FAKE NEWS

Current State of the art in Disinformation Generation and Detection

Alexander Schindler

Scientist

Information Management

Center for Digital Safety & Security

AIT Austrian Institute of Technology GmbH

Mina Schütz

Intern

Information Management

Center for Digital Safety & Security

AIT Austrian Institute of Technology GmbH

Agenda

- Introduction to Fake News
- Generative Processes
  - GAN / VAE
  - Sequence to Sequence
- Fake News Detection approaches
  - Visual
  - Accoustic
  - Textual
- Discussion and Conclusion

Who shares Fake News?

- Do you / have you shared **Fake News**?
- Are you likely to share **Fake News**?
- Can you identify **Fake News**?

George Bush Reading Upside-Down

- George W. Bush (US Pres.)
- Rick Perry (Gov. Texas)
- George Sanchez Charter School (Houston)
- 2002

Fake Image

Original Image

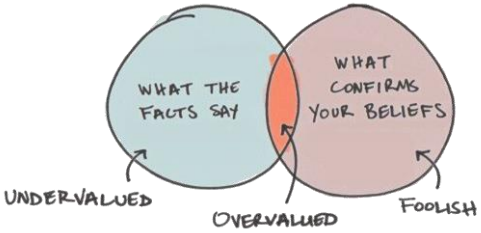

Flipping Error


Additional Proof

Ideological Priors

or

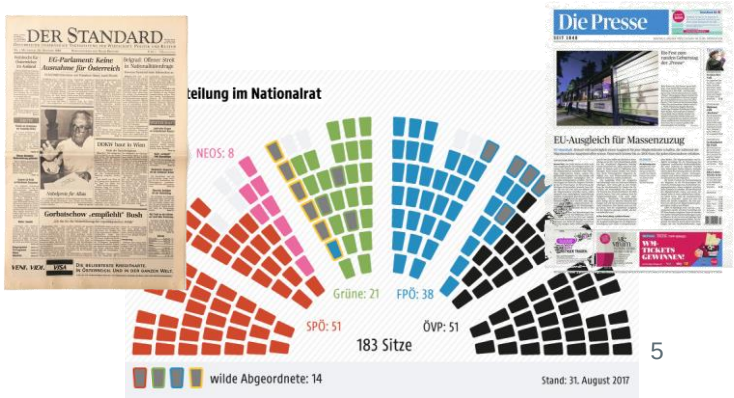
Why do we fall for Fake News?



Confirmation Bias

- **Naive Realism**
  - Believe information that is aligned with your views
- **Confirmation Bias**
  - Seek information that confirms your existing views
- **Normative Influence Theory**
  - Consume/Share socially safe options ➔ for social acceptance, affirmation

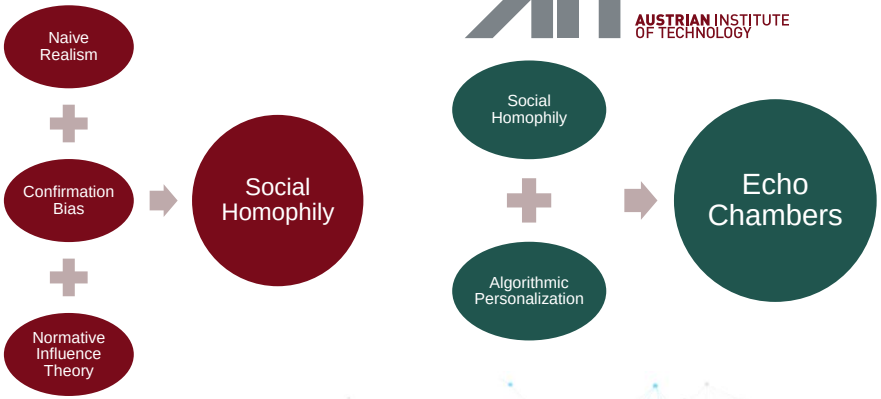
➔ Individual level of Fake News

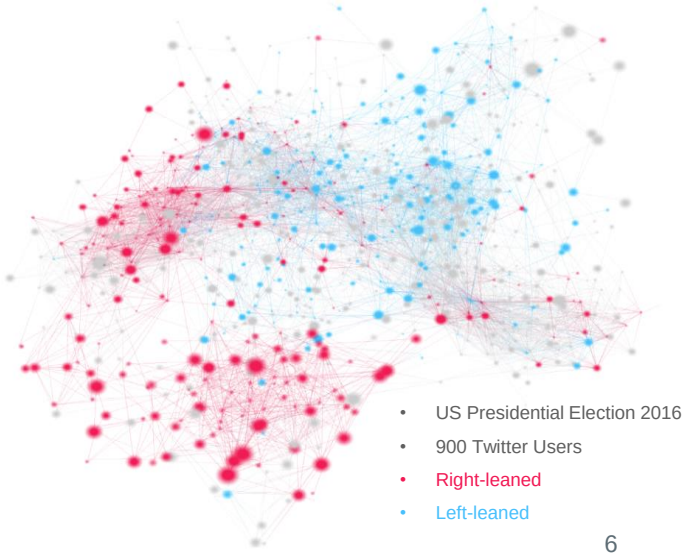


Nature / Characteristics

- **Social Level**

- **Echo Chambers / Filter Bubbles**
  - Social homophily
    - Form connections with ideologically similar individuals
  - Algorithmic personalization
    - Read content
    - Follow / befriend persons
  - ➢ Consequences
    - ➢ Less exposure to conflicting viewpoints
    - ➢ Isolation in own filter bubble
    - ➢ Improve survival / spread of fake news





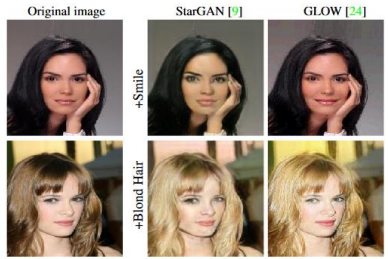
Gillani, Nabeel, et al. "Me, my echo chamber, and I: Introspection on social media polarization." Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018.

Types of Fake News

- **Fabricated content**
  - *Completely false*
- **Misleading content**
  - *Misleading use / framing of issue*
- **Imposter content**
  - *Genuine source impersonated with false sources*
- **Manipulated content**
  - *For deception (e.g. images)*
- **False connection**
  - *Headlines, visuals do not support content*
- **False context**
  - *Genuine content shared with false context information*

Fabricated Content

False Context



Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Manipulated Content

Imposter Content

How Your Audience Will Believe Anything: The Psychology Behind the Fake News https://www.click.co.uk/blog/how-your-audience-will-believe-anything-the-psychology-behind-the-fake-news-bas-van-den-belds-benchmark-2018-talk-review/

Sepcial Types of Fake News

- **Satire**
  - *Excluded from definition*
- **Hoax**
  - *False story – masquerade truth*
- **Rumors**
  - *Unverified sources*
  - *Not nescessarily false*
  - *May be verified later as true / false*

Satire

Hoax

Singapore dismisses Lee Kuan Yew death report as hoax
https://edition.cnn.com/2015/03/18/world/singapore-lee-kuan-yew/index.html

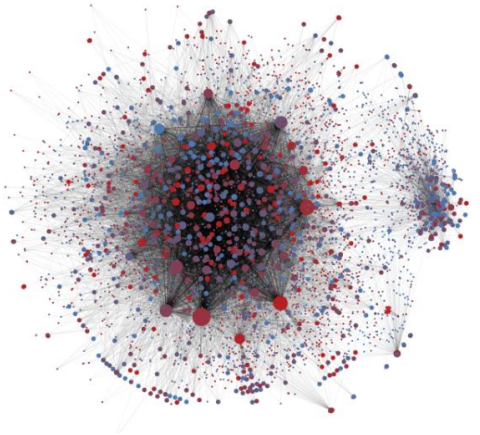
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

FAKE NEWS DETECTION

Primary Characteristics

- **Source / Promoters**
  - Who posts fake news? Who shares it?
  - Bots
- **Information Content**
  - Content
  - Linguistic style
- **User Responses**
  - Reactions to articles
  - Positive/Negative/Netral

Bots spreading false information in social media



Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. arXiv preprint arXiv:1804.08559.

Challenges

- High stakes and multiple players
- Adversarial intent
- Lack of awareness
- Propagation dynamics
- Constant change

General Remarks

- No general Fake News Detection Tool available
- Modality dependant approaches
  - Text Domain
  - Visual Domain
  - Accoustic Domain
  - Network Domain
- Hybrid / Multi-Modal approaches

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

GENERATIVE PROCESSES

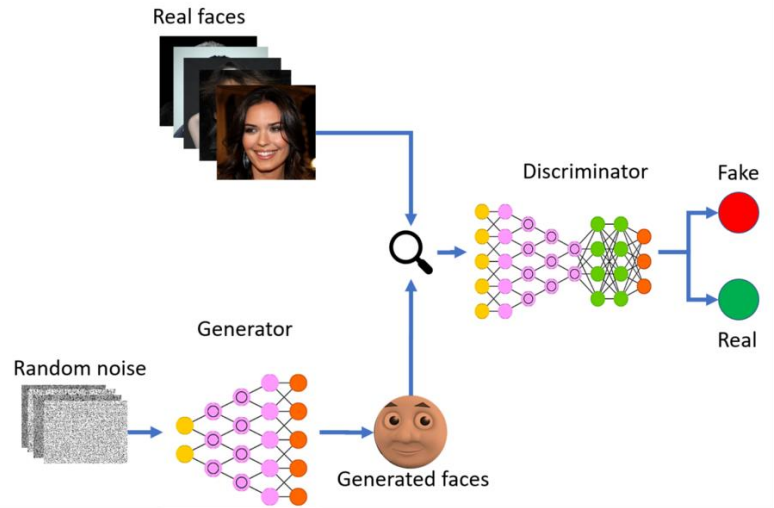
Content Synthesization and Manipulation

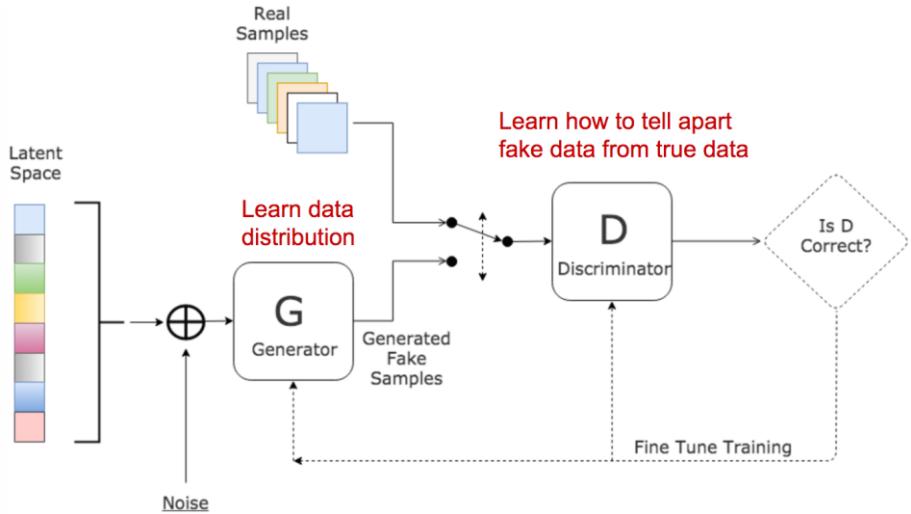
Generative adversarial nets

I Goodfellow, J Pouget-Abadie, M Mirza... - Advances in neural ..., 2014 · ... Like **generative adversarial networks**, variational autoencoders pair **network** with a second neural **network**. Unlike **generative adversarial network** in a VAE is a recognition model that performs approximate infer

☆ 99 Zitiert von: 13618 Ähnliche Artikel Alle 29 Versionen »

