Fake News detection
in press articles

Why?

- *"Infostrateg"* / NCBiR 2021
  - Fake News

- MLP-Mixer architecture advantages
  - Unlike RNNs, mixer operates in parallel on entire sequence
  - Unlike transformers, their complexity scales linearly with the sequence length and yet they can capture long-range dependencies
  - Conversly to CNN, MLP is not local as filter spans entire spacial area
  - Based on MLP blocks for which hardware acceleration is ubiquitous
  - much less number of parameters compared to CNN and transformer-based models, which leads to less training and serving computational cost

- Mixer for NLP task
  - Sequence text model to detect fake news that scales well
  - Comparison to CNN (the most frequent referred)

Fake News

FAKE NEWS:

- „False stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke"

  Cambridge Dictionary

- „False or misleading information presented as news. Fake news often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. […] first used in 1890s when sensational reports in newspapers were common"

  Wikipedia

- General idea for information distortion that employs **falsity** or **manipulation** to intentionally cause **gains** for the fake news issuer.

2016 Explosion of sensational reports
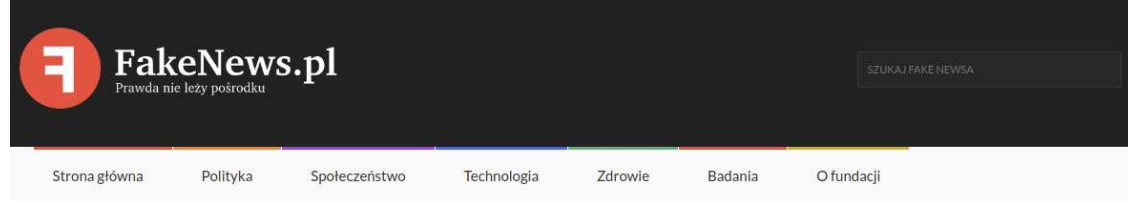
Who gets to decide?

Cartoon by Cathy Wilcox, drawn for UNESCO for World Press Freedom Day 2017.

Fact checking

Published by the Council of Europe
 (F-67075 Strasbourg Cedex)

- fact-checking database created by
   Duke University's Reporter's Lab
   https://reporterslab.org/fact-checking/

  - Agence France-Presse (AFP) AFP Fact Check
    (Poland) https://sprawdzam.afp.com/list

  - https://pravda.org.pl/

  - https://fakenews.pl/ FUNDACJA
    "PRZECIWDZIAŁAMY DEZINFORMACJI"

- List of countries:
  Poland:

  - Demogog is a project of the Demogog Association,
    the first professional fact-checking organization in
    Poland.

  - OKO.press fact-checks statements made by Polish
    public figures. It was founded in 2016 and is
    supported entirely by individual donations.

    - Indyvidual donations...

FakeNews.pl
Prawda nie leży pośrodku

SZUKAJ FAKE NEWSA

Strona główna   Polityka   Społeczeństwo   Technologia   Zdrowie   Badania   O fundacji

Inwazja Rosji na Ukrainę. Tropimy dezinformację [Artykuł aktualizowany]

⊙ Utworzone: 24 lutego 2022   ⊙ Aktualizacja: 11 marca 2022   ⚙ przez Karol Orzeł   💬 Komentarze

W związku z atakiem Rosji na Ukrainę w mediach społecznościowych pojawił się ogromny szum informacyjny, w którym informacje prawdziwe mieszają się z dezinformacją. Obserwujemy również towarzyszący tym wydarzeniom nagły wzrost ilości nieprawdziwych informacji i narracji prorosyjskiej w przestrzeni internetowej. Ze względu na zaistniałą sytuację postanowiliśmy zebrać w tym miejscu nasze posty z mediów społecznościowych związane z konfliktem rosyjsko-ukraińskim.

10.03.2022

Combat Fake News

- Polish Press Agency (PAP)

  https://fakehunter.pap.pl/

- www.truthorfiction.com

- Which fake news was the most read each year on different platforms such as Twitter or Facebook,

Fake News detection

Fake news comes from lack of editorial standards

- Fake news use the mechanizm of **gossiping** and **chinese wispers**

- Many people now get news from social media sites and networks → no standards

  - Filter bubbles
  - Echo chambers

- It can be difficult to tell whether stories are credible or not

- Information overload + general lack of understanding about how the internet works

- Social media sites can play a big part in increasing the reach of these type of stories → my friend likes it so am I – social proof

The economics of social media favour gossip, novelty, speed and "shareability"

False / Fake news?

fake - udawać

- False information is news, stories or hoaxes created to deliberately misinform or deceive readers.

- Usually, these stories are created to:

  - influence people's views,
  - push a political agenda or
  - cause confusion and
  - can often be a profitable business for online publishers.

