Multimodal content analysis in the service of Journalism

Dr. Symeon Papadopoulos
Principal Researcher @ CERTH/ITI
Head of MeVer group

mever
Media Analysis, Verification and Retrieval Group

Overview

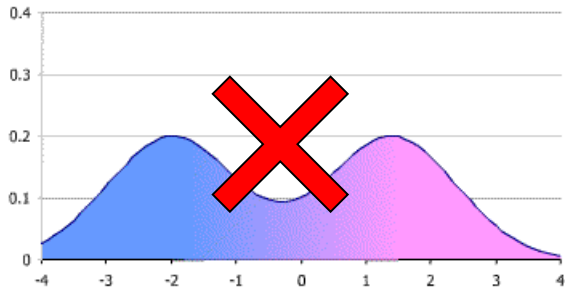
- multimodal content analysis - basics
- internet meme analysis
- multimodal misinformation detection
- multimodal deepfake detection

Modalities

The different forms/formats in which input information is available for analysis. (of course, output can also be multimodal and this is esp. true for generative AI)

- Text
- Image
- Audio
- Video
- 3D

Senses map to modalities, e.g. sight → image, hearing → audio, etc.



NOT multimodal distribution!

Single-modality video games (back in the 80s)

```
Copyright (c) 1981, 1982, 1983 Infocom, Inc. All rights reserved.
ZORK is a registered trademark of Infocom, Inc.
Revision 88 / Serial number 840726

West of House
You are standing in an open field west of a white house, with a boarded front
door.
There is a small mailbox here.

>open mailbox
Opening the small mailbox reveals a leaflet.

>take the leaflet
Taken.

>read leaflet
"WELCOME TO ZORK!

ZORK is a game of adventure, danger, and low cunning. In it
some of the most amazing territory ever seen by mortals. No
without one!"

>_
```

Text in
Text out

https://www.museumofplay.org/blog/new-interest-in-old-text-based-adventures/

Multimodal Interactions (2023)

Prompt:

Localize each person in the image using bounding box. What is the image size of the input image?

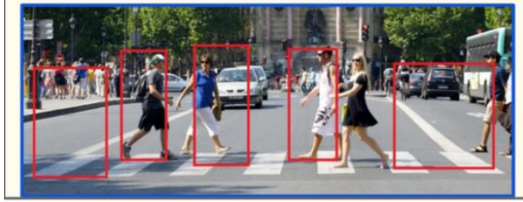
GPT-4V:

The image size of the input image is 1000 x 333 pixels.

Bounding boxes for each person:
- Person 1 (leftmost): (x1, y1) = (22, 108), (x2, y2) = (170, 300)
- Person 2: (x1, y1) = (200, 80), (x2, y2) = (290, 270)
- Person 3: (x1, y1) = (350, 70), (x2, y2) = (460, 280)
- Person 4: (x1, y1) = (540, 80), (x2, y2) = (640, 270)
- Person 5 (rightmost): (x1, y1) = (760, 100), (x2, y2) = (920, 290)

Note: These bounding box coordinates are approximate and may not be exact.

Visualization of Detected Bbox:



Prompt:

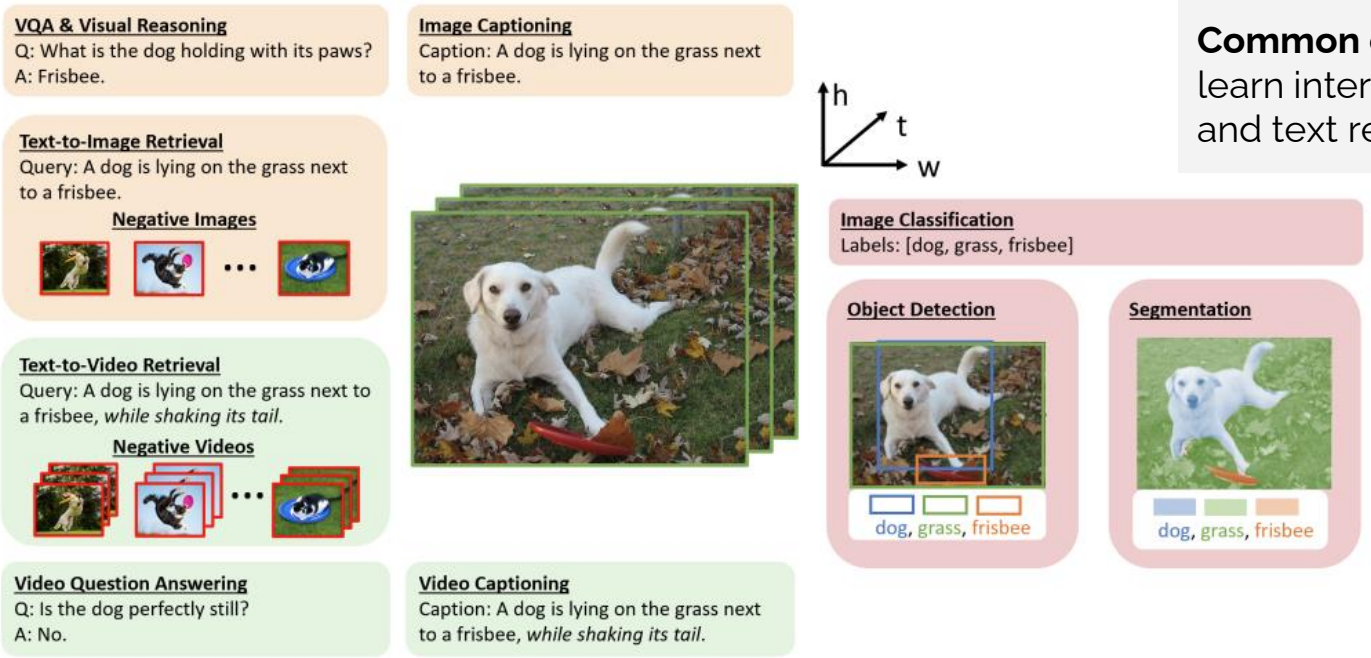
Count the number of apples in the image.

