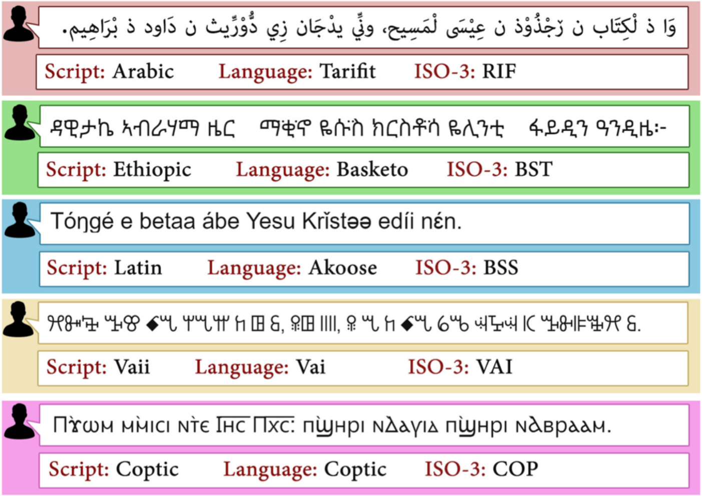
AfroLID Coverage



Languages x Country

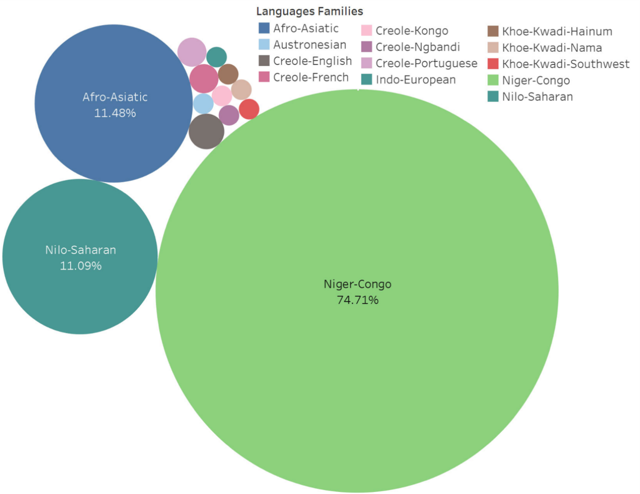


Scripts



وَا ذْ لْكِتَابْ نْ زّجْذُوْذْ نْ عِيْسَى لْمَسِيحْ، وْنِّي يِدْجَانْ زِي دُّوْرِّيثْ نْ دَاوِدْ ذْ بْرَاهِيمْ.
Script: Arabic Language: Tarifit ISO-3: RIF
ዳዊታኬ ኣብራሃማ ዜር ማቂኖ ዬሱ̈ስ ክርስቶሳ ዬሊንቲ ፋይዲን ዓንዲዜ፡-
Script: Ethiopic Language: Basketo ISO-3: BST
Tóŋgé e betaa ábe Yesu Krǐstəə edíi nɛ́n.
Script: Latin Language: Akoose ISO-3: BSS
Script: Vaii Language: Vai ISO-3: VAI
Ⲡϫⲱⲙ ⲙ̀ⲙⲓⲥⲓ ⲛ̀ⲧⲉ Ⲓⲏⲥ Ⲡⲭⲥ: ⲡ̀ϣⲏⲣⲓ ⲛ̀Ⲇⲁⲩⲓⲇ ⲡ̀ϣⲏⲣⲓ ⲛ̀Ⲁⲃⲣⲁⲁⲙ.
Script: Coptic Language: Coptic ISO-3: COP

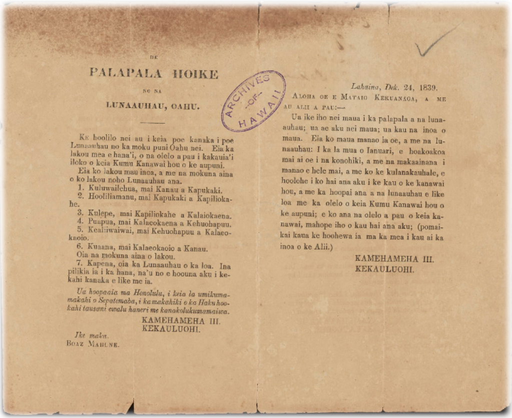
Language Families



OCR & HWR


بسم الله الرحمن الرحيم
الم ذلك الكتاب لا ريب فيه هدى
للمتقين الذين يؤمنون بالغيب
ويقيمون الصلوة ومما رزقنا هم
ينفقون والذين يؤمنون
بما أنزل إليك وما أنزل من قبلك
وبالآخرة هم يوقنون

Multilingual OCR (Hawaii Languages)



HE

PALAPALA HOIKE

NO NA

LUNAAUHAU, OAHU.