- False information can deceive people by:

  - looking like trusted websites or
  - using similar names and web addresses to reputable news organisations.
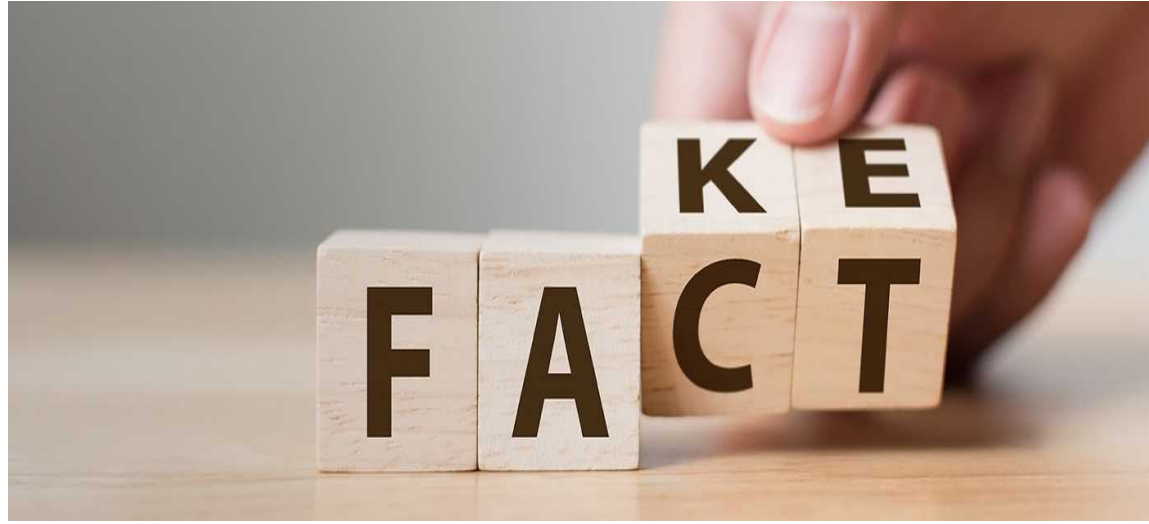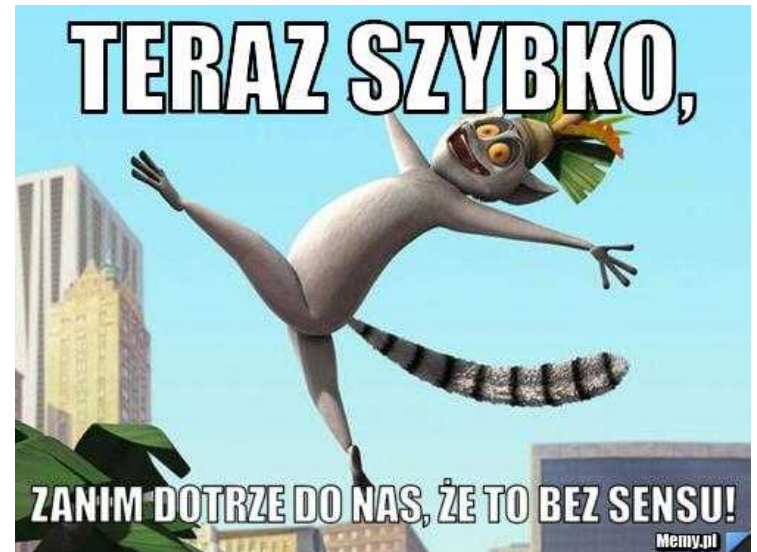
Call to action

- Engade reader to prompt response

- "Best" response is to:

  - share
  - comment
  - even object

    the more action.. the better

- More traffic

  → the more visibility

    → more spread

- Kind like virus infection

Hurry up! Before we all come to our senses!

TERAZ SZYBKO, ZANIM DOTRZE DO NAS, ŻE TO BEZ SENSU!
Memy.pl

Fake News
evaluation characteristics

- **Emotions**

- Negativity

- Provocation

- Stereotypical

- Fulfilling demand

- Use recent events to authenticate itself

- Judging / evaluation statements

- False support from bot / fake via social networks

- Easy to generate

- Targeted to specified social group audience focused on thier daily limitations or new constraints, threats

Fake news for dummies

General problem

- Spreads false information

- Can cause panic

- Can defame people / institutions / countries

- Spread false medical facts and manipulations

- Market / financial market manipulation

Information disorder reports - 2017

- Published by the Council of Europe (F-67075 Strasbourg Cedex)

- Fake news inadequate to describe the complex phenomena of information pollution

- Politicians around the world to describe news organisations whose coverage they find disagreeable

- Conceptual framework for examining information disorder, identifying the three different types: mis-, dis- and mal-information

- Use the dimensions of harm and falseness

Framework:
types of information disorder

- **Mis-information** is when false information is shared, but no harm is meant.

- **Dis-information** is when false information is knowingly shared to cause harm.