GPT-4V:

There are 7 apples in the image.

https://huyenchip.com/2023/10/10/multimodal.html

Multimodal tasks - Vision Language Understanding



Common challenge:
learn interplay between visual and text representations

Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this paper: image-text tasks , vision tasks as VL problems , and video-text tasks .

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends in Computer Graphics and Vision, 14(3–4), 163-352.

Multimodal tasks and Journalism

A few typical journalistic tasks:

- Monitoring (tens or hundreds of posts, articles, images and videos per day)
- “Parsing” (read text, inspect image, skim through video)
- Understanding (who? what? where? ….)
- Searching (find additional evidence, reverse image search, ….)
- Verifying (trying to establish the authenticity of image/video)

Today's menu of methods and Journalism

Internet Meme Analysis

image, text*
(*embedded in image)

monitoring / search

Misinformation Detection

image, text

monitoring / verification

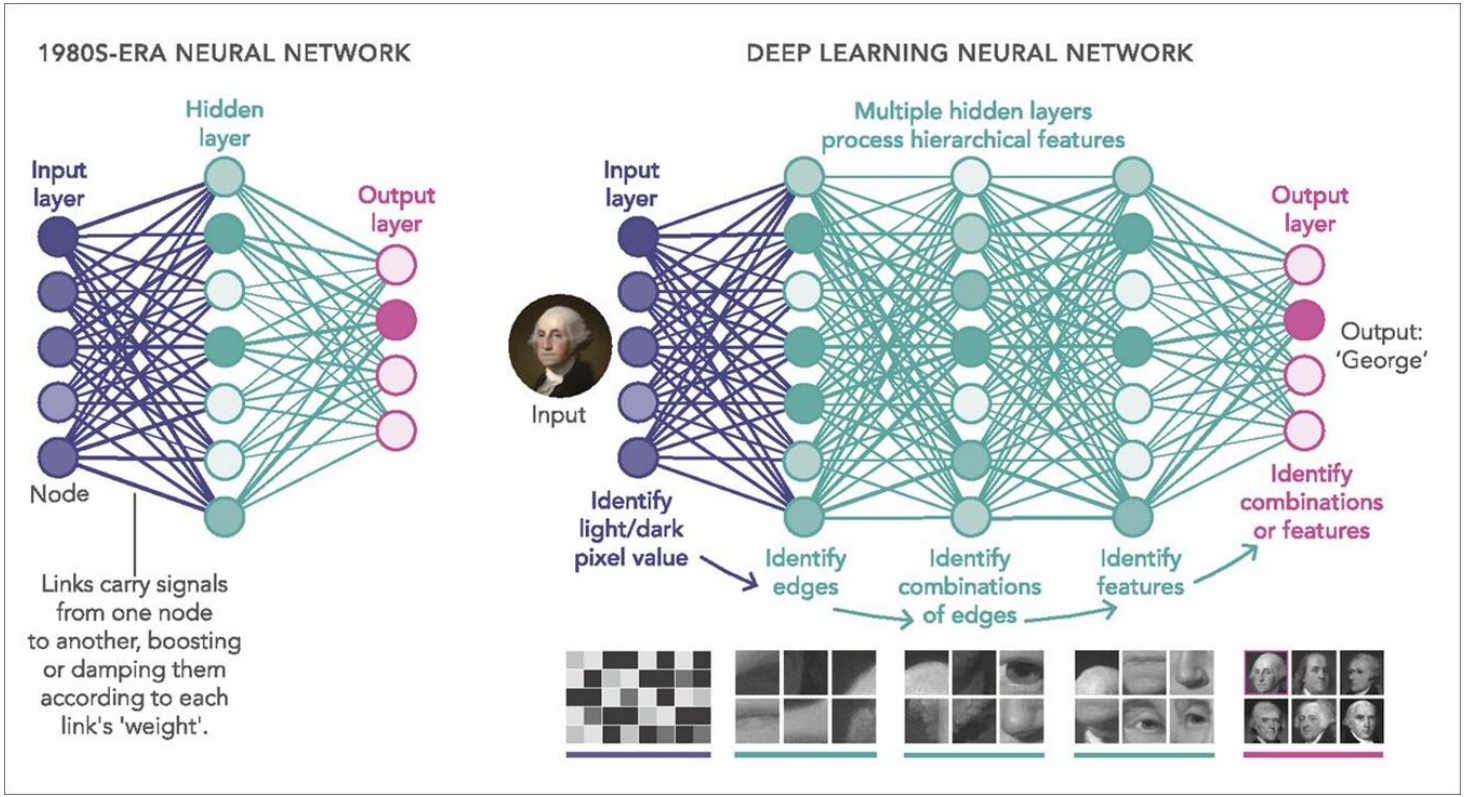
DeepFake Detection

video, audio

verification

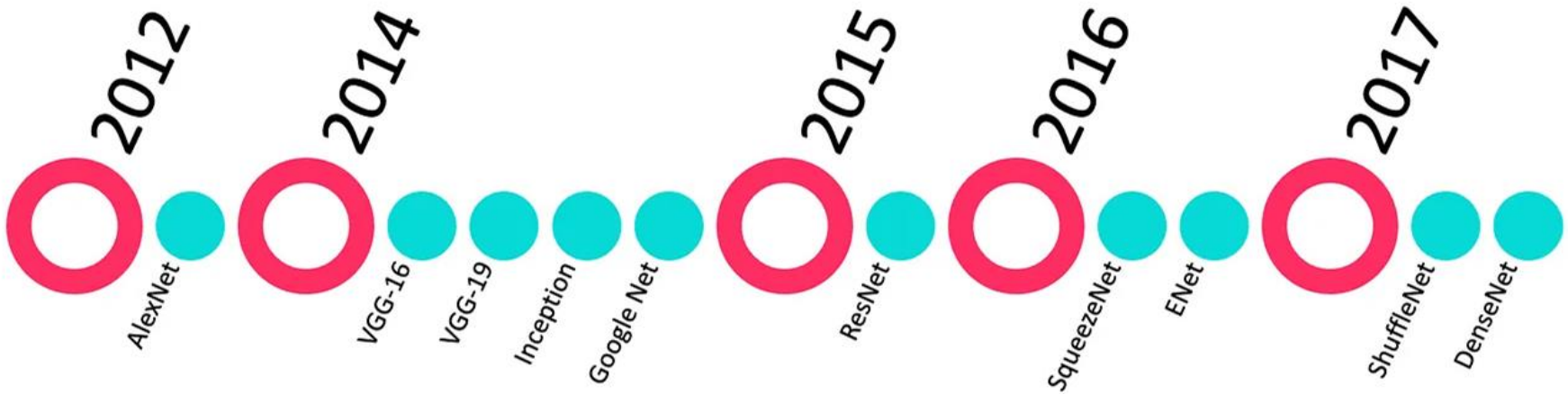
A 5-min Primer to the Latest Advances in Deep Learning

Neural networks



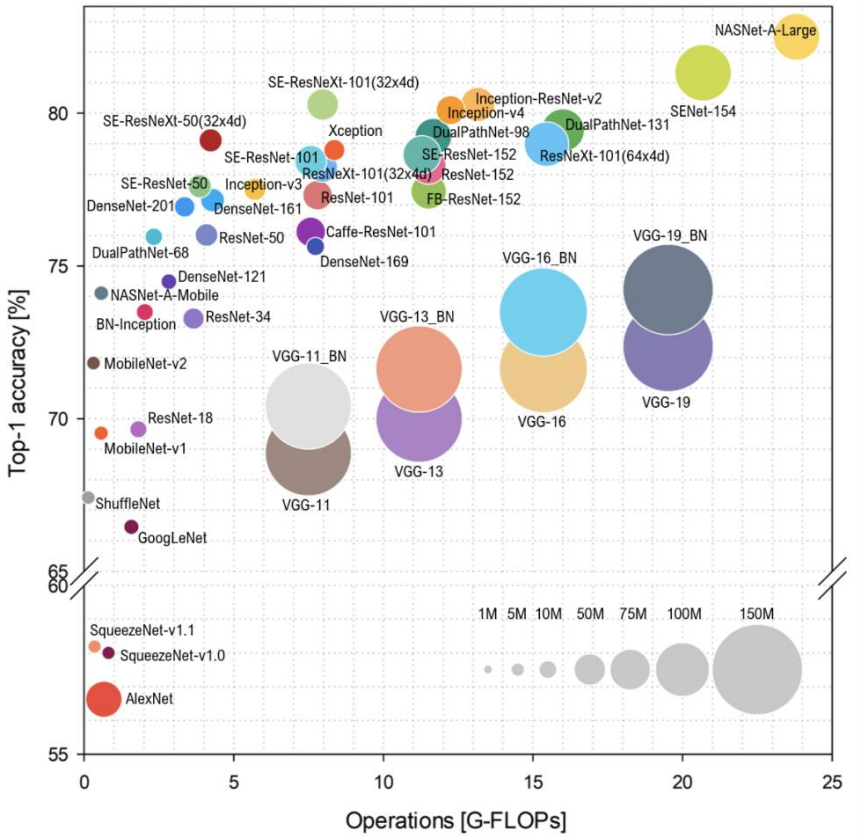
https://www.pnas.org/content/116/4/1074 (artist Lucy Reading-Ikkanda)

The Era of Deep Neural Networks



https://towardsdatascience.com/top-10-cnn-architectures-every-machine-learning-engineer-should-know-68e2b0e07201