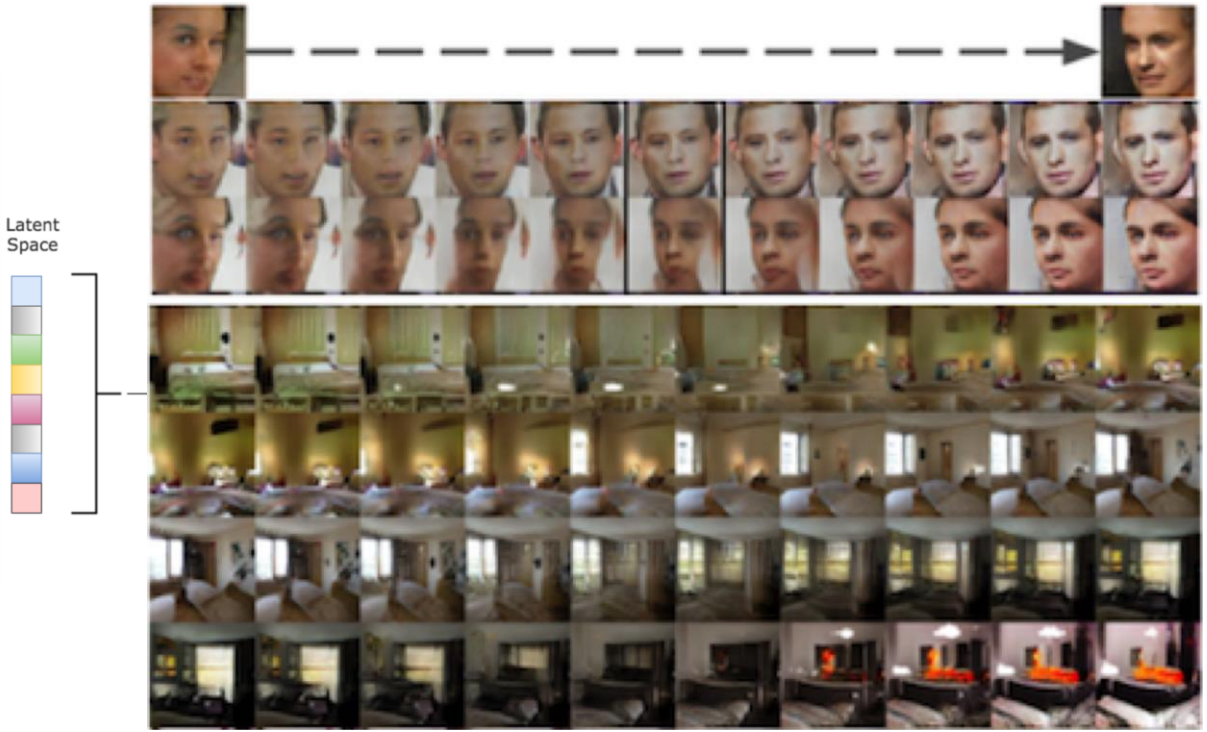
Generative Adversarial Networks (GAN)





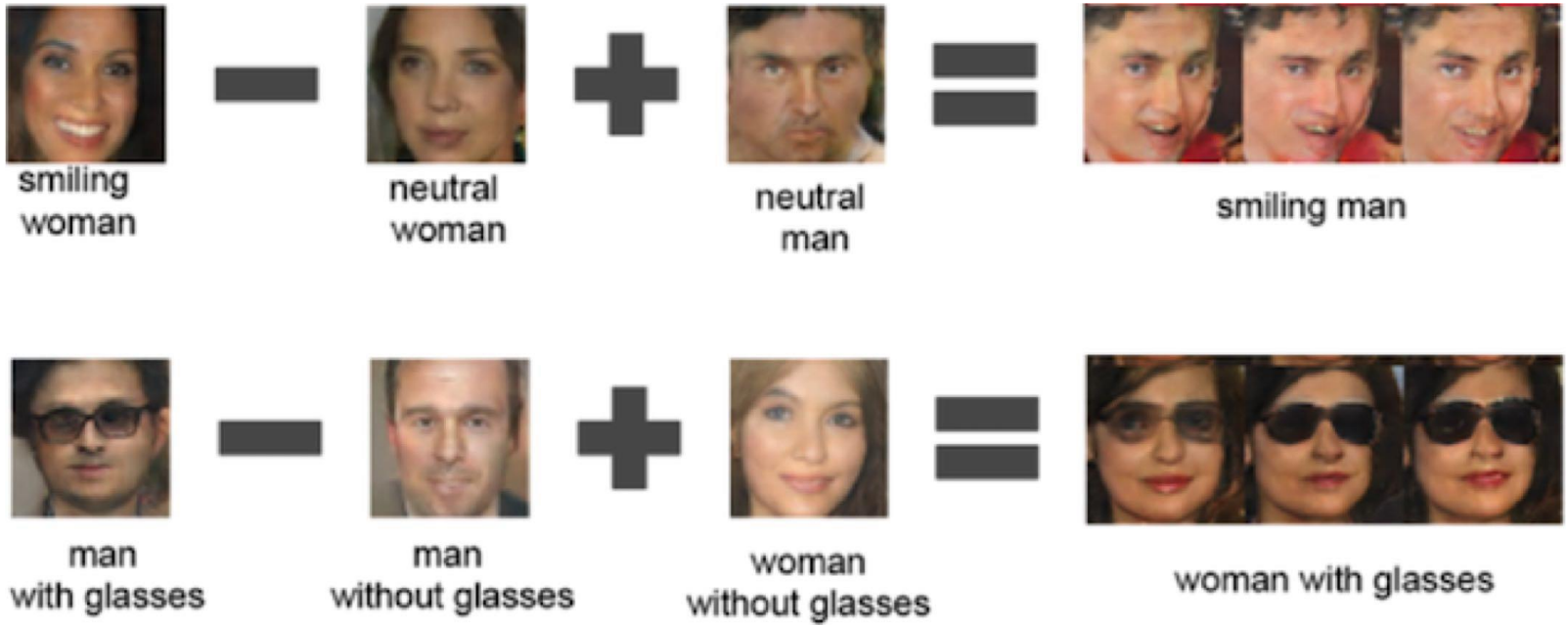
A Brief Introduction To GANs
https://medium.com/sigmoid/a-brief-introduction-to-gans-and-how-to-code-them-2620ee465c30

DCGAN: Interpolation in Latent Space



Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

DCGAN: Vector Arithmetic for Visual Concepts



Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Image Generation (2015)

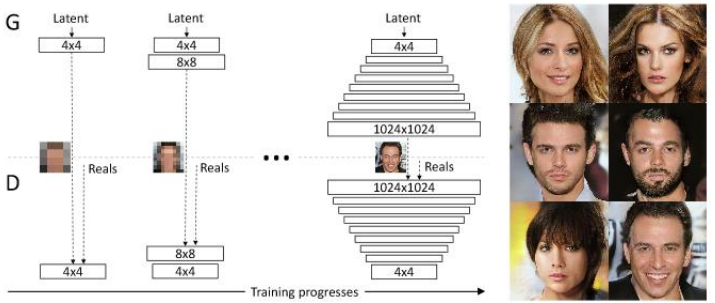
Generative Adversarial Networks: One network transforms random noise into images, a second one tries to distinguish generated from real images. Both are trained at the same time

Photographs of bed rooms that do not actually exist

Nov 2015: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, http://arxiv.org/abs/1511.06434, https://github.com/Newmu/dcgan_code

PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION

Tero Karras NVIDIA
Timo Aila NVIDIA
Samuli Laine NVIDIA
Jaakko Lehtinen NVIDIA and Aalto University
{tkarras,taila,slaine,jlehtinen}@nvidia.com

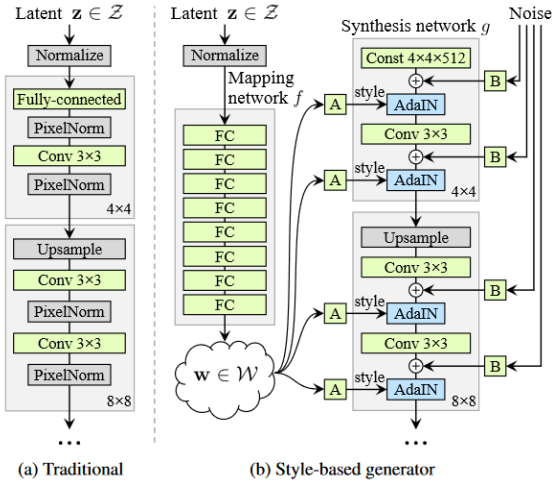


A Style-Based Generator Architecture for Generative Adversarial Networks

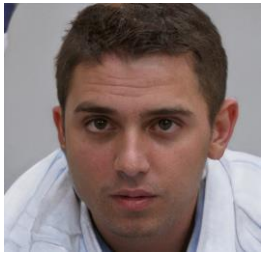
Tero Karras
NVIDIA
tkarras@nvidia.com

Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com



(a) Traditional (b) Style-based generator

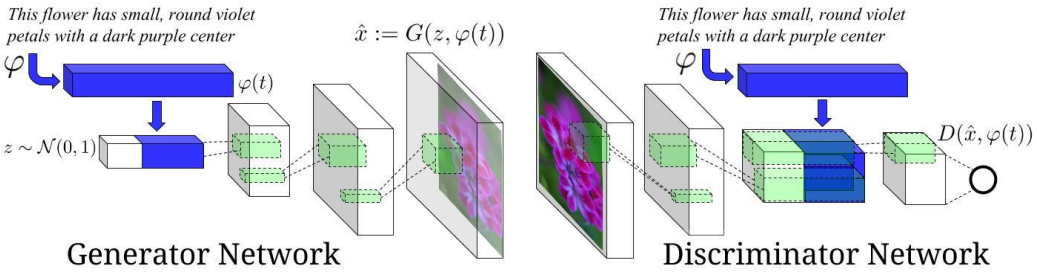


Generative Adversarial Text to Image Synthesis

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran
Bernt Schiele, Honglak Lee

REEDSCOT[1], AKATA[2], XCYAN[1], LLAJAN[1]
SCHIELE[2],HONGLAK[1]

[1] University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)
[2] Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)



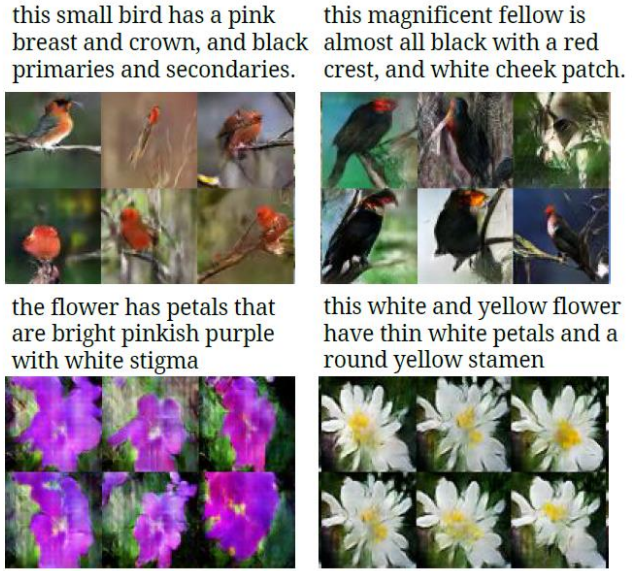
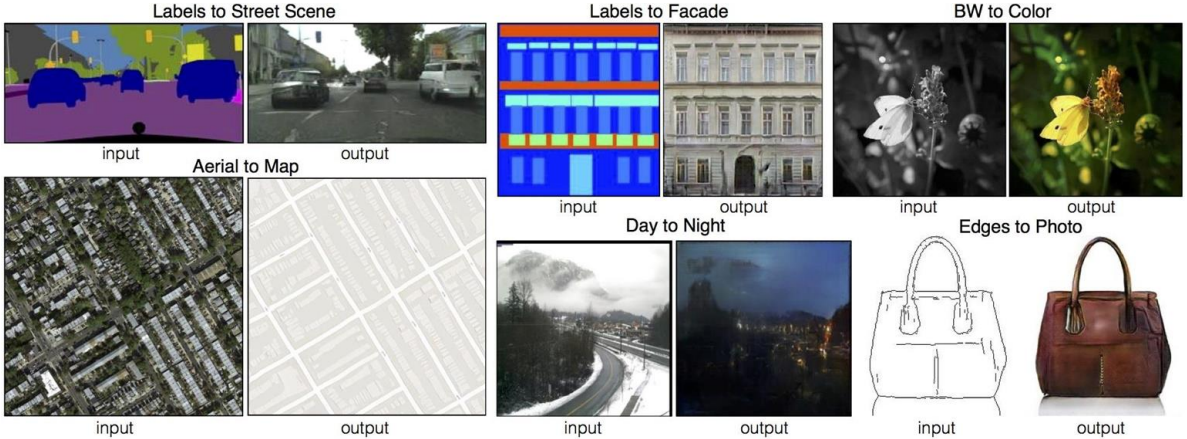


Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories, unseen text. Right: captions are from the training set.