- **Mal-information** is when genuine information shared to cause harm, often by moving information designed to stay private into the public sphere

INFORMATION DISORDER

FALSE   HARMFUL

Mis-Information

False Connection
Misleading Content

Dis-Information

False Context
Imposter Content
Manipulated Content
Fabricated Content

Mal-Information

Leaks
Harassment
Hate speech
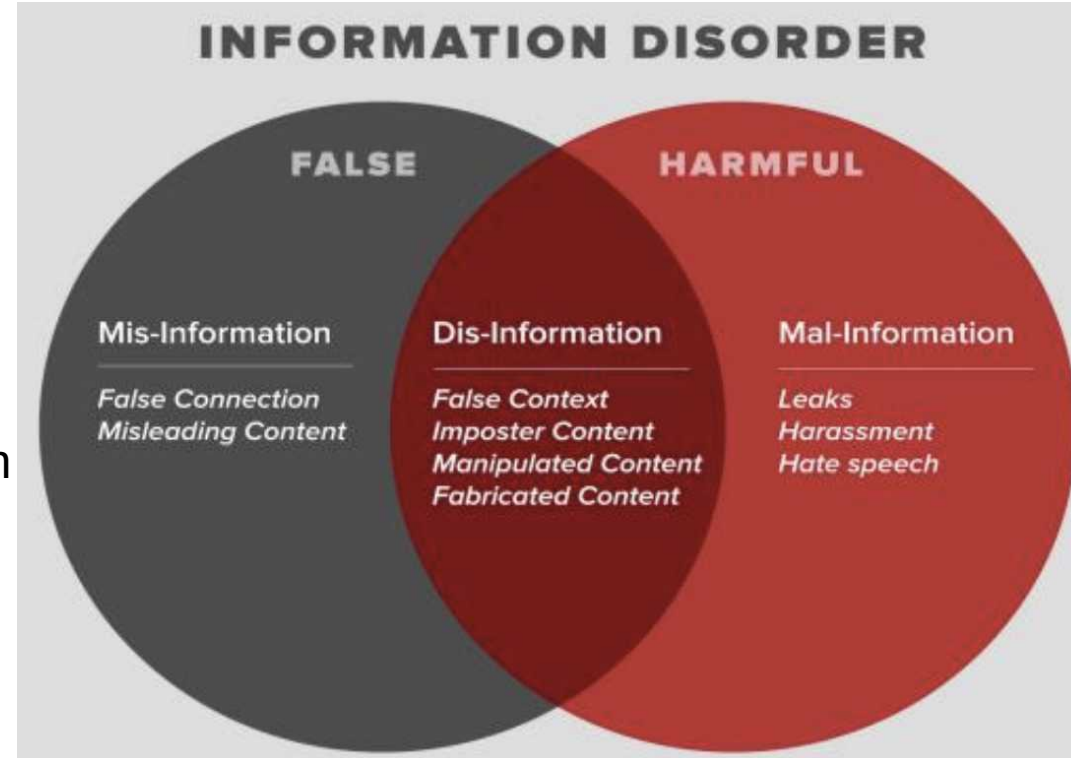
Framework: The components

Agent	Actor Type: Level of Organisation: Type of Motivation: Level of Automation: Intended Audience: Intent to Harm: Intent to Mislead:	Official / Unofficial None / Loose / Tight / Networked Financial / Political / Social / Psychological Human / Cyborg / Bot Members / Social Groups / Entire Societies Yes / No Yes / No
Message	Duration: Accuracy: Legality: Imposter Type: Message Target:	Long term / Short-term / Event-based Misleading/ Manipulated / Fabricated Legal / Illegal No / Brand / Individual Individual / Organisation / Social Group / Entire Society
Interpreter	Message reading: Action taken:	Hegemonic / Oppositional / Negotiated Ignored / Shared in support / Shared in opposition

Framework: phases

- 'Agent' who creates a fabricated message might be different to the agent who produces that message—who might also be different from the 'agent' who distributes the message.

- Similarly, we need a thorough understanding of who these agents are and what motivates them.
- Understanding the different types of messages being distributed by agents, helps estimating the scale of each and addressing them