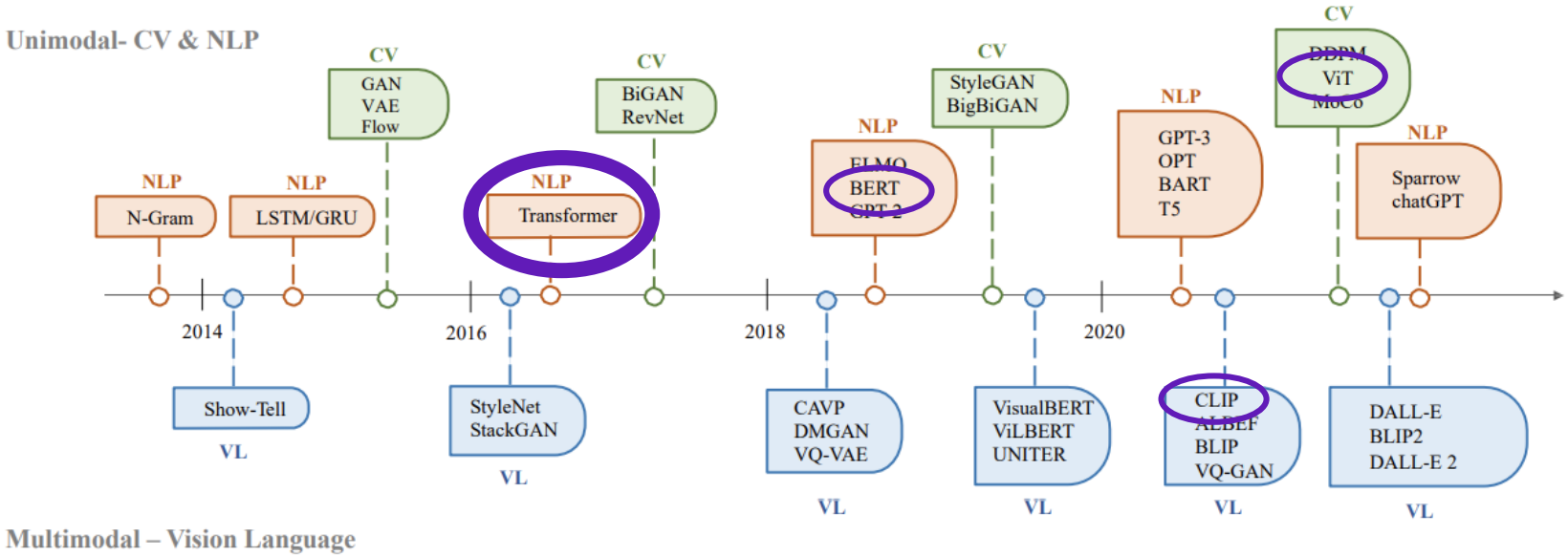
Increasing performance over time



CNNs
ResNet
InceptionNet
MobileNet
EfficientNet
Transformers
DeiT
LeViT
EfficientFormer
…..

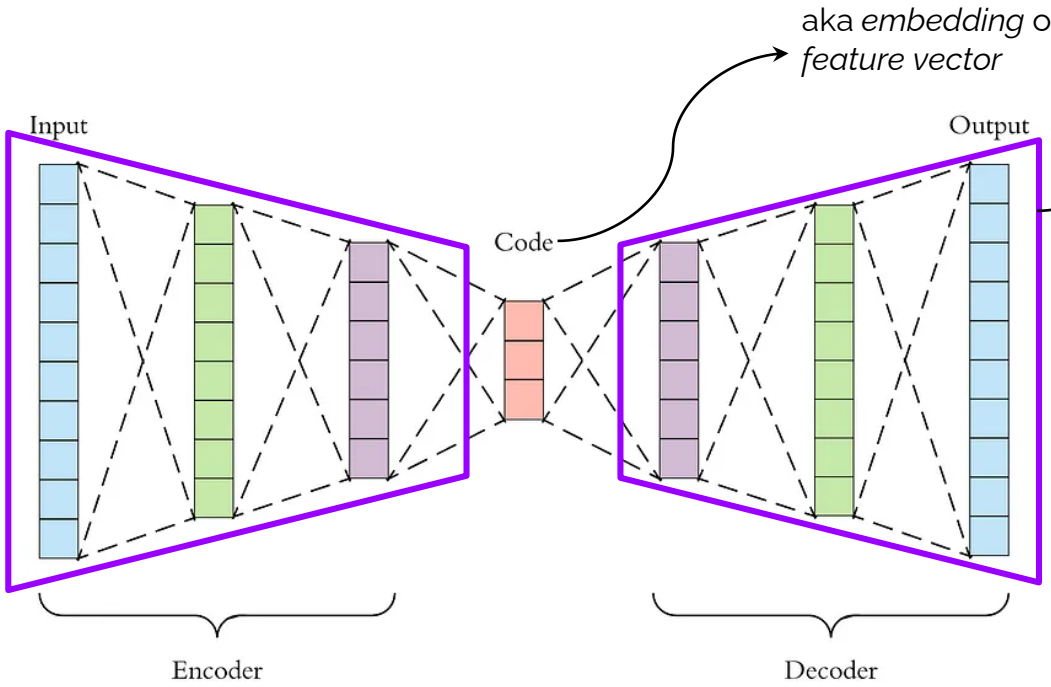
https://redding.dev/architecture-benchmarks/

A Brief History of CV, NLP & VL Architectures



Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Encoder - Decoder



Neural Network

Images

Convolutional Neural Network (CNN)
Vision Transformer (ViT)

Text

Recurrent Neural Network (RNN)
Transformer
(note that within such Text Encoders Word Embeddings may be used, e.g. Word2Vec)

What does an Encoder do?

Compression of information
Feature extraction
(or representation learning)

https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Transformer (or Attention is All You Need)

Perhaps the most influential work of the last decade in AI and the backbone of most modern AI-based VLU systems. At its core, a **Seq-to-Seq** model.