- Paper: http://arxiv.org/abs/1605.05396
- Github: https://github.com/paarthneekhara/text-to-image

CycleGAN

- Based on pix2pix: Fully-convolutional architecture to transform an image into another



- Interactive demo: https://affinelayer.com/pixsrv/

https://github.com/phillipi/pix2pix

Image super-resolution through deep learning

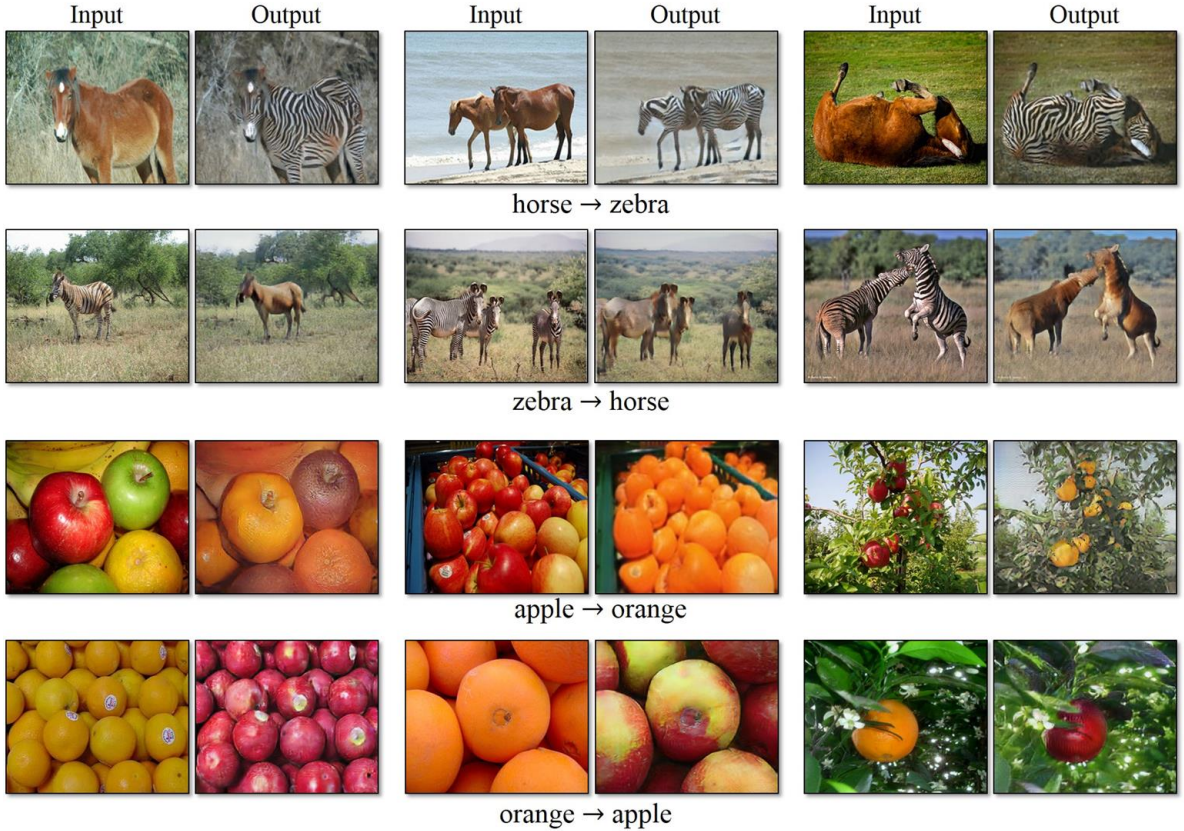
AUSTRIAN INSTITUTE OF TECHNOLOGY

- Upscale 16x16 images by factor 4

- Deep Convolutional Generative Adversarial Network (DCGAN)
  - Image as input instead of gaussian noise
  - Loss function measures difference between input and scaled version
  - Generator uses Residual Network (ResNet) modules

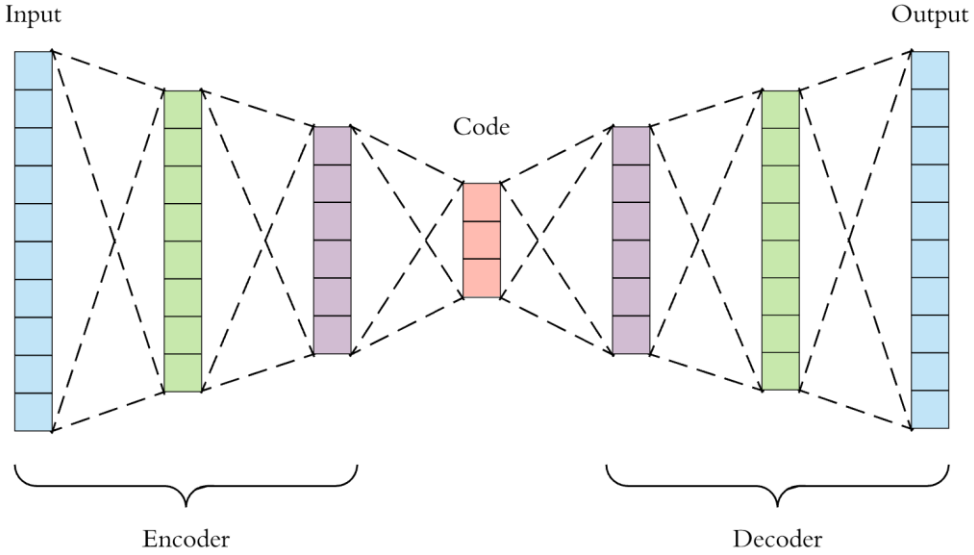
- Code on Github
  - https://github.com/david-gpu/srez

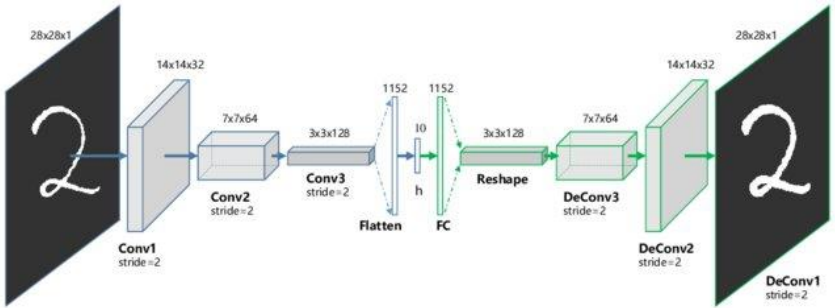
CycleGAN



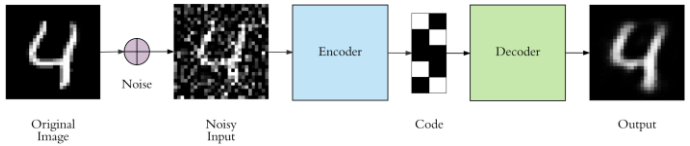
https://arxiv.org/pdf/1703.10593, https://github.com/junyanz/CycleGAN

Auto Encoders



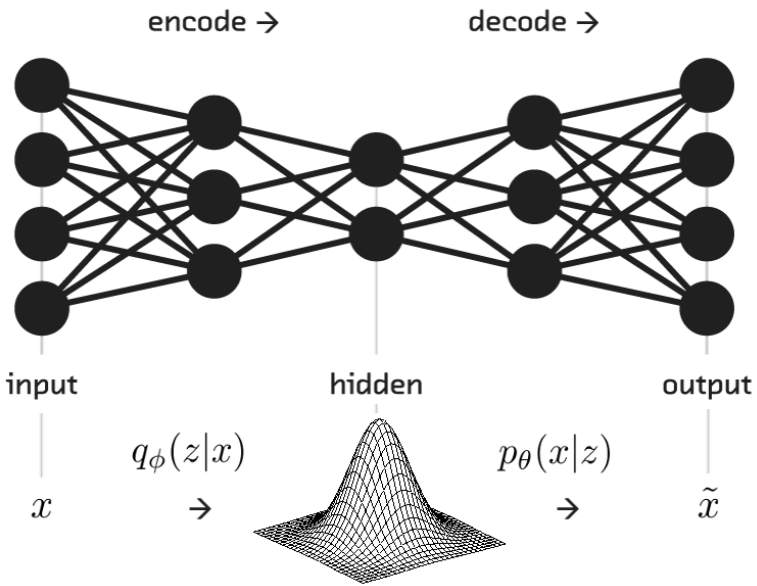


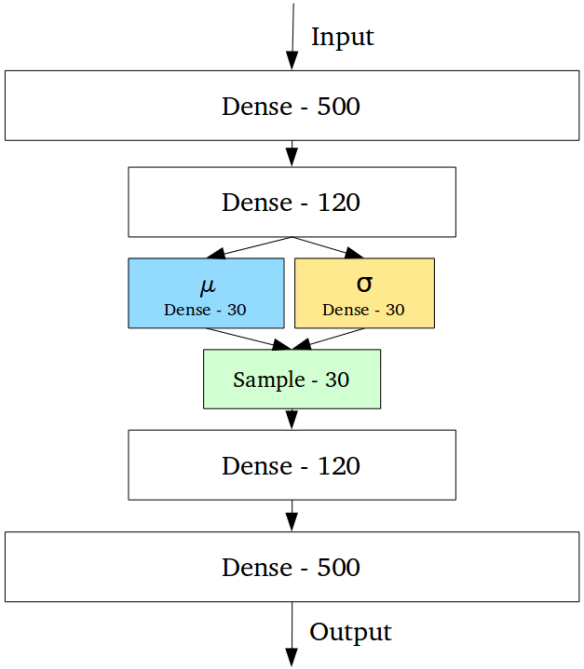
Guo, Xifeng, et al. "Deep clustering with convolutional autoencoders." International Conference on Neural Information Processing. Springer, Cham, 2017.