Creation when the message is created

(Re)Production when the message is turned into a media product

Distribution when the product is distributed or made public
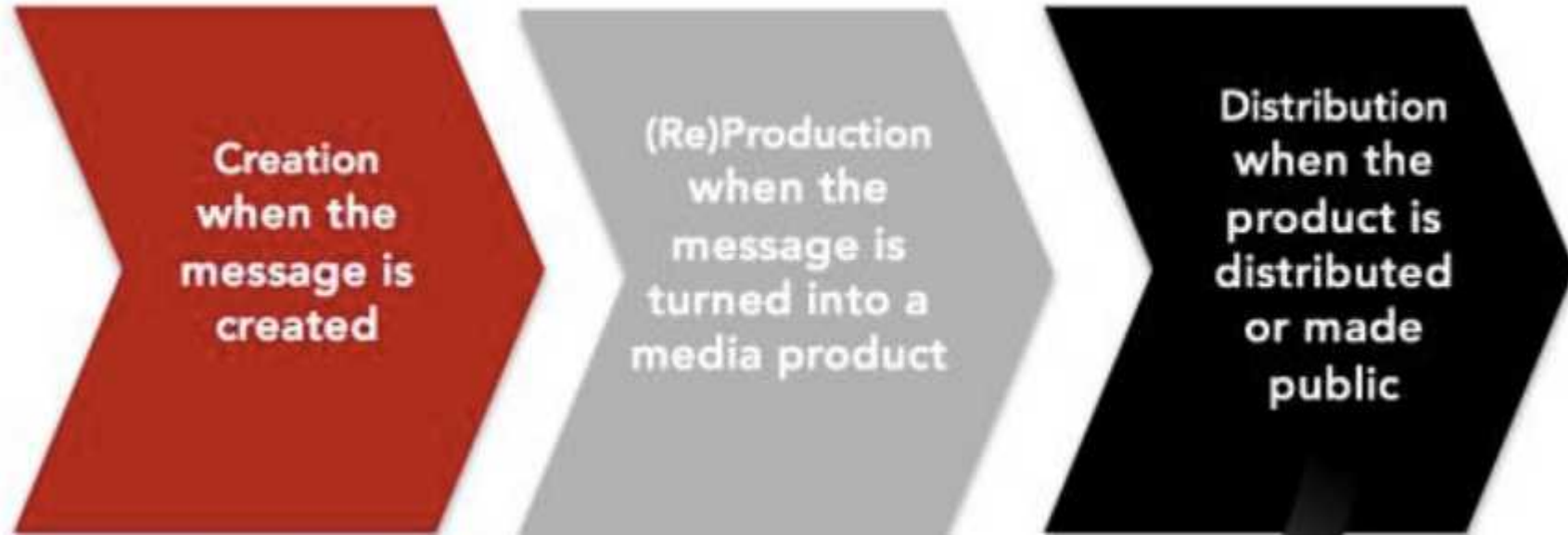
Three phases of information disorder

Creation

Article conceived by an unidentified person.

Production

Article published on the site WTOE5 News, part of a network of 43 fabricated news sites, which published over 750 articles.

Distribution

Article shared on Facebook by someone working for this network of fabricated sites.

Re-Production

Article shared by people connected to the fabricated news site network to amplify the impact of the article to make more profit.

Article shared on Facebook by Trump supporters.

Article shared by forces who had an interest in Trump winning (e.g. became content amplified by Russian troll factories or bot networks).

Article shared by Hillary Clinton supporters as evidence of how easily Trump supporters could be fooled

WTOE 5 NEWS
YOUR LOCAL NEWS NOW

TOP STORIES   COMMUNITY   ENTERTAINMENT   SPORTS   LIFE   ABOUT   LATEST NEWS

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

TOPICS: Pope Francis Endorses Donald Trump

Fake News phase analysis

Agent

Type of Actor	Motivation	Level of organization	Use of automation	Intended audience	Intent to harm	Intent to mislead
Unofficial actor	Financial	Connected to a network of sites	No	US voters	Yes	Yes

Message

Format	Durability	Level of accuracy	Imposter	Legality	Target of message	
Text	Election campaign	Fabricated	No	Yes	Hillary Clinton	

Interpreter

How is the message interpreted by different people?	What action is taken? Is the message re-produced?	

3 x 3 dimensions of FN

- The three types: mis-, dis- and mal-information
- The three elements: agents, messages and interpreters
- The three phases: creation, production and dissemination

FIRSTDRAFT

7 TYPES OF MIS- AND DISINFORMATION

SATIRE OR PARODY

No intention to cause harm but has potential to fool

MISLEADING CONTENT

Misleading use of information to frame an issue or individual

IMPOSTER CONTENT

When genuine sources are impersonated

FABRICATED CONTENT

New content is 100% false, designed to deceive and do harm

FALSE CONNECTION

When headlines, visuals or captions don't support the content

FALSE CONTEXT

When genuine content is shared with false contextual information

MANIPULATED CONTENT

When genuine information or imagery is manipulated to deceive

How to defend against Fake News

- <u>Be sceptical</u>: well-read titles, capital letters. incredibly shocking and appears to be untrue, it usually is.

- Are you sure it's not <u>satire</u>?: can be a joke. Fake news is much more serious information.

- <u>Verify the sender</u>

- <u>Source</u>: Many fake news sites cite a source – verify

- <u>Evidence</u>: quoted source - verifying the evidence or research cited by the author

- <u>Linguistic correctness</u>: check for errors

- <u>Pictures</u>: alterations of old photos, attributing an image to a completely different situation

- <u>Dates</u>: timeline is very often wrong. The chronology of events often does not make sense.

- **Compare information**: Has similar information also appeared elsewhere?