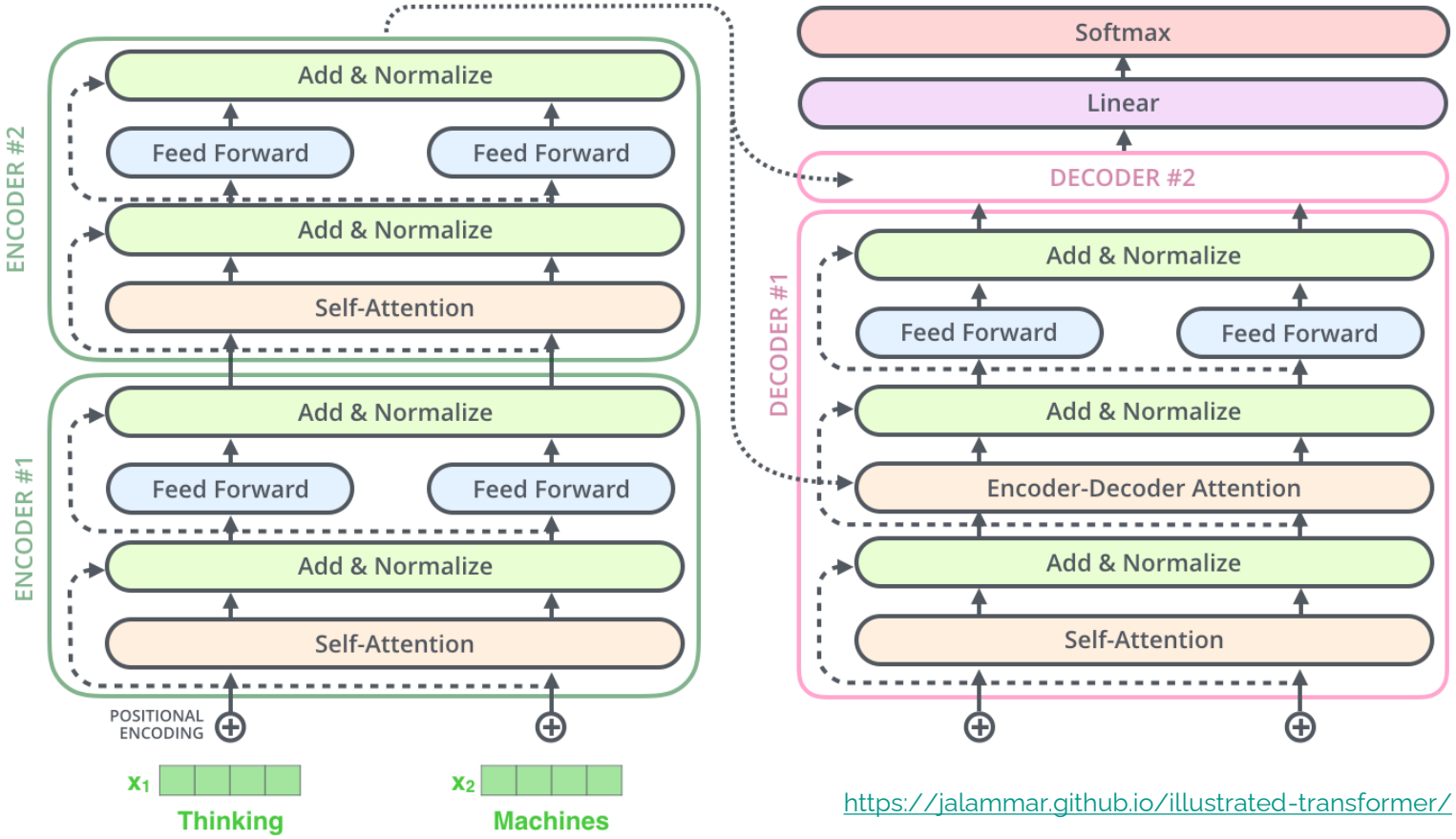
Several innovations:

- Self-attention
- Multi-attention head
- Positional encoding
- Largely parallel computations

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

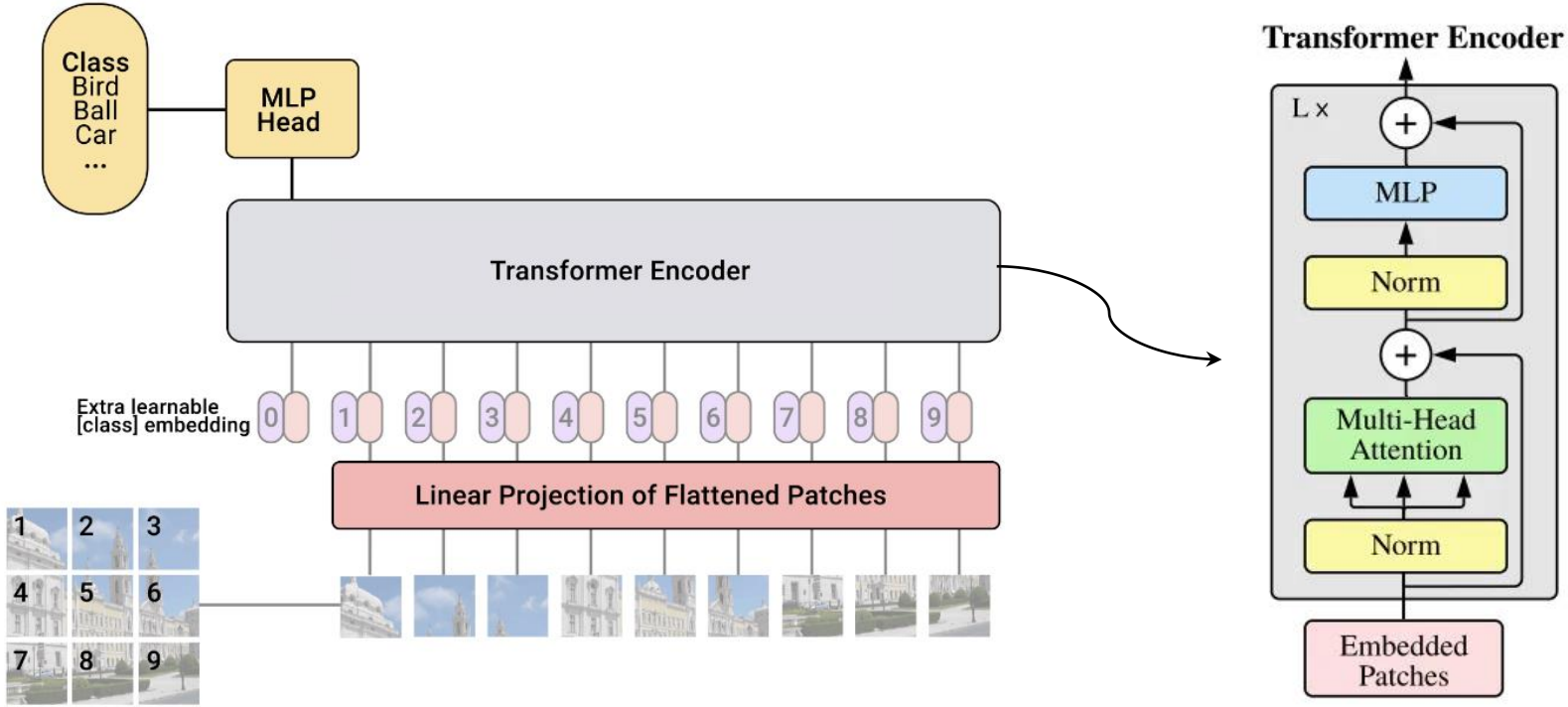
https://jalammar.github.io/illustrated-transformer/