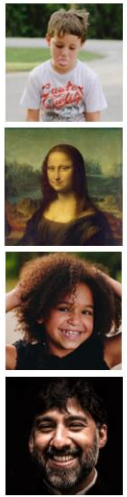
Applied Deep Learning - Part 3: Autoencoders
https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

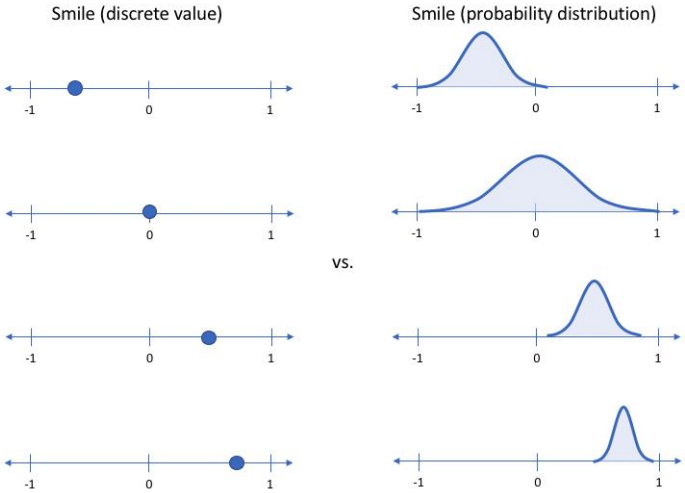
Variational Auto Encoders





Probability Distribution of Latent Factors





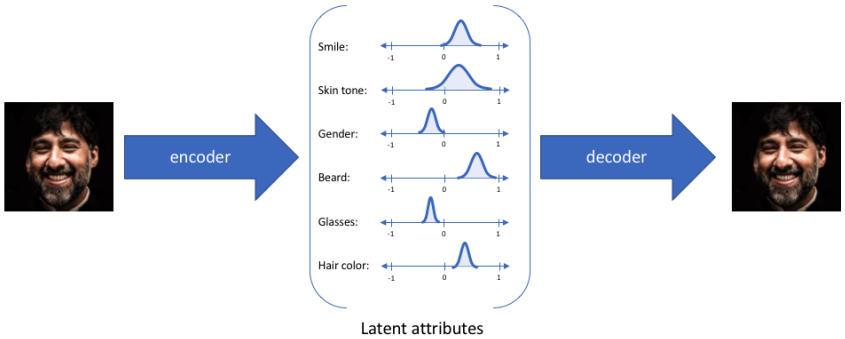
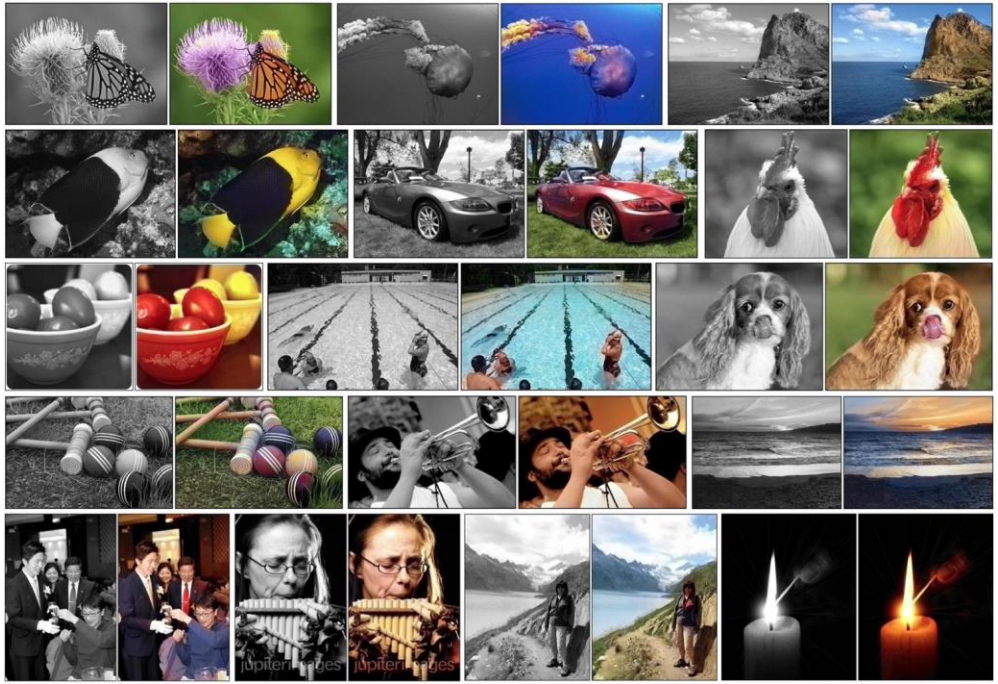


Image Colorization

Input: Gray-scale image. **Output:** Colored image.
Method: fully-convolutional network

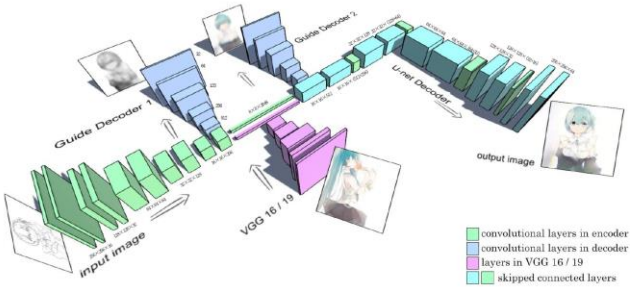


Mar 2016: Colorful Image Colorization, http://arxiv.org/abs/1603.08511, http://richzhang.github.io/colorization/

Style Transfer for Anime Sketches with Enhanced Residual U-net and Auxiliary Classifier GAN

AUSTRIAN INSTITUTE OF TECHNOLOGY

- Apply the style of a painting to a grayscale sketch
- Residual U-Net
  - Image segmentation
  - Large hints from VGG 19 outputs
- with Auxiliary Classifier Generative Adversarial Network (AC-GAN)



https://arxiv.org/abs/1706.03319

WaveNet

ConvNet that can **predict next step of time sequence**, using a clever architecture for processing a large temporal context (about 3000-6000 past time steps)

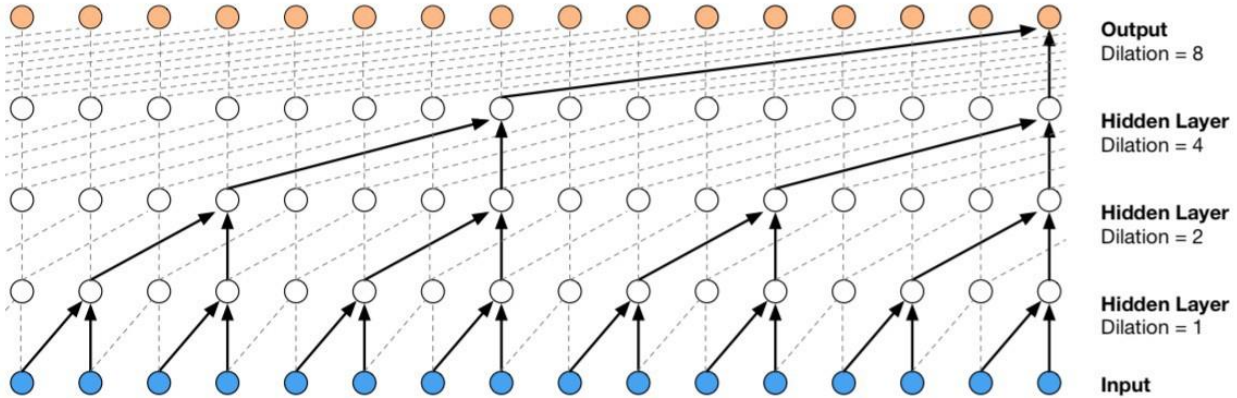


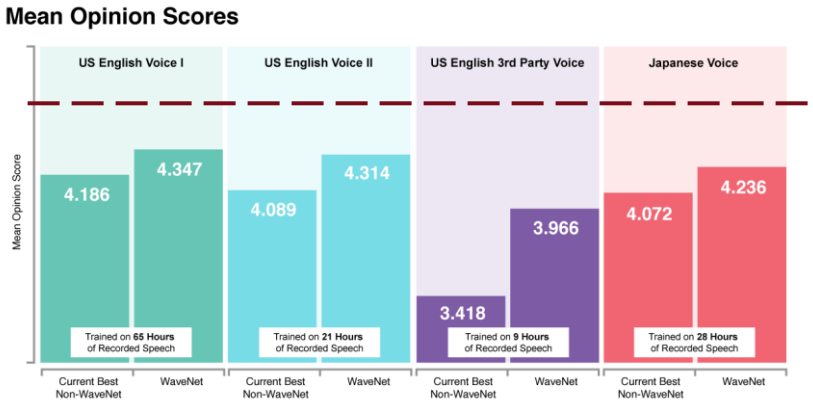
Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Text-to-Speech WaveNet in production

AUSTRIAN INSTITUTE OF TECHNOLOGY

- Previously discussed:
  - WaveNet (DeepMind, Sep 2016, arXiv 1609.03499)
  - Deep Voice 1 (Baidu, Feb 2017, arXiv 1702.07825)
  - Tacotron (Google, Mar 2017, arXiv 1703.10135)
  - Deep Voice 2 (Baidu, May 2017, arXiv 1705.08947)
- Oct 2017: Improved WaveNet used for Google Assistant
  - Generate 20 s of audio in 1 s (old WaveNet: 20 ms in 1 s)
  - 24 kHz, 16 bit (old WaveNet: 16 kHz, 8 bit)



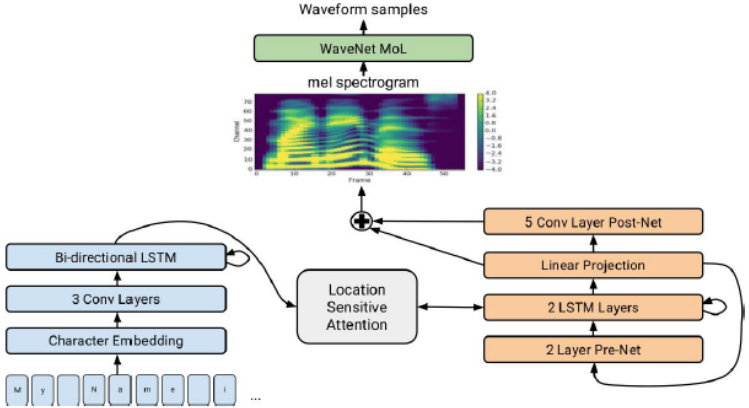
Human voice scores at 4.667.

Side note: Univ. of Hamburg just released aligned Wikipedia audio corpus with hundreds of hours (German, English, Dutch): http://nats.gitlab.io/swc/

https://deepmind.com/blog/wavenet-launches-google-assistant/

Tacotron 2

- Text-to-speech directly from characters, learned end-to-end
- Improvement over Tacotron: mel spectrogram + WaveNet instead of linear spectrogram + Griffin-Lim
- Improvement over some earlier WaveNet-based systems: no need to explicitly predict phones, timing and f0
- Comparison to Deep Voice 3: higher MOS, more mispronounciations, fewer repetitions or skipped words



Name	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground Truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

https://arxiv.org/abs/1712.05884

Tacotron-2 Examples

- Synthetic or Real?

"That girl did a video about Star Wars lipstick."

"She earned a doctorate in sociology at Columbia University."

"George Washington was the first President of the United States."

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

- Stress and intonation
- Questions
- Prosody
  - Intonation, rhythm, tone

Style / Reference

Reference text: Alice was not much surprised at this, she was getting so used to queer things happening.

Result

Perturbed text: Eric was not much surprised at this, he was getting so used to TensorFlow breaking.

Base Voice Model

Singing

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. arXiv preprint arXiv:1803.09047.

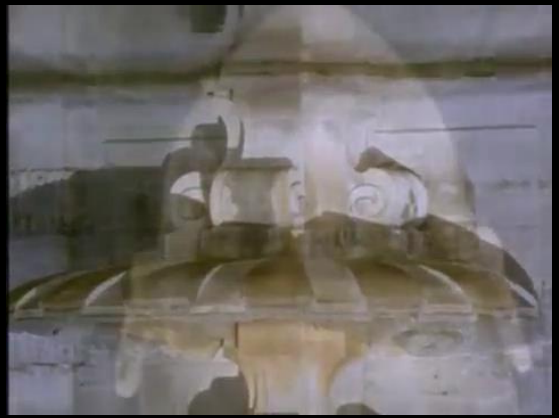
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

VISUAL CONTENT

Finally, Deep Fakes!