Ke hoolilo nei au i keia poe kanaka i poe Lunaauhau no ka moku puni Oahu nei. Eia ka lakou mea e hana'i, o na olelo a pau i kakauia'i iloko o keia Kumu Kanawai hou o ke aupuni.

Eia ko lakou mau inoa, a me na mokuna aina o ko lakou noho Lunaauhau ana.

1. Kuluwailehua, mai Kanau a Kapukaki.
2. Hooliliamanu, mai Kapukaki a Kapiliokahe.
3. Kulepe, mai Kapiliokahe a Kalaiokaena.
4. Puapua, mai Kalaeokaena a Kehuohapuu.
5. Kealiiwaiwai, mai Kehuohapuu a Kalaeokaoio.
6. Kuaana, mai Kalaeokaoio a Kanau.

Oia na mokuna aina o lakou.

7. Kapena, oia ka Lunaauhau o ka loa. Ina pilikia ia i ka hana, na'u no e hoouna aku i kekahi kanaka e like me ia.

Ua hoopaaia ma Honolulu, i keia la umikumamakahi o Sepatemaba, i ka makahiki o ka Haku hookahi tausani ewalu haneri me kanakolukumamaiwa.

KAMEHAMEHA III.
KEKAULUOHI.

Ike maka.
Boaz Mahune.

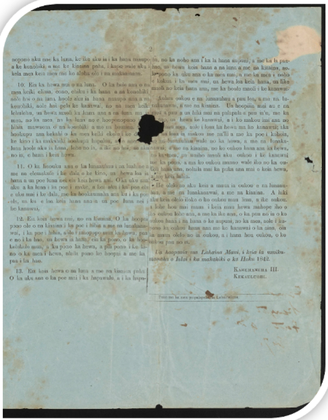
ARCHIVES -OF- HAWAII

Lahaina, Dek. 24, 1839.

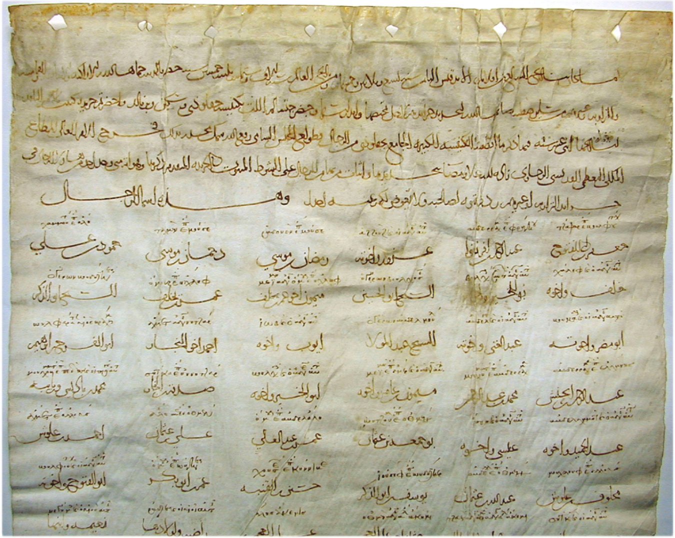
Aloha oe e Mataio Kekuanaoa, a me au alii a pau:—

Ua ike iho nei maua i ka palapala a na lunaauhau; ua ae aku nei maua; ua kau na inoa o maua. Eia ko maua manao ia oe, a me na lunaauhau: I ka la mua o Ianuari, e hoakoakoa mai ai oe i na konohiki, a me na makaainana i manao e hele mai, a me ko ke kulanakauhale, e hoolohe i ko hai ana aku i ke kau o ke kanawai hou, a me ka hoopai ana a na lunaauhau e like loa me ka olelo o keia Kumu Kanawai hou o ke aupuni; e ko ana na olelo a pau o keia kanawai, mahope iho o kau hai ana aku; (pomaikai kaua ke hoohewa ia ma ka mea i kau ai ka inoa o ke Alii.)