Just a Glimpse into the Transformer



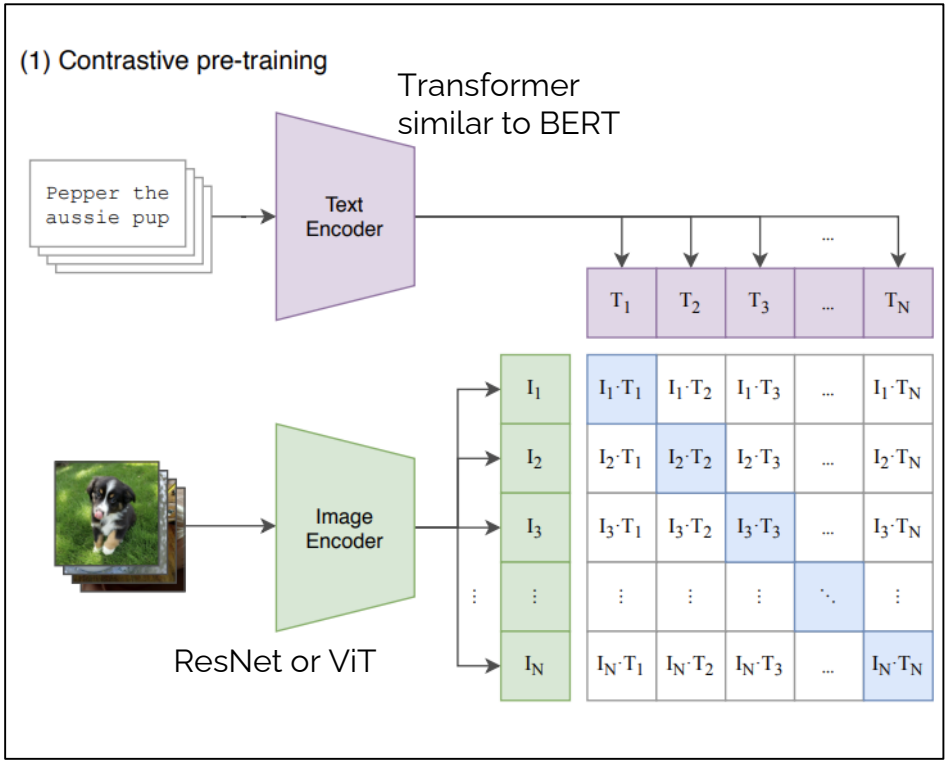
https://jalammar.github.io/illustrated-transformer/

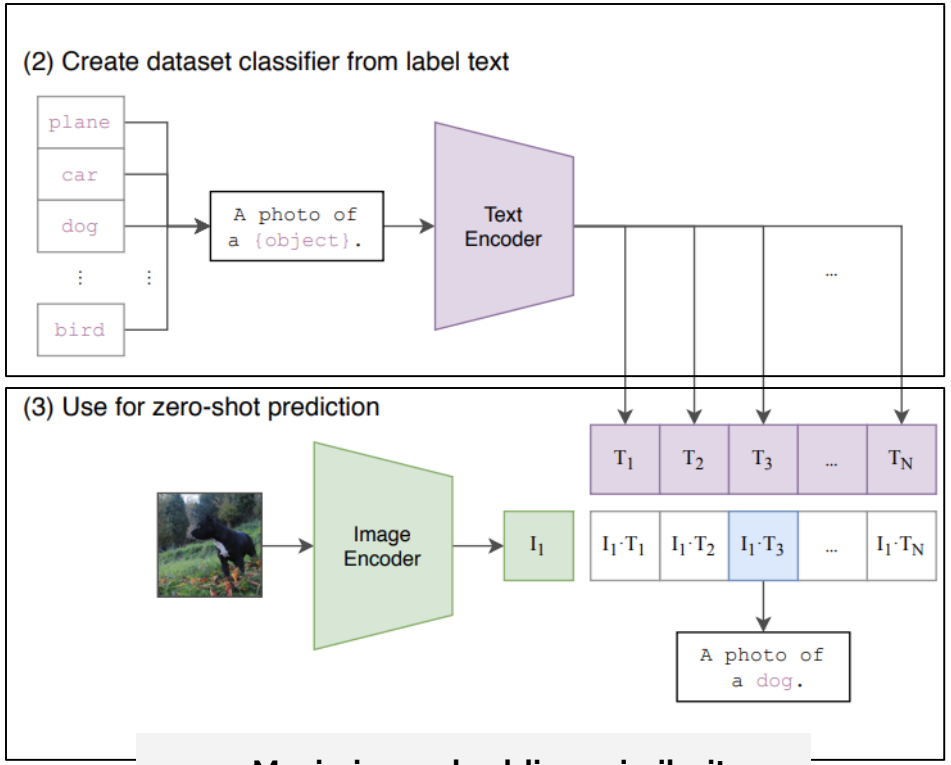
Vision Transformer (ViT)



https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html

Contrastive Language Image Pretraining (CLIP)