Steps to create a Deep Fake with Deep Face Lab

1. Define Source and Destination Videos
2. Extract faces
3. Remove all but destination faces (other persons)
4. Clean source faces (false detections, occlusions, etc.)
5. Align faces
6. Train model
7. Apply model ➔ Convert destination to source faces
8. Assemble video from all frames



Effort to create this Deep Fake

~ 10min manual interaction (for everything)

~ 3h computation time – pre-processing

~ 1h computation time – model training (10K epochs / 10% of normal)

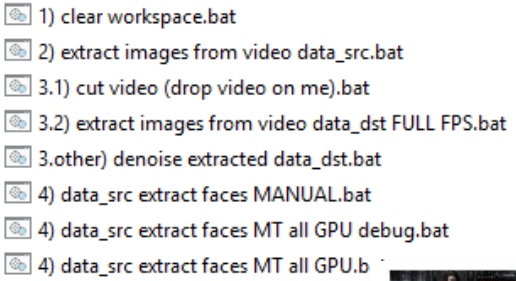
~ 1h computation time – convertion

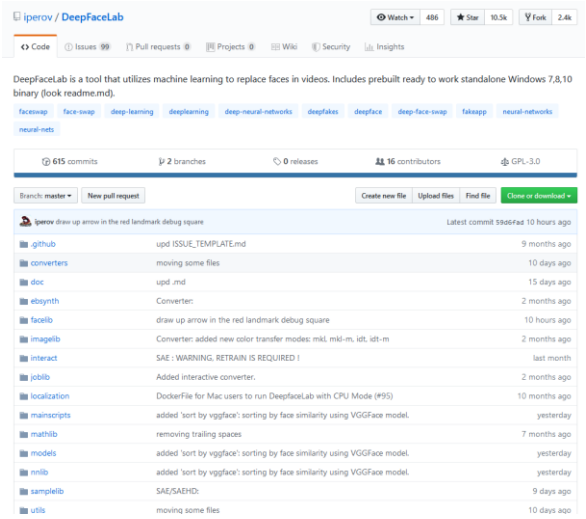
~ 10min computation time – re-assembling video

➔ ~ 6h to create this video!

Deep Face Lab

- Available on Github
- Step-wise documentation
- Prepared, enumerated scripts for Windows

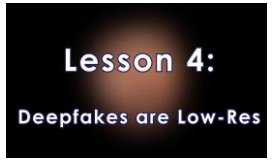




Corridor Crew – Deep Fake Demonstrations

We Made The Best Deepfake on The Internet
https://www.youtube.com/watch?v=3vHvOyZ0GbY&t=886s

How We Faked Keanu Reeves Stopping a Robbery
https://www.youtube.com/watch?v=IzEFnbZ0Zd4



Never hold a sign into a camera

- No AI required!

Original

Fake

Fake

Original

Fake – Das Mädchen mit dem Schild „Will trade racists for rapists“, https://www.mimikama.at/allgemein/fake-schild-will-trade-racists-for-rapists/

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

DETECTION APPROACHES

Visual tampering / forgery

Forgery Detection

- Identify GAN generated / forged images
- Encoder-Decoder Network
- Learn disentangled partitioned latent spaces
  - Real image embedding space
  - Fake image embedding space
  - Difference used to detect fakes
- Disentangled loss
- VGG-like architecture
- Class Activation Mapping (CAM)
  - Visualize / explain faked image regions

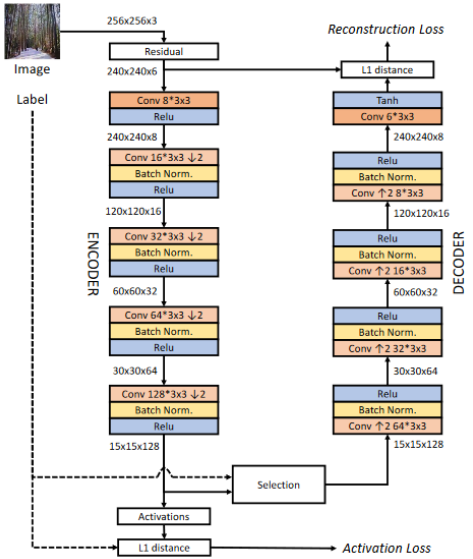


Figure 2: Two examples of images manipulated with Face2Face [39] and FaceSwap [2] (left) and their corresponding class activation maps, when the network (XceptionNet [10]) is trained on Face2Face forgeries (middle) and when it is trained on FaceSwap ones (right).

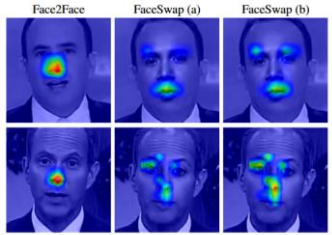


Figure 3: Class activation maps for our method when it is trained on Face2Face [39] and tested on Face2Face forgeries (left) or FaceSwap [2] (middle), and finally trained on Face2Face but fine-tuned using only four images manipulated with FaceSwap and tested on FaceSwap (right).

Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

Detecting GAN generated Faces

- Identify GAN generated faces
- VGG-like architecture / Inception variant
- Train on DeepFake Dataset
- High accuracy with simple approach
- Observation
  - Activations
    - Fake ➔ Background
    - Real ➔ Eyes

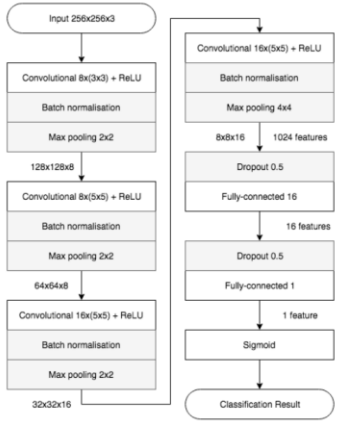
Network	Deepfake classification score		
Class	forged	real	total
Meso-4	0.882	0.901	0.891
MesoInception-4	0.934	0.900	0.917

Table 3. Classification scores of several networks on the *Deepfake* dataset, considering each frame independently.

Network	Aggregation score	
Dataset	Deepfake	Face2Face (23)
Meso-4	**0.969**	**0.953**
MesoInception-4	**0.984**	**0.953**

Table 5. Video classification scores on the two dataset using image aggregation, with the *Face2Face* dataset compressed at rate 23.

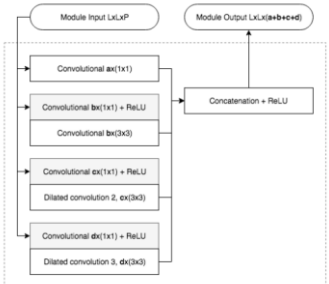
Model Architecture



mean layer output of 100 *deepfake* faces

mean layer output of 100 real faces

Inception Variant



Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018, April). Detection of GAN-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.

Nimble Challenge

- Motivation
  - Media Forensic Technology Development
  - Develop Tools, Evaluation Tasks, Datasets
- 4 Tasks
  - Manipulation Detection and Localization
  - Splice Detection and Localization
  - Provenance Filtering
  - Provenance Graph Building
- Hosted in 2017, 2018
- Successor
  - Media Forensics Challenge 2018, 2019

Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nimble_medifor_july17pimeeting-merge_v12.pdf

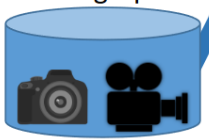
Nimble Challenge 2017

Manipulation Detection and Localization Evaluation Task

System Input

Image(s) + (metadata)

Camera Fingerprint DB



Algorithm

Opt Out

Processed

System Output

e.g., "No illuminated faces in the image for the system to analyze"

Metrics

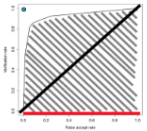
"Not Scored"

Detection

Confidence score

27.58

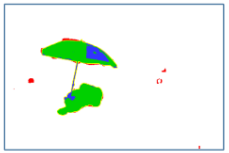
Receiver operating characteristic



Area Under the Curve: 0.85 (AUC)

Localization

Probe Mask (If a manipulation)

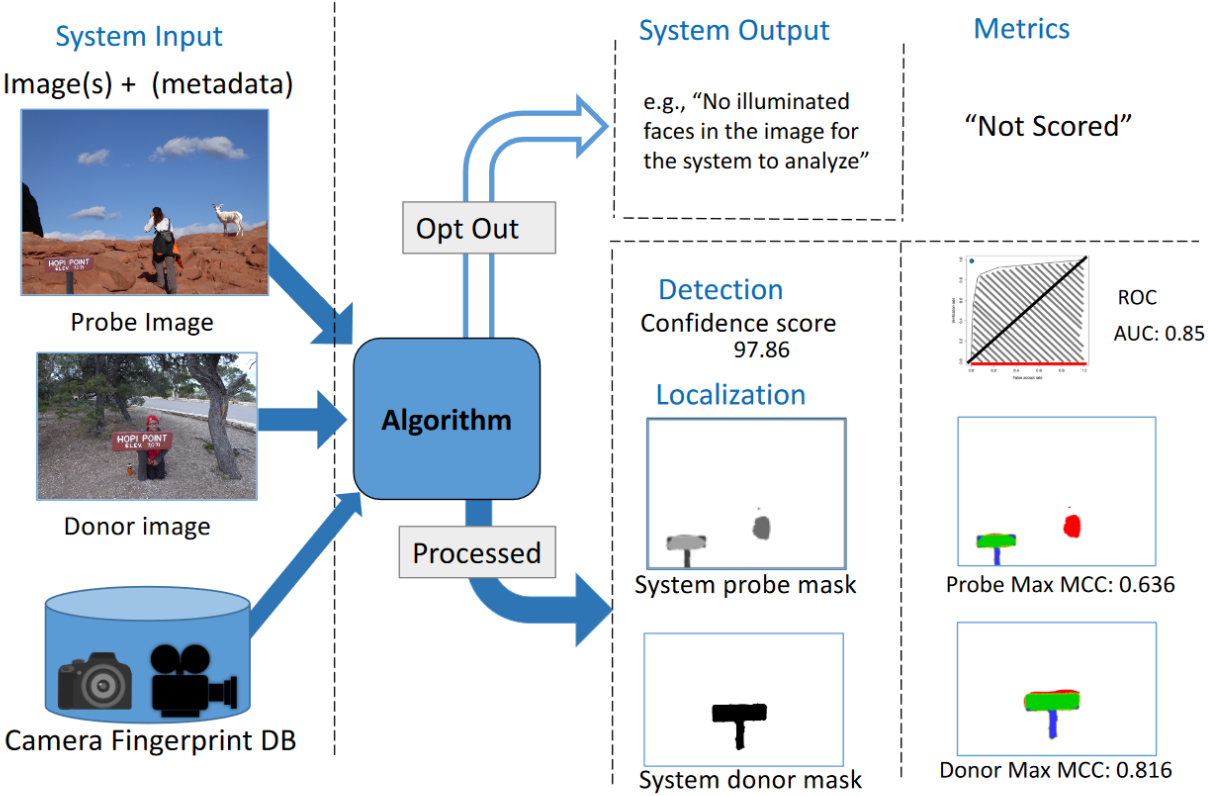


Maximum Matthews Correlation Coefficient 0.873343237591

Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nist_medifor_july17pimeeting-merge_v12.pdf