KAMEHAMEHA III.
KEKAULUOHI.

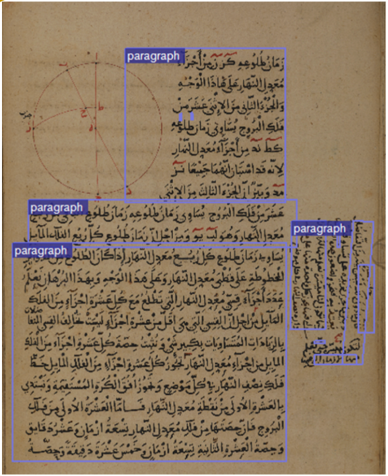


Multilingual Parchments



Layout Analysis

Layout Analysis



https://blogs.bl.uk/digital-scholarship/2019/09/rasm2019-results.htm

Generation for OCR and HWR

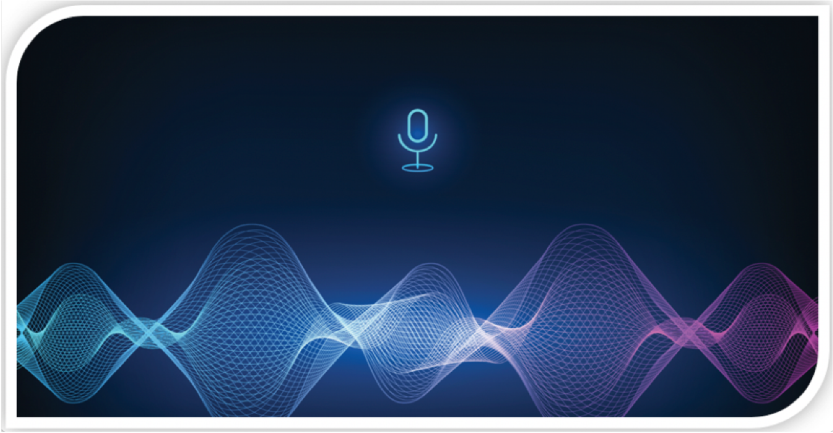
OCR Tutorial

OCR

- Tesseract OCR Tutorial
- TrOCR Finetuning and Inference Tutorial

Figure: [Link]

Voice Technologies



Whisper

Robust Speech Recognition via Large-Scale Weak Supervision

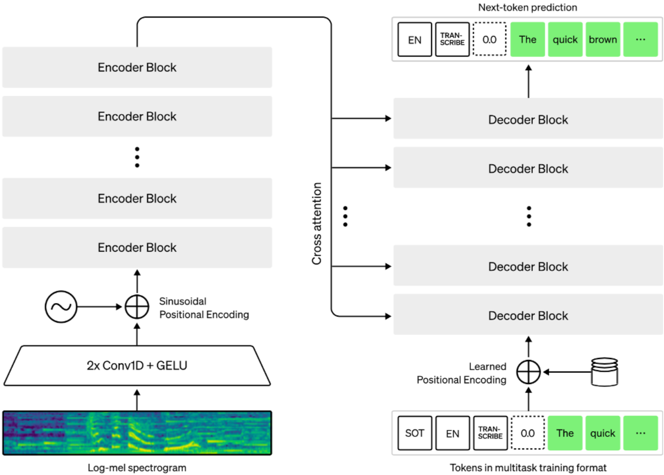
Alec Radford[*,1] Jong Wook Kim[*,1] Tao Xu[1] Greg Brockman[1] Christine McLeavey[1] Ilya Sutskever[1]

Abstract

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

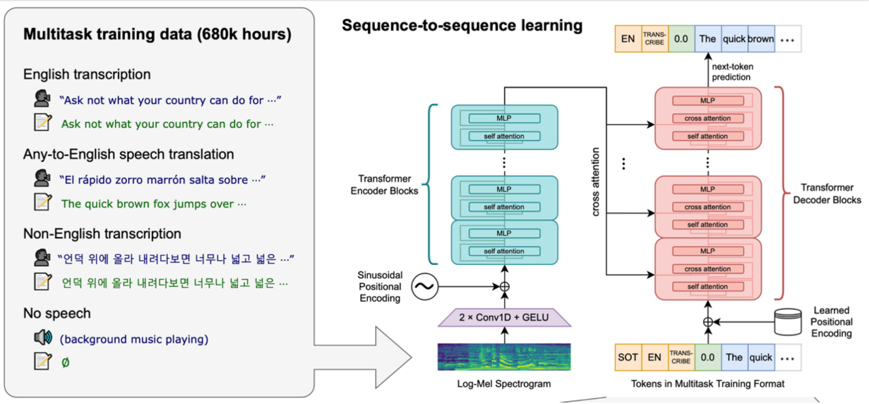
methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