- <u>Fake news works because of the reach</u>:
  Not so much an advice on how to detect fake news, but how to fight it
  - remember that it also depends on us whether the fake news spreads further
  - we share it on social media, we talk about it and we spread information whose credibility we have not checked in any way.
  - put some effort into verifying the news we pass on.

Fog of war
FN as a weapon

- No way to verify information

- Most messages are emotional

- Most messages spread as a fake news as mostly they are

JEŚLI W CIĄGU MIESIĄCA WALKI SIĘ NIE ZAKOŃCZĄ

TO W KWIETNIU NADAL BĘDZIE WOJNA

rdo DiCaprio donating

którego po dam, Ukrainą!

obnizce VAT i cen paliw

SL, MNIE 01.02.2022. 10:55 / aktualizacja: 12:15
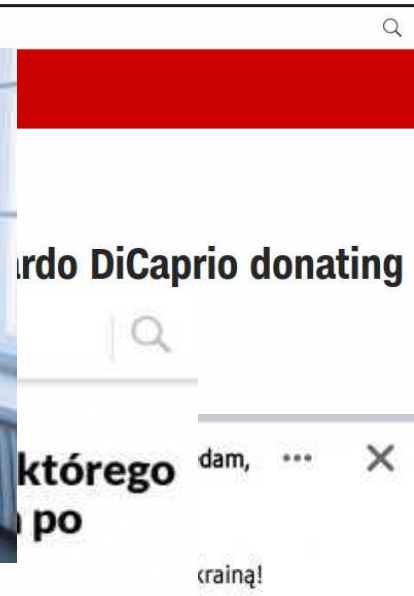
These videos a

Washington (C

Ukraine was re
thousands of p

Articles and sc
maternal grand
that DiCaprio's
Visegrad Fund

Benzyna dużo tańsza. Zdjęcia trafiają do sieci (fot. Twitter/Mateusz Pluta, Robert Gontarz, Artur Kosicki, Paweł Szefernaker, Rafał Mundry)

CNN politics   The Biden Presidency   Facts First   US Elections

5 styczeń 2022

Szef Polityka Insight Centrum Analityczne

Dziennikarz onet
Dziennikarz onet

Dziennikarz newsweek

Profesor prawa SGH

Challnage

- Manual fact-checking is time consuming → to long

  - Fake news spreads fast
  - Hard to rectify

- In need to find automatic assisting tool for FN detection

- MLP architecture is just

  - FFN MLP
  - No convolutions (? controversy)
  - No attention mechanizm

- Consequently MLP Mixer is:

  - Matrix multiplicactions
  - Non-linearities
  - Normalization
  - Skip connections

- General idea:

  - Apply MLP repeatadly for spatial locations and feature channels
- Why

MLP-Mixer: An all-MLP Architecture for Vision

Ilya Tolstikhin[*], Neil Houlsby[*], Alexander Kolesnikov[*], Lucas Beyer[*],

Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner,

Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy

[*]equal contribution

Google Research, Brain Team

{tolstikhin, neilhoulsby, akolesnikov, lbeyer,
 xzhai, unterthiner, jessicayung[†], andstein,
keysers, usz, lucic, adosovitskiy}@google.com

[†]work done during Google AI Residency

Abstract

Convolutional Neural Networks (CNNs) are the go-to model for computer vision. Recently, attention-based networks, such as the Vision Transformer, have also become popular. In this paper we show that while convolutions and attention are both sufficient for good performance, neither of them are necessary. We present *MLP-Mixer*, an architecture based exclusively on multi-layer perceptrons (MLPs). MLP-Mixer contains two types of layers: one with MLPs applied independently to image patches (i.e. "mixing" the per-location features), and one with MLPs applied across patches (i.e. "mixing" spatial information). When trained on large datasets, or with modern regularization schemes, MLP-Mixer attains competitive scores on image classification benchmarks, with pre-training and inference cost comparable to state-of-the-art models. We hope that these results spark further research beyond the realms of well established CNNs and Transformers.[1]

MLP-Mixer

- MLP Mixer architecture key points:

  - Performs adequatly at large scale

    - Not SOTA but comparable

  - Benefit much more from scaling than previous architectures
    - Better throughput img / sec / core (x3 ViT-16)