Nimble Challenge 2017

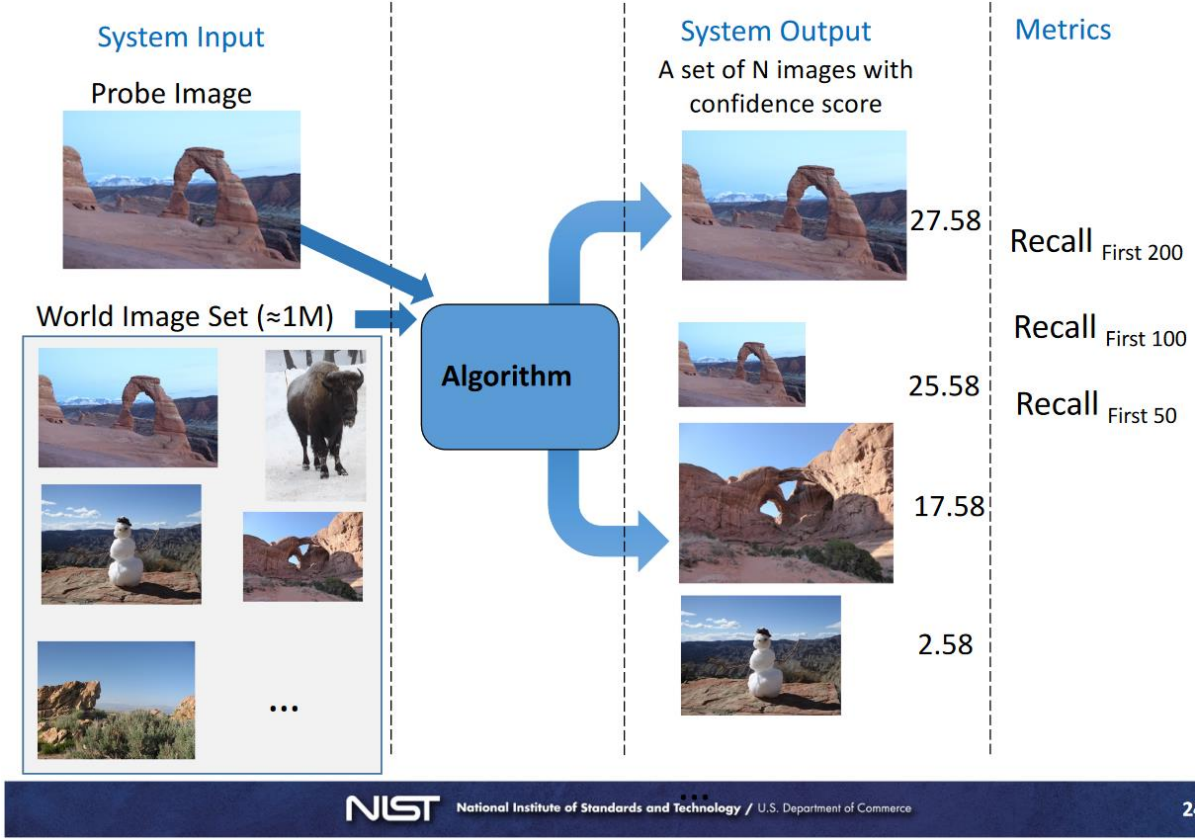
Splice Detection and Localization Evaluation Task



Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nist_medifor_july17pimeeting-merge_v12.pdf

Nimble Challenge 2017

Provenance Filtering Evaluation Task

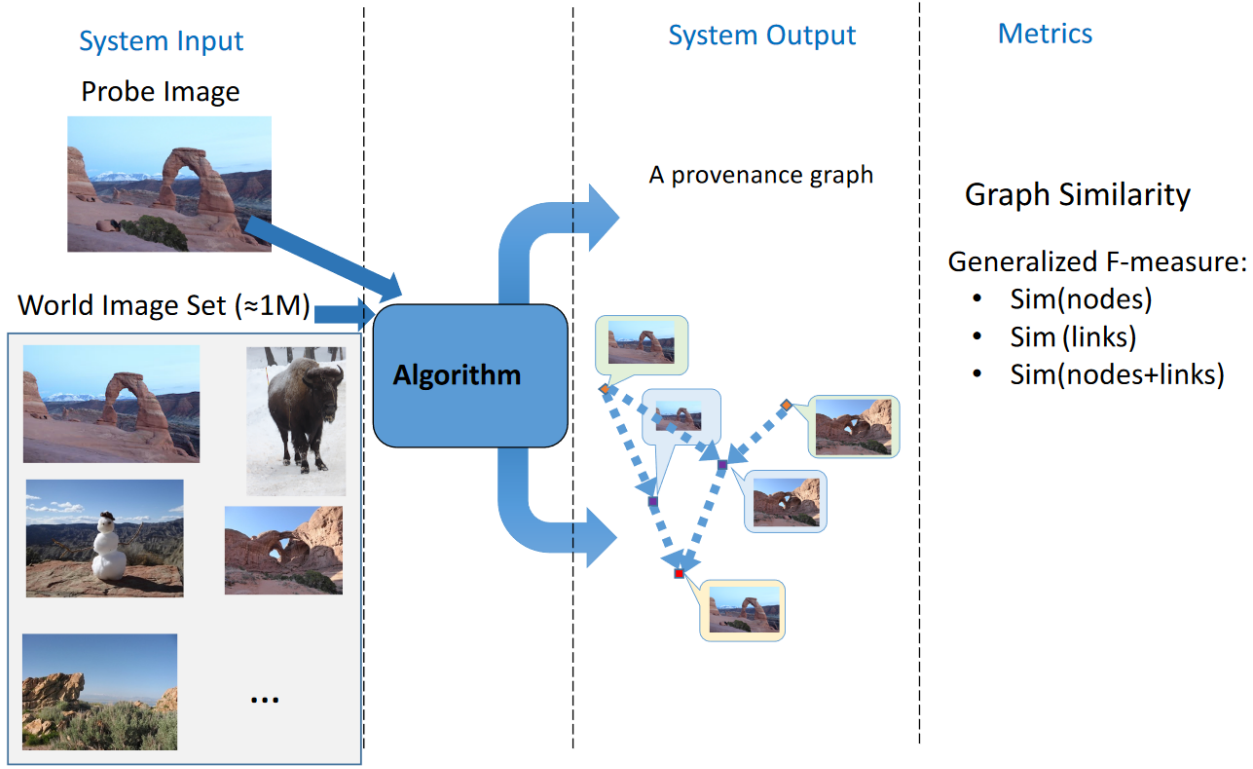


NIST National Institute of Standards and Technology / U.S. Department of Commerce

Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nist_medifor_july17pimeeting-merge_v12.pdf

Nimble Challenge 2017

Provenance Graph Building Evaluation Task

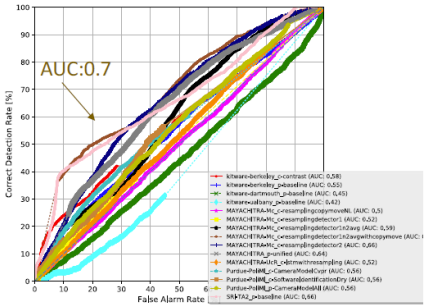


Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nist_medifor_july17pimeeting-merge_v12.pdf

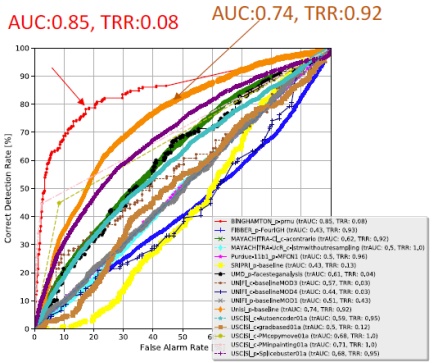
Results

NC17 Image Manipulation Detection Results

Systems That Processed All Probes



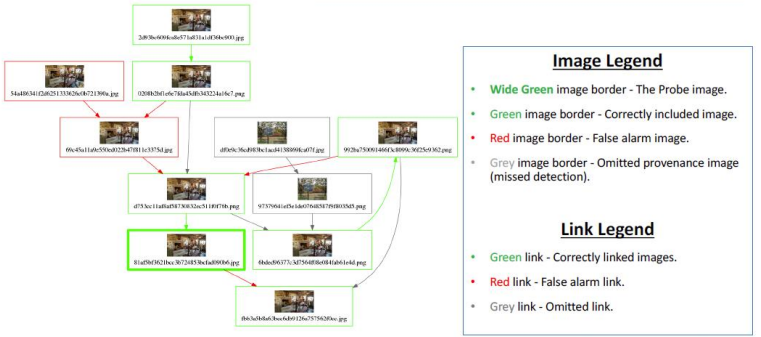
Systems that "Opted In" to Process Some of the Probes



7/28/17 NIST National Institute of Standards and Technology / U.S. Department of Commerce 16

Provenance Graph Evaluation Example: ND-Purdue, Baseline System

	Mean Similarity		
Mean Node Recall	Node Overlap	Link Overlap	Node and Link Overlap
0.778	0.778	0.375	0.588



7/28/17 NIST National Institute of Standards and Technology / U.S. Department of Commerce 29

Nimble Challenge 2017 Evaluation - Slides: for the 1 year PI meeting https://www.nist.gov/sites/default/files/documents/2017/09/05/nist_medifor_july17pimeeting-merge_v12.pdf

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AUDIO TAMPERING

Fake Audio

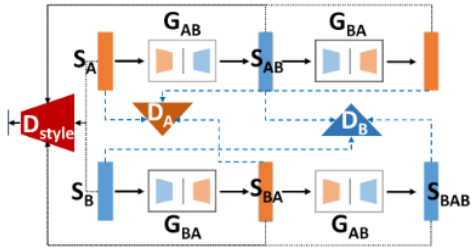
- **No Deep Fake for Audio!!!**
  - Deep Fakes with voice require
    - Traditional audio manipulation
    - Impressionist
- **Approaches to style-transfer in audio**
  - Not promising

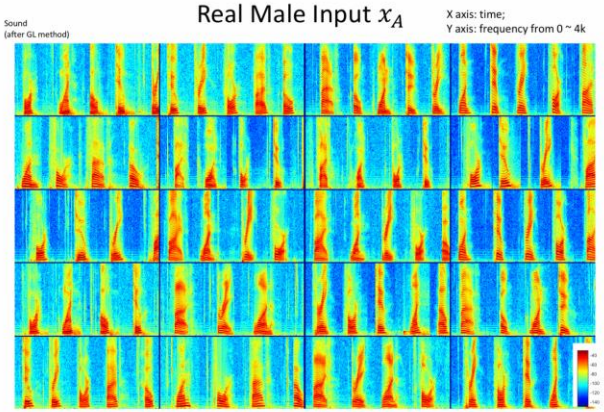
Deep Fake VFX - Pity the poor impressionist by Jim Meskimen https://www.youtube.com/watch?v=Wm3squcz7Aw

Voice-GAN

- Spectrogram-based-GAN

- Examples
  - Female
  - Female ➔ Male
  - Female ➔ Male ➔ Female





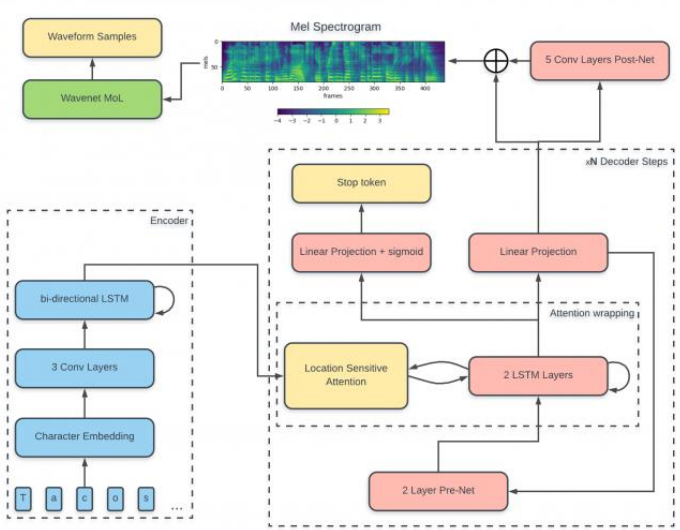
Gao, Y., Singh, R., & Raj, B. (2018, April). Voice impersonation using generative adversarial networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2506-2510). IEEE.

Speech Synthesis

- **Text to Speech**
  ➔ different approach

- **Tacotron-2**
  - Conditioned Wavenets

Model Architecture:



Tacotron-2 https://github.com/Rayhane-mamah/Tacotron-2

Tacotron-2 Examples

- Synthetic or Real?