Whisper



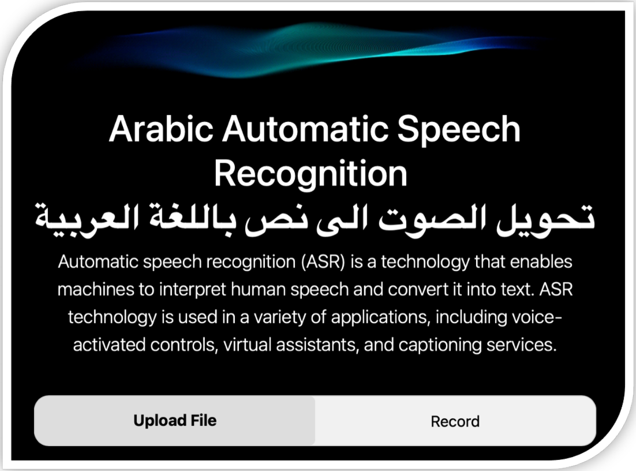
Next-token prediction
EN
TRAN-SCRIBE
0.0
The
quick
brown
...
Encoder Block
Encoder Block
Encoder Block
Encoder Block
Decoder Block
Decoder Block
Decoder Block
Decoder Block
Cross attention
Sinusoidal Positional Encoding
2x Conv1D + GELU
Log-mel spectrogram
Learned Positional Encoding
SOT
EN
TRAN-SCRIBE
0.0
The
quick
...
Tokens in multitask training format

Whisper



Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
cross attention
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
Log-Mel Spectrogram
SOT
Tokens in Multitask Training Format

Arabic ASR (Demo)



One Model to Pronounce Them All: Multilingual Grapheme-to-Phoneme Conversion With a Transformer Ensemble

Kaili Vesik[1,2], Muhammad Abdul-Mageed[1,2,3], Miikka Silfverberg[2]

Language	Source	Target (IPA)
Alphabet:		
arm	ահեղ	ɑ h ɛ ʁ
	լիարժեք	l j ɑ ɾ ʒ ɛ k^h
fre	front	f ʁ ɔ̃
	vêtu	v e t y
Alphasyllabary:		
hin	दिखावा	d ɪ k^h ɑː ʋ ɑː
	हटना	ɦ ə ʈ n ɑː
kor	개벽	k e̝ b j ʌ̹ k̚
	오빠	o̝ p͈ a̠
Syllabary:		
jpn	いなり	i n a̠ ɾ^j i
	やせん	j a̠ s ẽ̝ ɴ

Table 1: Sample pairs from training data

	Multilingual		Self-trained	
Lang	WER	PER	WER	PER
ady	28.44	6.46	29.11	6.46
arm	13.11	2.98	12.89	3.07
bul	27.11	5.91	30.89	6.92
dut	15.78	2.98	16.89	3.07
fre	5.33	1.24	5.78	1.36
geo	26.00	5.25	26.67	5.23
gre	16.67	2.68	15.78	2.60
hin	6.44	1.58	6.67	1.66
hun	4.67	1.05	4.22	0.98
ice	9.56	2.11	9.11	1.83
jpn	6.00	1.44	6.00	1.40
kor	32.22	8.54	32.44	8.86
lit	19.33	3.63	20.00	3.68
rum	9.33	1.96	10.44	2.23
vie	4.89	1.66	4.00	1.28
avg	14.99	3.30	15.39	3.37

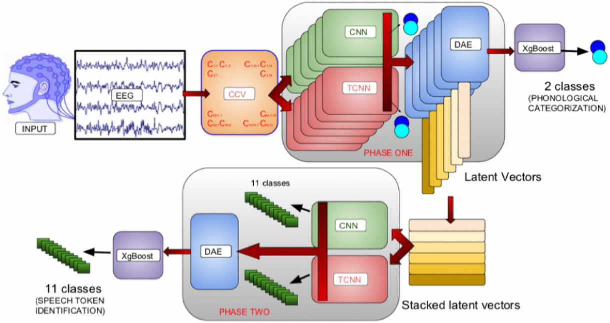
Table 6: Blind test set results for *fully-supervised multilingual* and *self-trained multilingual* models.