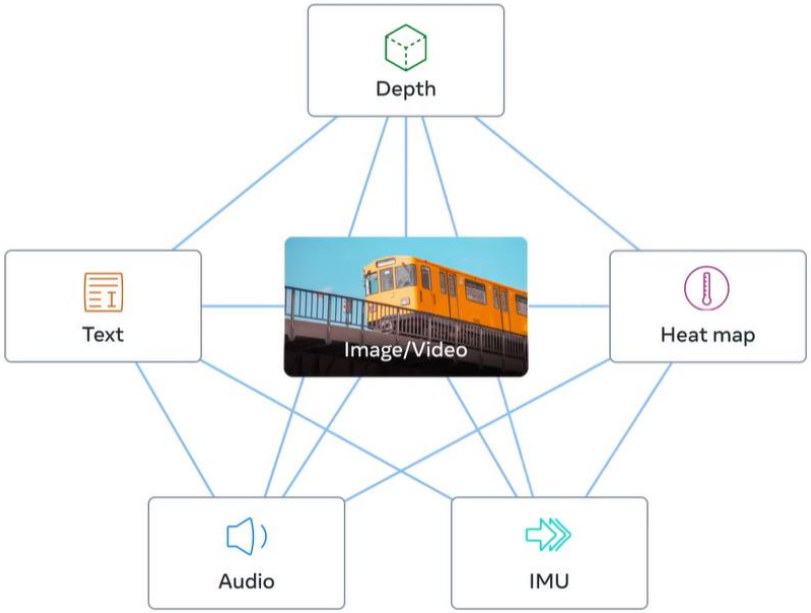


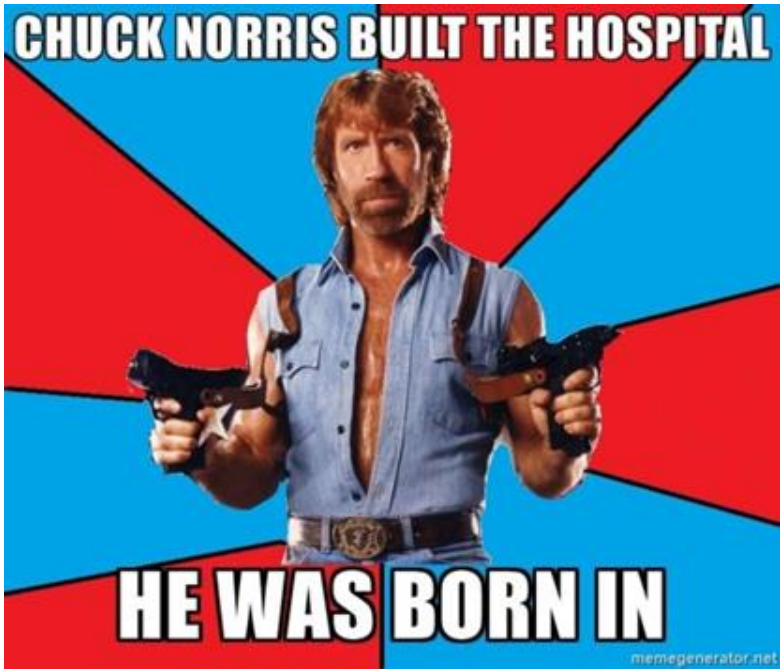
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

- **Maximize embeddings similarity**
- **Natural language supervision**
- **Contrastive learning**

The Era of Large Language Models

- CNNs, Encoders-Decoders, and mostly Transformers have been the stepping stone for Large Language Models
- ChatGPT is based on the Generative Pretrained Transformer (GPT) architecture
- Recently, there’s a trend towards Large Multimodal Models (LMMs)
- Open models: FLAVA, KOSMOS-2, LXMERT





Internet meme analysis

Internet image memes analysis

Memes: a popular visual means of communication in the Internet used to express humour, irony, sarcasm and even hate.

(a) image macros

(b) object labelling

(c) screenshots

(d) text-out-of-image

Rogers, R., & Giorgi, G. (2023). What is a meme, technically speaking? Information, Communication & Society, 1-19.

Internet image memes analysis

Meme detection

Discriminate between meme images and regular images. Primarily a vision task.

Binary classification

Classification multimodal memes into categories such as hateful/ non-hateful (FHM), troll/non-troll (TamilMemes), offensive/ non-offensive (MultiOFF).

Multi-class/-label classification

Classify memes in >2 classes, such as sentiment recognition (Memeotion7k) or propaganda type (Propaganda Memes).

Koutlis, C., Schinas, M., & Papadopoulos, S. (2023). MemeTector: Enforcing deep focus for meme detection. International Journal of Multimedia Information Retrieval, 12(1), 11.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in neural information processing systems, 33, 2611-2624.