"That girl did a video about Star Wars lipstick."

"She earned a doctorate in sociology at Columbia University."

"George Washington was the first President of the United States."

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

- Stress and intonation
- Questions
- Prosody
  - Intonation, rhythm, tone

Style / Reference

Reference text: Alice was not much surprised at this, she was getting so used to queer things happening.

Base Voice Model

Result

Perturbed text: Eric was not much surprised at this, he was getting so used to TensorFlow breaking.

Singing

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. arXiv preprint arXiv:1803.09047.

Simple but effective / No AI

Pelosi videos manipulated to make her appear drunk are being shared on social media
https://www.youtube.com/watch?v=sDOo5nDJwgA

FFmpeg

Suche

Anmelden | Einstellungen | Hilfe/Anleitung | Über Trac | Register

Wiki | Journal | Tickets anzeigen | Suche | Tags

Wiki: How to speed up / slow down a video

+1 | Aufwärts | Startseite | Inhaltsverzeichnis | Änderungshistorie

Speeding up/slowing down video

You can change the speed of a video stream using the setpts video filter. Note that in the following examples, the audio stream is not changed, so it should ideally be disabled with `-an`.

To double the speed of the video, you can use:

```
ffmpeg -i input.mkv -filter:v "setpts=0.5*PTS" output.mkv
```

The filter works by changing the presentation timestamp (PTS) of each video frame. For example, if there are two succesive frames shown at timestamps 1 and 2, and you want to speed up the video, those timestamps need to become 0.5 and 1, respectively. Thus, we have to multiply them by 0.5.

Note that this method will drop frames to achieve the desired speed. You can avoid dropped frames by specifying a higher output frame rate than the input. For example, to go from an input of 4 FPS to one that is sped up to 4x that (16 FPS):

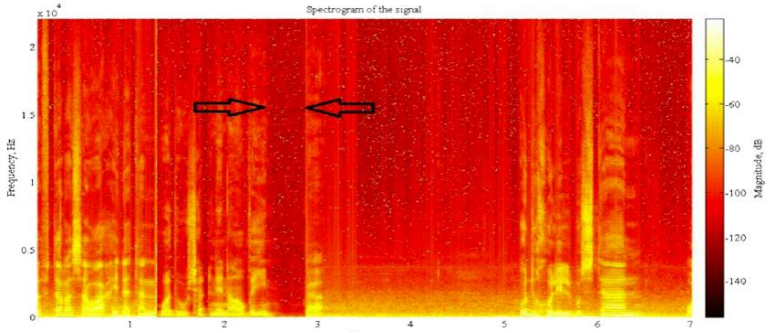
```
ffmpeg -i input.mkv -r 16 -filter:v "setpts=0.25*PTS" output.mkv
```

To slow down your video, you have to use a multiplier greater than 1:

```
ffmpeg -i input.mkv -filter:v "setpts=2.0*PTS" output.mkv
```

Audio tampering research

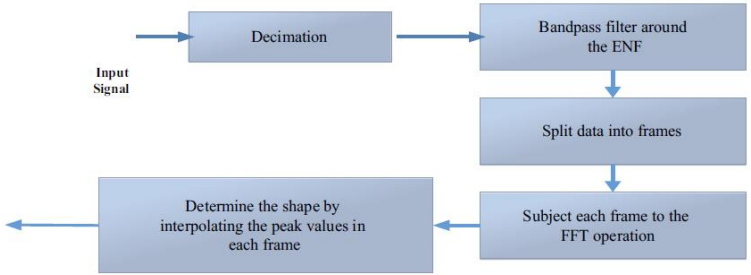
- Motivation
  - Identify modified or fabricated content
  - E.g. Forged / faked evidence in court cases
- Common Modifications
  - Splicing, copying, moving, insertions
- Common Countermeasures
  - Local noise level estimation: Splicing ➔ different noise levels
  - Exploring pitch similarity: Copy-move
  - Electric Network Frequency (ENF)
- Problems
  - Many traces get lost in compression (e.g. mp3)



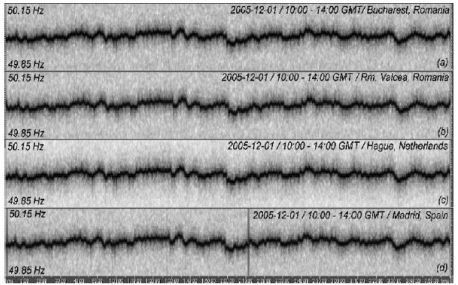
Zakariah, M., Khan, M. K., & Malik, H. (2018). Digital multimedia audio forensics: past, present and future. Multimedia tools and applications, 77(1), 1009-1040.

Electronic Network Frequency (ENF)

- Traces of ENF in the recording
  - ENF deviates from 50 to 60Hz
  - Distinct pattern
  - Similar on different networks
  - Captured in recording
- Compares
  - Estimated ENF signature of recording
  - Reference frequency database by power supply company
- Max offset for cross correlation (MOCC)
  - For query ENF and reference ENF



Zakariah, M., Khan, M. K., & Malik, H. (2018). Digital multimedia audio forensics: past, present and future. Multimedia tools and applications, 77(1), 1009-1040.

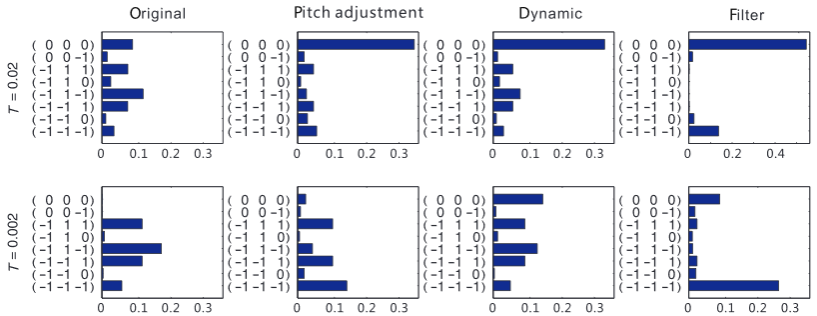


Grigoras, C. (2007). Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic science international, 167(2-3), 136-145.

Audio Postprocessing Detection

- Amplitude Cooccurrence Vectors (ACV)
- Gray Level coocurrence matrix (GLCM)
  - Joint probability of two pixels
  - Texture analysis / characterize coocurrence patterns
  - Adapted to audio
- Distinguish tampered audio
- Identify type of modification

https://www.youtube.com/watch?v=sDOo5nDJwgA



Luo, D., Sun, M., & Huang, J. (2016). Audio postprocessing detection based on amplitude cooccurrence vector feature. IEEE Signal Processing Letters, 23(5), 688-692.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

TEXT BASED FAKE NEWS DETECTION

TEXT CONTENT ANALYSIS

Mina Schütz

FAKE NEWS DETECTION: TEXT CONTENT

- **Classification task**
  - Binary or multi-classification
- **Regression task**
  - Output as a numeric score of truthfulness

- **Similar NLP concepts**
  - Fake product reviews
  - Online resumes
  - Opinion spamming
  - Fake profiles
  - Spamming and phishing

CONTENT-BASED APPRoAChES

- Headline & body text

- **Linguistic cues and patterns**
  - Character, word, sentence or document level
  - Linguistic Inquiry and Word Count (LIWC)

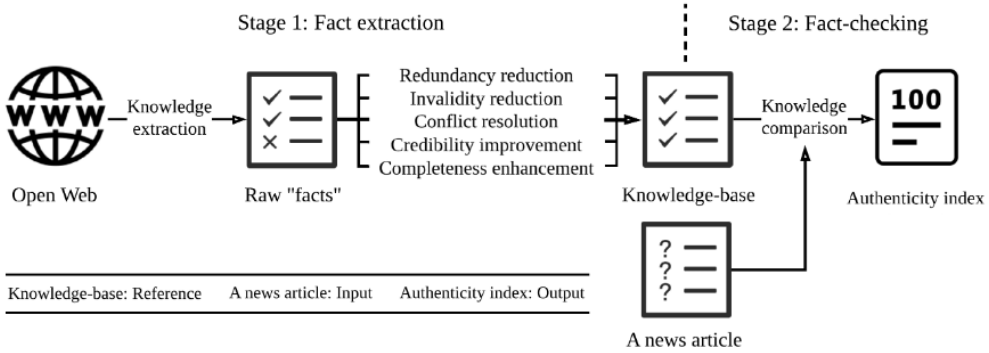
- **Style-based**
  - Hashtags, mentions, punctuation marks, sentiment
  - Topics languages, domains

Attribute Type	Feature
Quantity	Character count
	Word count
	Noun count
	Verb count
	Number of noun phrases
	Sentence count
	Paragraph count
	Number of modifiers (e.g., adjectives and adverbs)
Complexity	Average number of clauses per sentence
	Average number of words per sentence
	Average number of characters per word
	Average number of punctuations per sentence
Uncertainty	Percentage of modal verbs
	Percentage of centainty terms
	Percentage of generalizing terms
	Percentage of tentative terms
	Percentage of numbers and quantifiers
	Number of question marks
Subjectivity	Percentage of subjective verbs
	Percentage of report verbs
	Percentage of factive verbs
	Percentage of imperative commands

Knowledge-based approaches

- Check for truthfulness of claims

- **Manual fact-checking**
  - Expert-based
  - Crowdsources
  - politifact.com, snopes.com

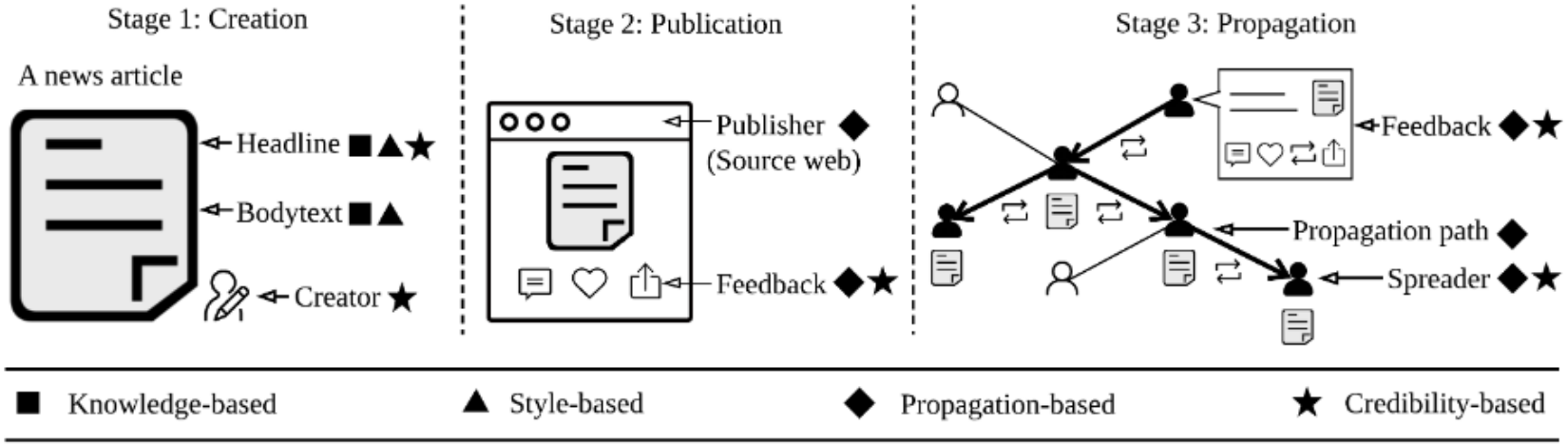


- **Automatic fact-checking**
  - Linked Open Data (i.e. DBpedia)
  - Fact-extraction (Knowledge base construction)
  - Fact-checking (Knowledge comparison