Imagined Speech Recognition

SPEAK YOUR MIND!
Towards Imagined Speech Recognition With Hierarchical Deep Learning

Pramit Saha[1], *Muhammad Abdul-Mageed*[2], *Sidney Fels*[1]

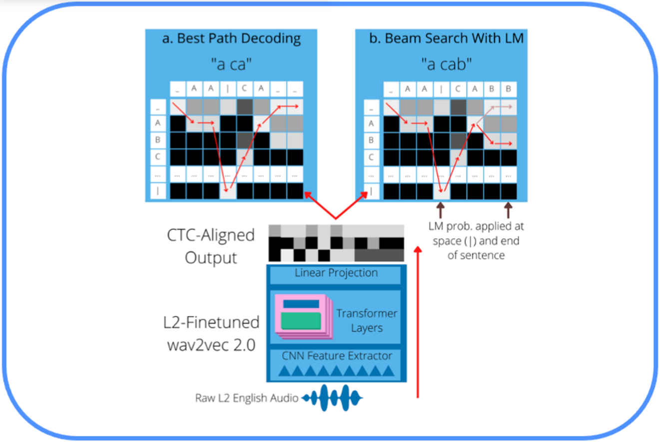
[1]Human Communication Technologies Lab, University of British Columbia [2]Natural Language Processing Lab, University of British Columbia
pramit@ece.ubc.ca, muhammad.mageed@ubc.ca, ssfels@ece.ubc.ca



Automatic Speech Recognition for L2 English

Improving Automatic Speech Recognition for Non-Native English with Transfer Learning and Language Model Decoding

Peter Sullivan, Toshiko Shibano, Muhammad Abdul-Mageed



Vision Transformers

Published as a conference paper at ICLR 2021

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

ViT

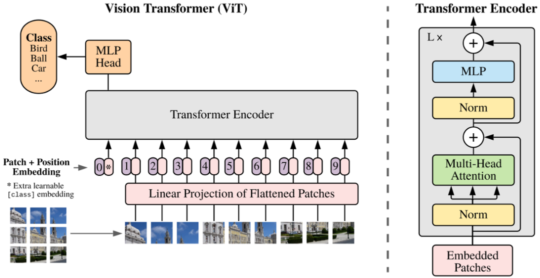


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

BEiT: BERT Pre-Training of Image Transformers

Hangbo Bao[†,*] **Li Dong**[‡], **Songhao Piao**[†], **Furu Wei**[‡]
† Harbin Institute of Technology
‡ Microsoft Research
https://aka.ms/beit

Abstract

We introduce a self-supervised vision representation model **BEiT**, which stands for **B**idirectional **E**ncoder representation from **I**mage **T**ransformers. Following BERT [DCLT19] developed in the natural language processing area, we propose a *masked image modeling* task to pretrain vision Transformers. Specifically, each image has two views in our pre-training, i.e., image patches (such as 16×16 pixels), and visual tokens (i.e., discrete tokens). We first "tokenize" the original image into visual tokens. Then we randomly mask some image patches and fed them into the backbone Transformer. The pre-training objective is to recover the original visual tokens based on the corrupted image patches. After pre-training BEiT, we directly fine-tune the model parameters on downstream tasks by appending task layers upon the pretrained encoder. Experimental results on image classification and semantic segmentation show that our model achieves competitive results with previous pre-training methods.