  - Linear complexity in number of input pixels

    - Similar to CNN, in contrast to ViT $O(n^2)$

  - Conversly to CNNs is not localized as its filter spans entire spatial area

    - No resolution shrinking
    - Stack layers of same size
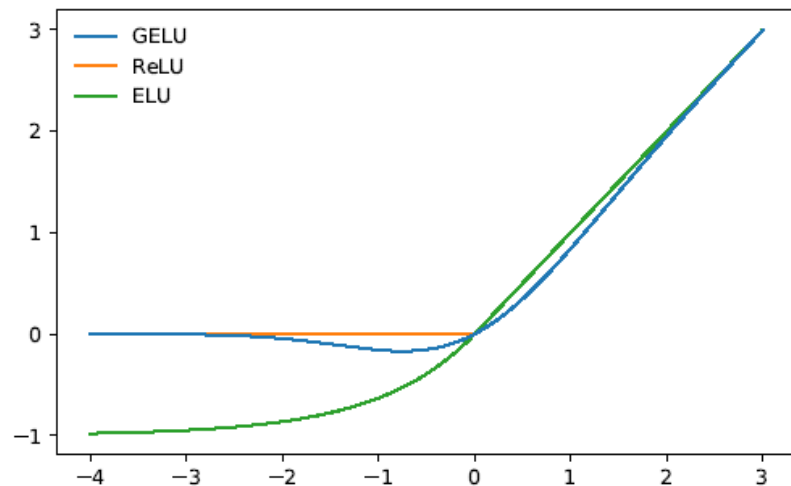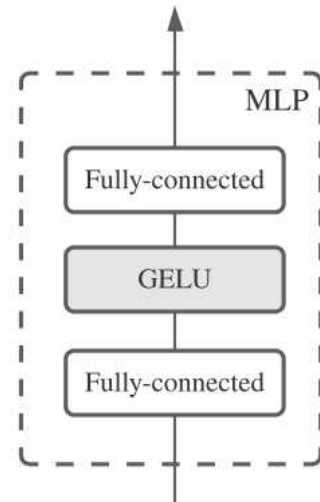
Architecture

```
# non-overlapping patches; conv 2D for a square
kernel_size = patch_size
strides = patch_size
num_patches_x = image_x_dims / patch_size
num_patches_y = image_y_dims / patch_size
# Input: x * y + (channels(filters#)
num_patches_x * num_patches_y * channel#
```
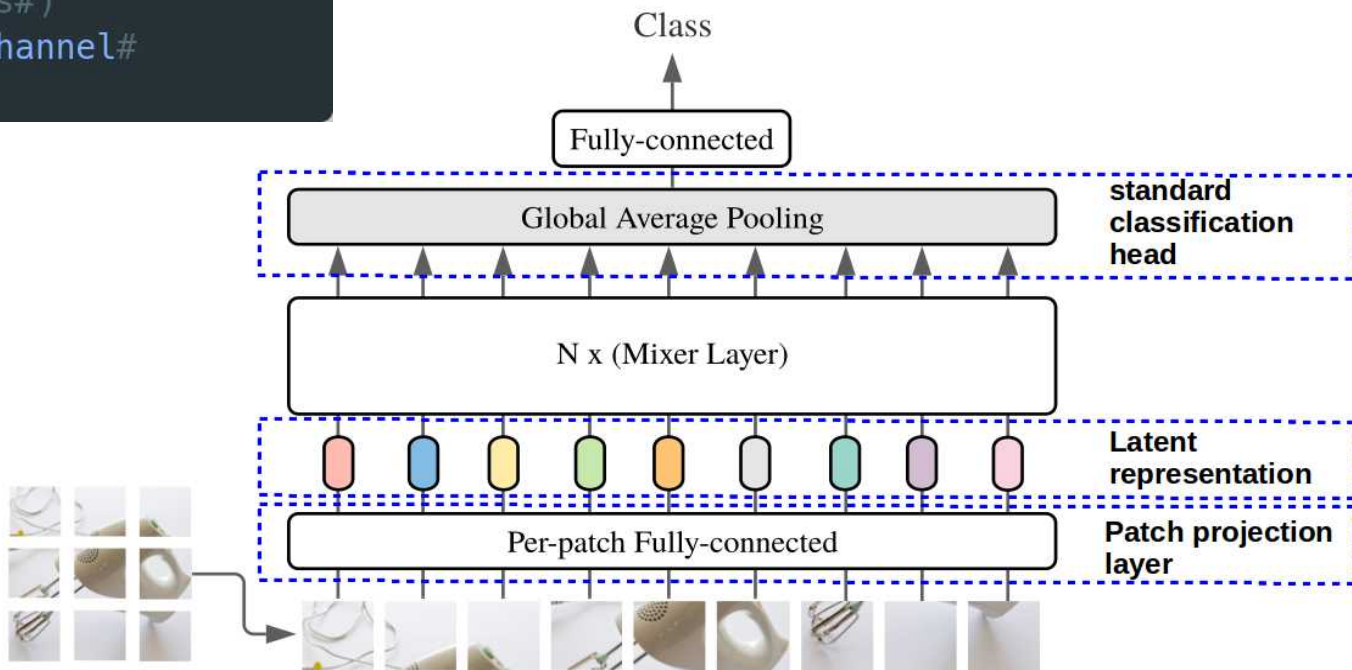
Architecture

- Token-mixing
  forward propagate all channels individually but in the same way

- Channel-mixing
  forward propagate all patches individually but in the same way