SOCIAL-CONTEXT Based APPROACHES

- **Stance-based**
  - Stance of body text relative to the headline claim
  - Viewpoint of user
  - Infer validity of original article
  - Support or refute claim
- **User-based**
  - Registration age
  - Numer of followers / followees
- **Propagation-based**
  - Propagation networks
  - i.e. Twitter shares / retweets
  - Likes
- **Credibility-based**
  - Headlines (clickbaits)
  - News source, spreaders, author

HYBRID APPROACHES



preprocessing

- Tokenization
- Stemming
- Lemmatization
- Part-of-Speech Tagging
- Generalization
- Vector representation
  - Bag-of-Words, N-Gram, LDA
  - TF, TF-IDF
  - Word2Vec, GloVe
  - WordPiece Model
  - FastText

- **Datasets**
  - FEVER
  - LIAR
  - CREDBANK
  - BuzzFace / BuzzFeedNews
  - PHEME
  - Some-like-it-hoax
  - Kaggle datasets (WSDM)
  - FakeNewsNet
  - Fake News Challenge (FNC-1)
  - GermanFakeNC

Classification MODELS & Evaluation

- **Supervised**
  - Naïve Bayes Classifier
  - Logistic Regression
  - Decision Trees / Random Forest
  - Support Vector Machines
- **Unsupervised**
  - Recurrent Neural Network
  - Long Short-Term Memory
  - Convolutional Neural Networks
  - Uni- or bidirectional Transformers

- **Evaluation Metrics**
  - Precision
  - Recall
  - F1
  - Accuracy
  - ROC
  - ROC AUC
  - Cross-fold-validation

CHALLENGES

- Datasets are biased, too small or not available anymore
  - Domain / dataset specific
  - Many features to choose from
  - Model learns writing style of a news source (not if it is fake news)
  - Often only linguistic cues or wordcount without context

- Almost no knowledge-based approaches

- No common fake news definition
  - No standardized approach available (Accuracy from ~ 20% to ~ 95%)
  - Early fake news identification

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

FAKE NEWS PROBLEM

Overestimated?

Technical Problem?

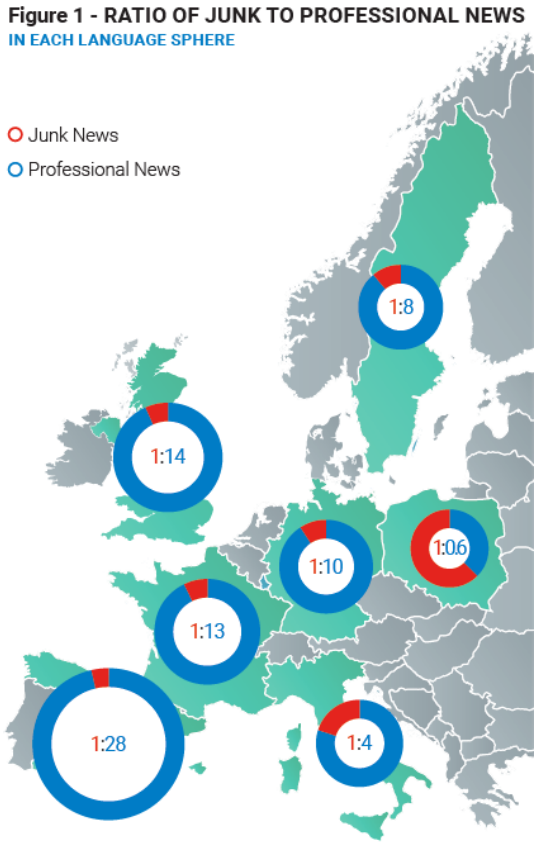
Solvable?

Conclusions

- **General**
  - Fake News detection currently a hyped topic
  - Fear mongering to create business opportunities
- **„Traditional“ attacks**
  - Media Forensic
    - traditional research fields
    - Many well established approaches
    - Ongoing evaluation campaigns
- **AI-Threat**
  - Low-Res GAN – no threat
  - GAN content identifiable (incl. Deep Fakes)
  - Challenge: future Hi-Res GANs
- **Biggest Challenge**
  - Text content analysis
  - Claim / Fact checking

Extent of Fake News in EU Parliament Elections

- Oxford Institute study
  - 7 different languages
- **Only 4% source articles ➔ Fake**
  - ~600M Tweets
  - ~140K Users
  - ~6K unique articles



Conclusions

- **General Discussion about Fake News**
  - Technological Problem?
  - Sociological Problem?
  - Political Problem?
- **Who is responsible?**
  - Tech-Companies (Twitter, Facebook, etc.)
- **Most cited solutions**
  - Renessaince of quality journalism
  - Promote media literacy

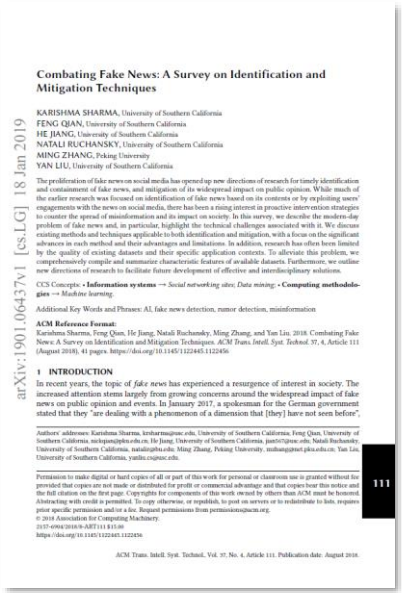
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

RECOMMENDED LITERATURE

Surveys

- Kumar, S., & Shah, N. (2018). **False information on web and social media: A survey**. *arXiv preprint arXiv:1804.08559*.

- Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). **Combating fake news: A survey on identification and mitigation techniques**. ACM Transactions on Intelligent Systems and Technology (TIST), 10(3), 21.



Combating Fake News: A Survey on Identification and Mitigation Techniques

KARISHMA SHARMA, University of Southern California
FENG QIAN, University of Southern California
HE JIANG, University of Southern California
NATALI RUCHANSKY, University of Southern California
MING ZHANG, Peking University
YAN LIU, University of Southern California

The proliferation of fake news on social media has opened up new directions of research for timely identification and containment of fake news, and mitigation of its widespread impact on public opinion. While much of the earlier research was focused on identification of fake news based on its contents or by exploiting users' engagements with the news on social media, there has been a rising interest in proactive intervention strategies to counter the spread of misinformation and its impact on society. In this survey, we describe the modern-day problem of fake news and, in particular, highlight the technical challenges associated with it. We discuss existing methods and techniques applicable to both identification and mitigation, with a focus on the significant advances in each method and their advantages and limitations. In addition, research has often been limited by the quality of existing datasets and their specific application contexts. To alleviate this problem, we comprehensively compile and summarize characteristic features of available datasets. Furthermore, we outline new directions of research to facilitate future development of effective and interdisciplinary solutions.

CCS Concepts: • **Information systems** → *Social networking sites*; *Data mining*; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: AI, fake news detection, rumor detection, misinformation

ACM Reference Format:
Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2018. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 37, 4, Article 111 (August 2018), 41 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In recent years, the topic of *fake news* has experienced a resurgence of interest in society. The increased attention stems largely from growing concerns around the widespread impact of fake news on public opinion and events. In January 2017, a spokesman for the German government stated that they "are dealing with a phenomenon of a dimension that [they] have not seen before",

Authors' addresses: Karishma Sharma, ksharma@usc.edu, University of Southern California, Feng Qian, University of Southern California, nickqian@phs.edu.cn, He Jiang, University of Southern California, jiaxi1@usc.edu, Natali Ruchansky, University of Southern California, natali@phs.edu, Ming Zhang, Peking University, mzhang@net.pku.edu.cn, Yan Liu, University of Southern California, yanliu.cs@usc.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
2157-6904/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

ACM Trans. Intell. Syst. Technol., Vol. 37, No. 4, Article 111. Publication date: August 2018.

111

arXiv:1901.06437v1 [cs.LG] 18 Jan 2019

False Information on Web and Social Media: A Survey

SRIJAN KUMAR, Computer Science, Stanford University, USA
NEIL SHAH*, Computer Science, Carnegie Mellon University, USA

False information can be created and spread easily through the web and social media platforms, resulting in widespread real-world impact. Characterizing how false information proliferates on social platforms and why it succeeds in deceiving readers are critical to develop efficient detection algorithms and tools for early detection. A recent surge of research in this area has aimed to address the key issues using methods based on feature engineering, graph mining, and information modeling. Majority of the research has primarily focused on two broad categories of false information: opinion-based (e.g., fake reviews), and fact-based (e.g., false news and hoaxes). Therefore, in this work, we present a comprehensive survey spanning diverse aspects of false information, namely (i) the actors involved in spreading false information, (ii) rationale behind successfully deceiving readers, (iii) quantifying the impact of false information, (iv) measuring its characteristics across different dimensions, and finally, (iv) algorithms developed to detect false information. In doing so, we create a unified framework to describe these recent methods and highlight a number of important directions for future research.[1]

Additional Key Words and Phrases: misinformation, fake news, fake reviews, rumors, hoaxes, web, internet, social media, social networks, bad actors, bots, propaganda, conspiracy, knowledge bases, e-commerce, disinformation, impact, mechanism, rationale, detection, prediction

ACM Reference Format:
Srijan Kumar and Neil Shah. 2018. False Information on Web and Social Media: A Survey. 1, 1 (April 2018), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The web provides a highly interconnected world-wide platform for everyone to spread information to millions of people in a matter of few minutes, at little to no cost [10]. While it has led to ground-breaking phenomenon such as real-time citizen journalism [34], at the same time it has led to increased visibility and impact of both true and false information [37]. False information on the web and social media has affected stock markets [17], slowed responses during disasters [32], and terrorist attacks [27, 96]. Recent surveys have alarmingly shown that people increasingly get their news from social media than from traditional news sources [75, 88], making it of paramount importance to curtail false information on such platforms. With primary motives of influencing opinions and earning money [1, 46, 56, 94], the wide impact of false information makes it one of the modern dangers to society, according to the World Economic Forum [39]. Understanding the reasons for why and how false information is created is important to proactively detect it and mitigate its impact. In this survey, we review the state of the art scientific literature on false information on the web and social media to give a comprehensive description of its ***mechanisms, rationale, impact, characteristics, and detection.*** While recent surveys have focused on fake

*Dr. Shah is now at Snap Inc.
[1]A previous version of this survey will appear in the book titled Social Media Analytics: Advances and Applications, by CRC press, in 2018.

Authors' addresses: Srijan Kumar, Computer Science, Stanford University, USA, srijan@cs.stanford.edu; Neil Shah, Computer Science, Carnegie Mellon University, USA, neil@cs.cmu.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2018 Copyright held by the owner/author(s).
XXXX-XXXX/2018/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

, Vol. 1, No. 1, Article . Publication date: April 2018.

arXiv:1804.08559v1 [cs.SI] 23 Apr 2018

Bibliography

- Graves, Lucas (2018). Understanding the Promise and Limits of Automated Fact-Checking. *Factsheet February 2018*, Reuters Institute, 1 – 8. https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking.
- Figueira, Álvaro; Oliveira, Luciana (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science Vol. 121*, 817 – 825. https://www.sciencedirect.com/science/article/pii/S1877050917323086. 10.1016/j.procs.2017.11.106.
- Mahid, Zaitul Iradah; Manickam, Selvakumar; Karuppayah, Shankar (2018). Fake News on Social Media: Brief Review on Detection Techniques. *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia, 1 – 5. https://ieeexplore.ieee.org/document/8776689. 10.1109/ICACCAF.2018.8776689.
- Oshikawa, Ray; Qian, Jing; Wang, William Yang (2018). A Survey on Natural Language Processing for Fake News Detection. 1 – 11. https://arxiv.org/abs/1811.00770.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Thank you!

Alexander Schindler and Mina Schütz

25.11.2019