BEiT

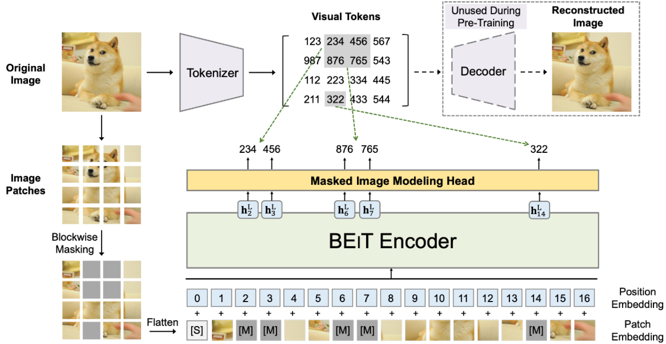


Figure 1: Overview of BEIT pre-training. Before pre-training, we learn an "image tokenizer" via autoencoding-style reconstruction, where an image is tokenized into discrete visual tokens according to the learned vocabulary. During pre-training, each image has two views, i.e., image patches, and visual tokens. We randomly mask some proportion of image patches (gray patches in the figure) and replace them with a special mask embedding [M]. Then the patches are fed to a backbone vision Transformer. The pre-training task aims at predicting the visual tokens of the *original* image based on the encoding vectors of the *corrupted* image.

MAE-ViT

Masked Autoencoders Are Scalable Vision Learners

Kaiming He[*,†] Xinlei Chen[*] Saining Xie Yanghao Li Piotr Dollár Ross Girshick

[*]equal technical contribution [†]project lead

Facebook AI Research (FAIR)

Abstract

This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric *encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image,* e.g., *75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3× or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well:* e.g., *a vanilla ViT-Huge model achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. Transfer performance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior.*

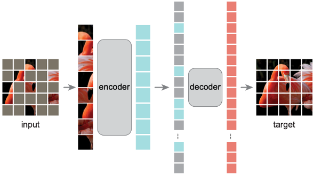


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

in vision [59, 46] preceded BERT. However, despite significant interest in this idea following the success of BERT.

MAE-ViT

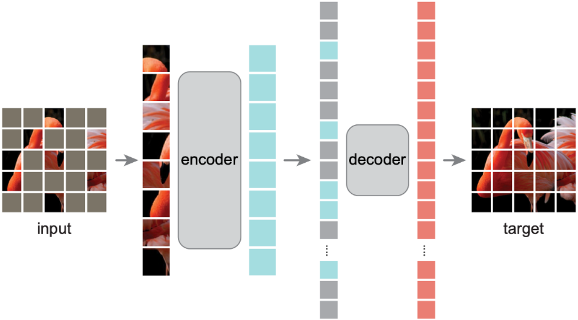
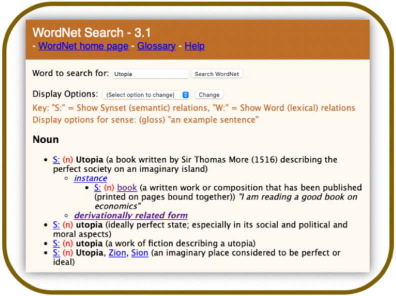


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Visual Object Recognition



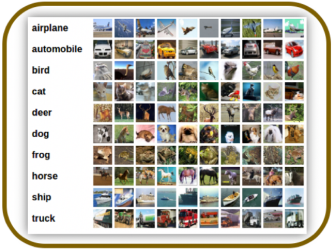
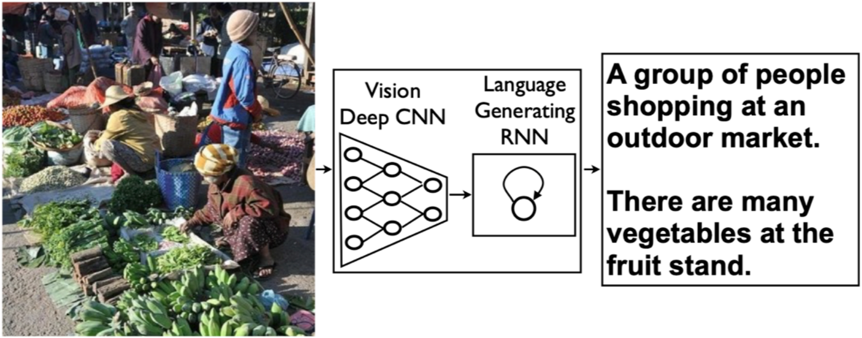


Image Captioning



(Vinyals et al., 2015)

Descriptions of Visual Archives

(Google image search)

Image Captioning

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

(Xu et al., 2016)

Museum Image Captioning

Captions: (1) Lion-Shaped Furniture Leg. (2) In both Egypt and Nubia the lion was associated with the sun god and symbolized royalty.

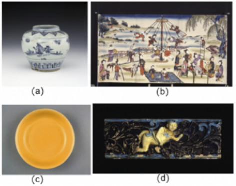


Figure 3: Four ancient Chinese artwork images.