- *Layer norm*:
  mean and variance of all summed inputs of a neuron are used to perform normalization

LN :normalize the distributions of intermediate layers. Enables smoother gradients, faster training, and better generalization accuracy.
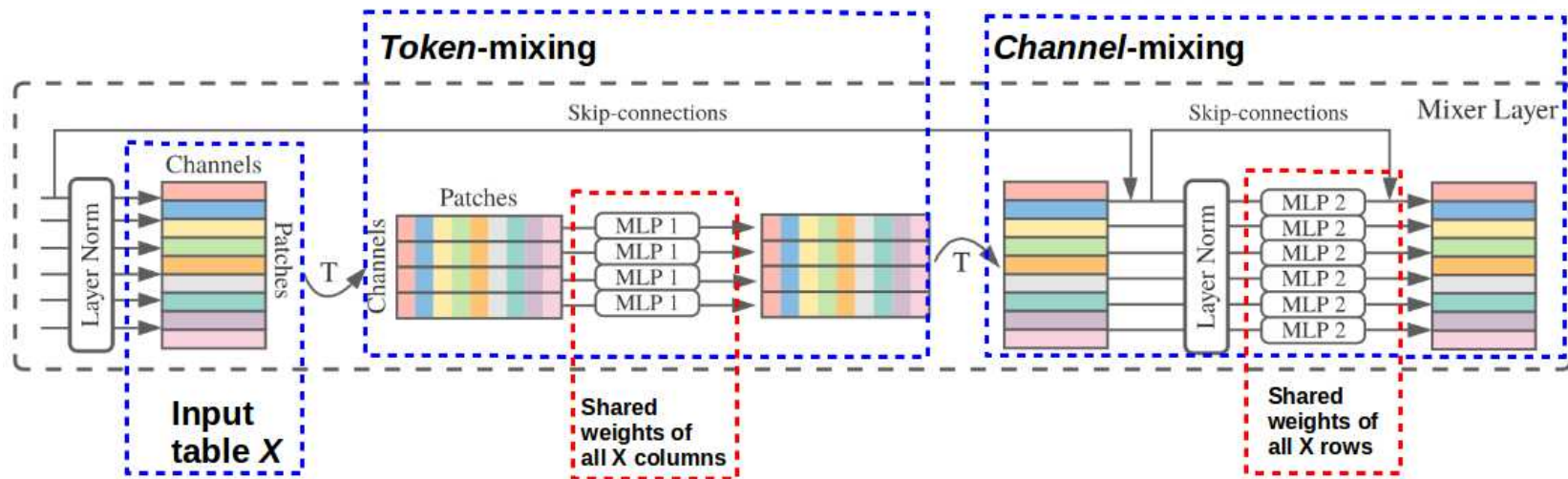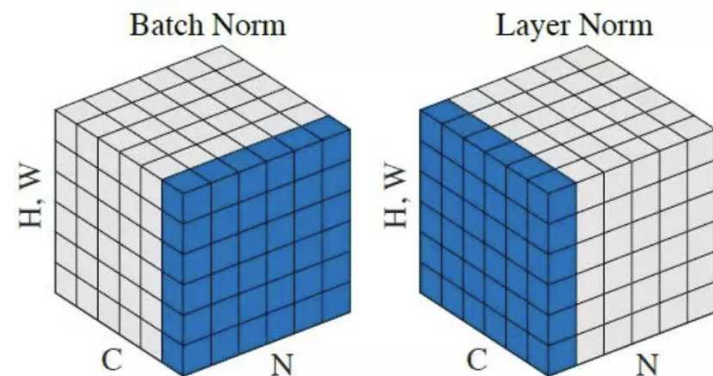
Why mixer?