Images	(a)	(b)	(c)	(d)
Artwork type	Jar	Print	Dish	Wall-tile
Model NIC	Porcelain bowl with underglaze blue decoration	Woodcut	There is a mark in underglaze blue on the base	There is an inscription on the base
Model SA	Porcelain jar with underglaze blue decoration	Woodcut	Porcelain dish with rounded sides	Made of blue glazed porcelain
Model LSTM-A_5-MC	This ovoid jar has a short neck with a thickened rim and a recessed base	Woodcut	Porcelain dish with rounded sides	Made of blue glazed porcelain
Model LSTM-MC-$OUT_{dynamic}$	Porcelain jar with underglaze blue decoration	Ink and colours on paper	Porcelain dish with rounded sides	Earthenware wall tile with moulded decoration and green glaze

Figure 4: Generated captions for the images in Figure 4 by model NIC, SA, LSTM-A_5-MC and LSTM-MC-$OUT_{dynamic}$.

(a)	Porcelain jar with underglaze blue decoration. This jar has a narrow raised neck, rounded body and tapering foot. On the unglazed base the spiral marks resulting from turning the jar on the potter's wheel are clearly visible. On the outside an old man is shown in a landscape on one side and another, walking with a servant carrying a 'qin' wrapped in a textile cloth, on the other side; they are separated by plants and rocks. Around the neck is a double ring with banana leaves forming a collar; stylized lappets, like vertical long and short stripes, surround the foot. The cobalt blue appears black where it has not been covered by the blue-green glaze. Originally the jar would have had a domed cover with lotus-bud finial
(b)	Woodcut. Recreation. Acrobats and jugglers and women playing musical instruments. Printed in ink, colour on paper
(c)	Porcelain dish has yellow enamel over thin plain felspathic glaze. There is an inscription on the base, which is glazed
(d)	Wall tile with relief moulded decoration and 'fahua'-palette glazes. This 'fahua'-palette tile is moulded and carved in relief with a celestial boy flying in a contorted pose through scrolling foliage. He is naked except for a torque necklace with three pierced beads. He holds the stem of lotus foliage in his right hand and raises the other hand. The tiny curl in the centre of his forehead indicates his youth. The turquoise glaze is fugitive

Figure 5: Ground-truth captions for the images in Figure 3.

(Sheng & Moens, 2019)


Better Language Models and Their Implications
We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

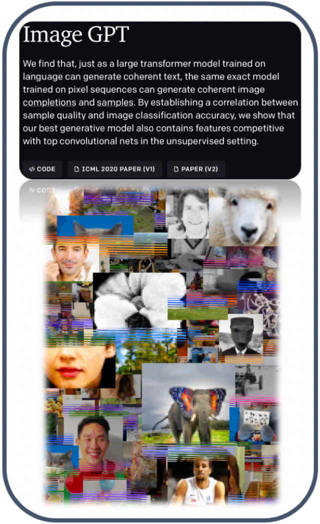
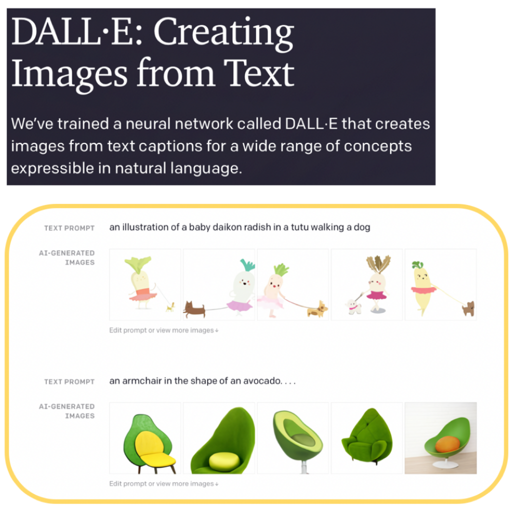
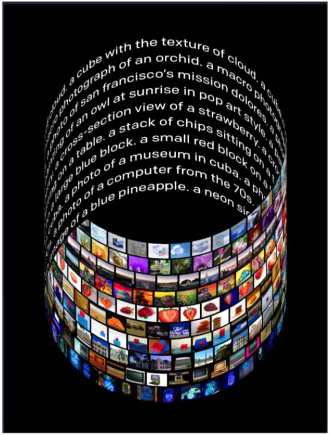
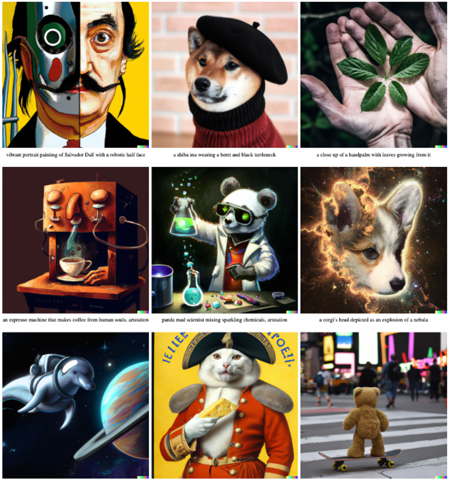