- Benefits more from data scale
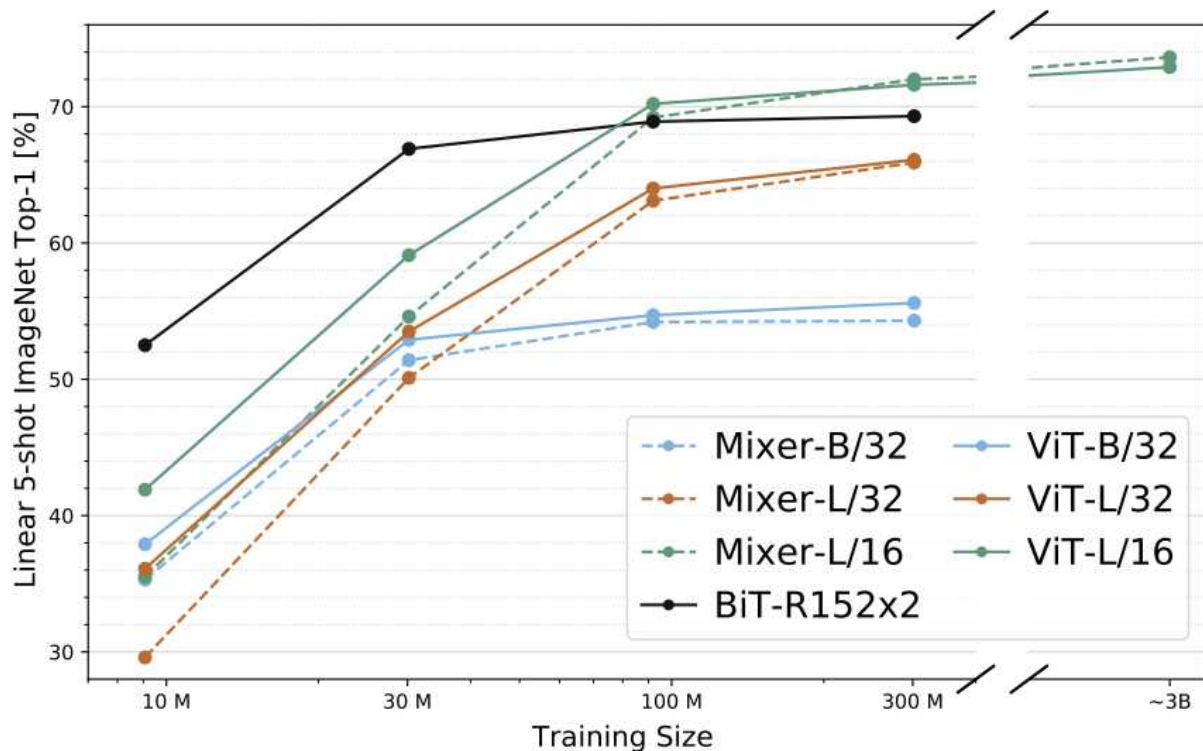
- Linear classifier on frozen representation
- Evaluate Top-1 accuracy
- 5 shoot classifier

- Top-1 accuracy is the conventional accuracy, model prediction (the one with the highest probability) must be exactly the expected answer.

Use it for Fake News classification

- Change the sequence of image patch tokens into sequence of text tokens

- Channels are now token embedding dimensions

- Patches are produced form sequence with Conv1D

- The fake news clasification mixer will have:

  embedding_dims: 64, token_mixing_mlp_dims: 32, channel_mixing_mlp_dims = 64
  patch_size = 5, num_mixer_layers = 8

Results

- Mixer vs CNN is comparable on same dataset in terms of precision / recall / f1-score

- Mixer is susceptible to overfitting → regularization is required

  - Best results with Dropout (0.4)

- Better results by increasing:

  - The embedding dimensions

  - Number of mixer blocks

  - Training model for longer

```python
models.Sequential([
    layers.Embedding(tk.num_words, 8, input_length=input_length
    layers.Conv1D(16,4,activation='relu'),
    layers.AveragePooling1D(),
    layers.Conv1D(32,4,activation='relu'),
    layers.AveragePooling1D(),
    layers.Flatten(),
    layers.Dense(1, activation='sigmoid')
])
```

```python
# Multilayer Perceptron with GeLU ( Gaussian Linear Units ) activation
def mlp( x , hidden_dims ):
    y = tf.keras.layers.Dense( hidden_dims )( x )
    y = tf.nn.gelu( y )
    y = tf.keras.layers.Dense( x.shape[ -1 ] )( y )
    y = tf.keras.layers.Dropout( 0.4 )( y )
    return y
```

	precision	recall	f1-score	support
not fake	0.79	0.72	0.75	898
fake	0.64	0.72	0.68	625

Future work

- Corpus creation for polish
  - More in-deph: tagged with manipulation techniques and conceptual framework

- Test for larger dataset
  - Multiclass classification → how malitious a fake news is
  - Multilabel mis-infomration, dis-information, mal-information

- Test the projection layer model instead of fixed learned embedding like using non-trainable hash functions
  - locality sensitive hashing (LSH) to create representations form text
  - min-hash

References:

1) *"The real story of 'fake news': The term seems to have emerged around the end of the 19th century". Merriam-Webster. Retrieved October 13, 2017.*

2) *"INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making" Council of Europe report DGI(2017)09 Claire Wardle, PhD; Hossein Derakhshan*

3) *Soll, Jacob (December 18, 2016). "The long and brutal history of fake news". Politico Magazine. Retrieved March 25, 2019.*

4) *"The science of fake news" David M. J. Lazer, Matthew A. Baum,Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, Jonathan L. Zittrain SCIENCE, 9 Mar 2018, Vol 359, Issue 6380, pp. 1094-1096, DOI: 10.1126/science.aao2998*

5) *Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D.,Uszkoreit, J., Lucic, M., and Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. CVPR 2021*

6) *Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", Complexity, vol. 2020, Article ID 8885861, 11 pages, 2020. https://doi.org/10.1155/2020/8885861*

7) *Indyk, P. and Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604–613, 1998.*