Image GPT
We find that, just as a large transformer model trained on language can generate coherent text, the same exact model trained on pixel sequences can generate coherent image completions and samples. By establishing a correlation between sample quality and image classification accuracy, we show that our best generative model also contains features competitive with top convolutional nets in the unsupervised setting.
CODE
ICML 2020 PAPER (V1)
PAPER (V2)

Generative Deep Learning *Contd.*







vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Figure 1: Selected 1024×1024 samples from a production version of our model.

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion.

Diffusion Models

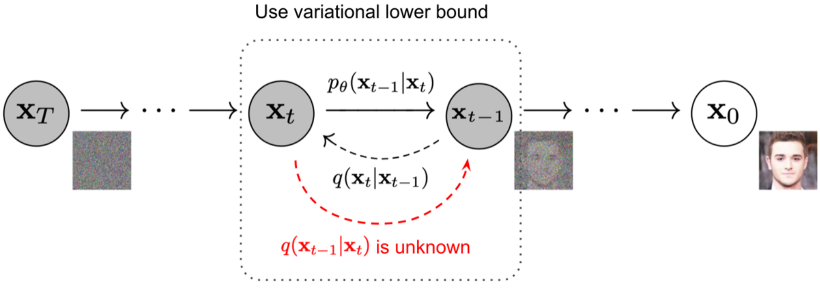


Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. 2020 with a few additional annotations)

Sample Generations

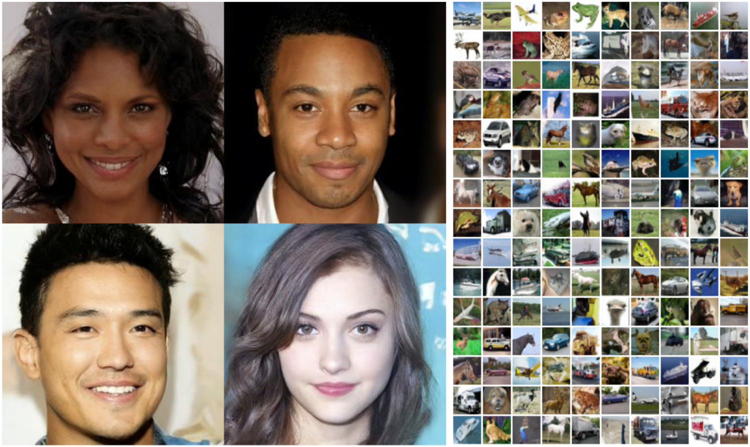


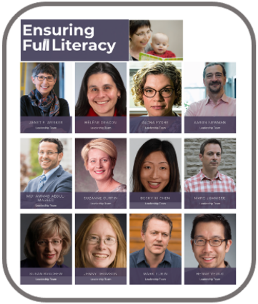
Figure 1: Generated samples on CelebA-HQ 256 $\times$ 256 (left) and unconditional CIFAR10 (right)

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

learnera.ai


LEARNERA.AI
Home
Workshops
UBC Deep Learning
& NLP Group
Workshops

Partnerships



Social Sciences and
Humanities Research
Council of Canada

Canada

Acknolwedgements

Natural Sciences and Engineering Research Council of Canada
Canada

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

Google

Canada Research Chairs
Chaires de recherche du Canada
Canada

Penn

INDIANA UNIVERSITY

UBC ARTS

COLUMBIA UNIVERSITY

Microsoft Research

compute canada

UBC Language Sciences

AMD

aws

Collaborators