Kumar, G. K., & Nandakumar, K. (2022). Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. In Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Meme detection with MemeTector

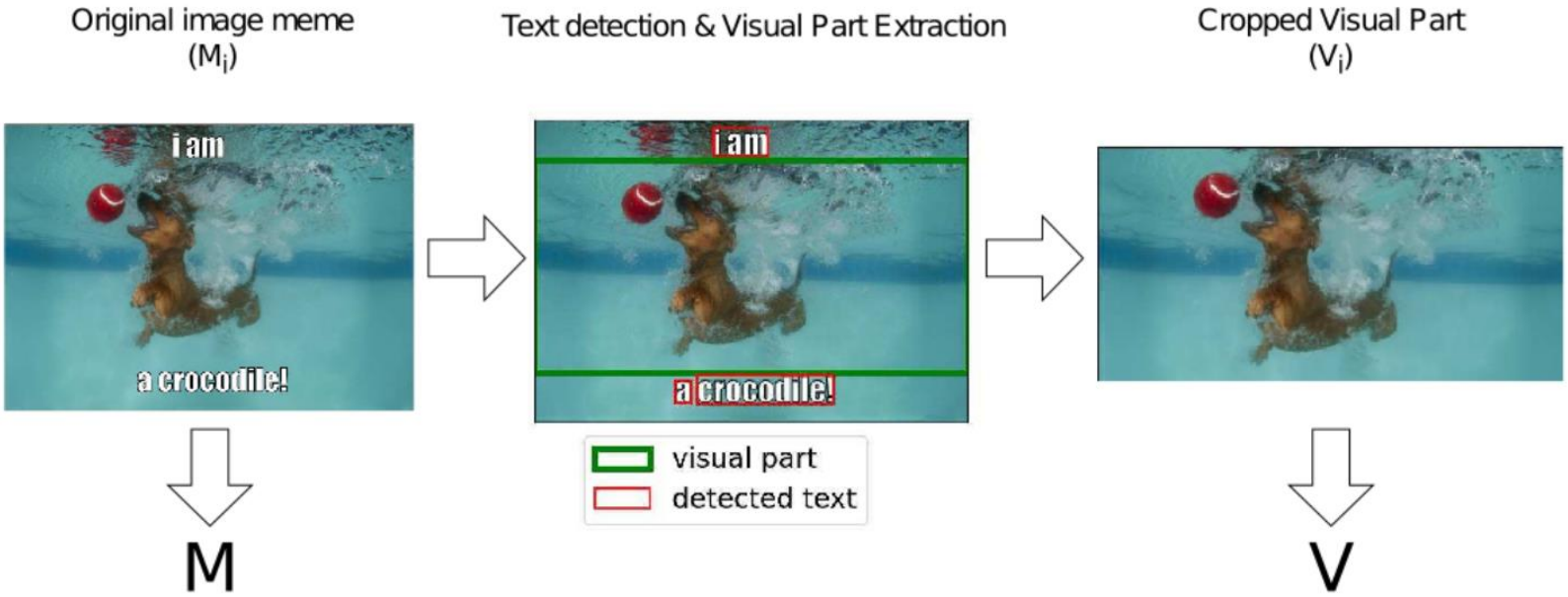
Memes and regular images have key visual differences easily detectable by humans

VS

Koutlis, C., Schinas, M., & Papadopoulos, S. (2023). MemeTector: Enforcing deep focus for meme detection. IJMR, 12(1), 11.

Meme detection with MemeTector

Considering memes' background images as regular forces the model to focus on differences.



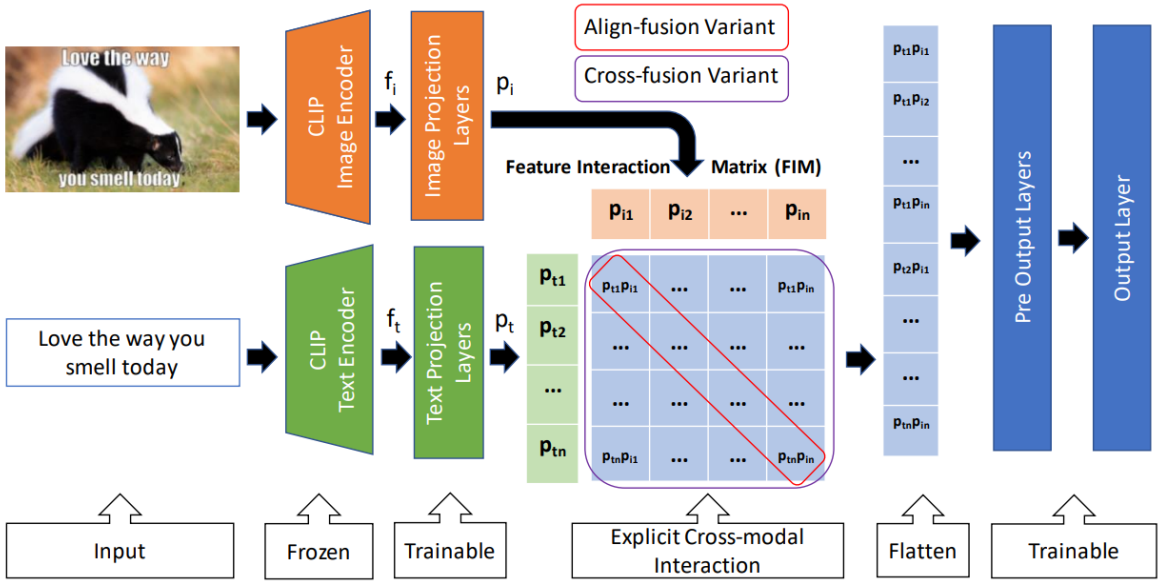
Koutlis, C., Schinas, M., & Papadopoulos, S. (2023). MemeTector: Enforcing deep focus for meme detection. IJMR, 12(1), 11.

Meme detection with MemeTector

What does the model see?
Text elements mostly - font family, size, position

Koutlis, C., Schinas, M., & Papadopoulos, S. (2023). MemeTector: Enforcing deep focus for meme detection. IJMR, 12(1), 11.

Multimodal hate-speech detection with Hate-CLIPper



Performance

85.8% test accuracy on Hateful Memes Challenge benchmark - with just 5M params!

Kumar, G. K., & Nandakumar, K. (2022). Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. In Proceedings of 2nd Workshop on NLP for Positive Impact (NLP4PI), pages 171–183, Abu Dhabi, United Arab Emirates. ACL.

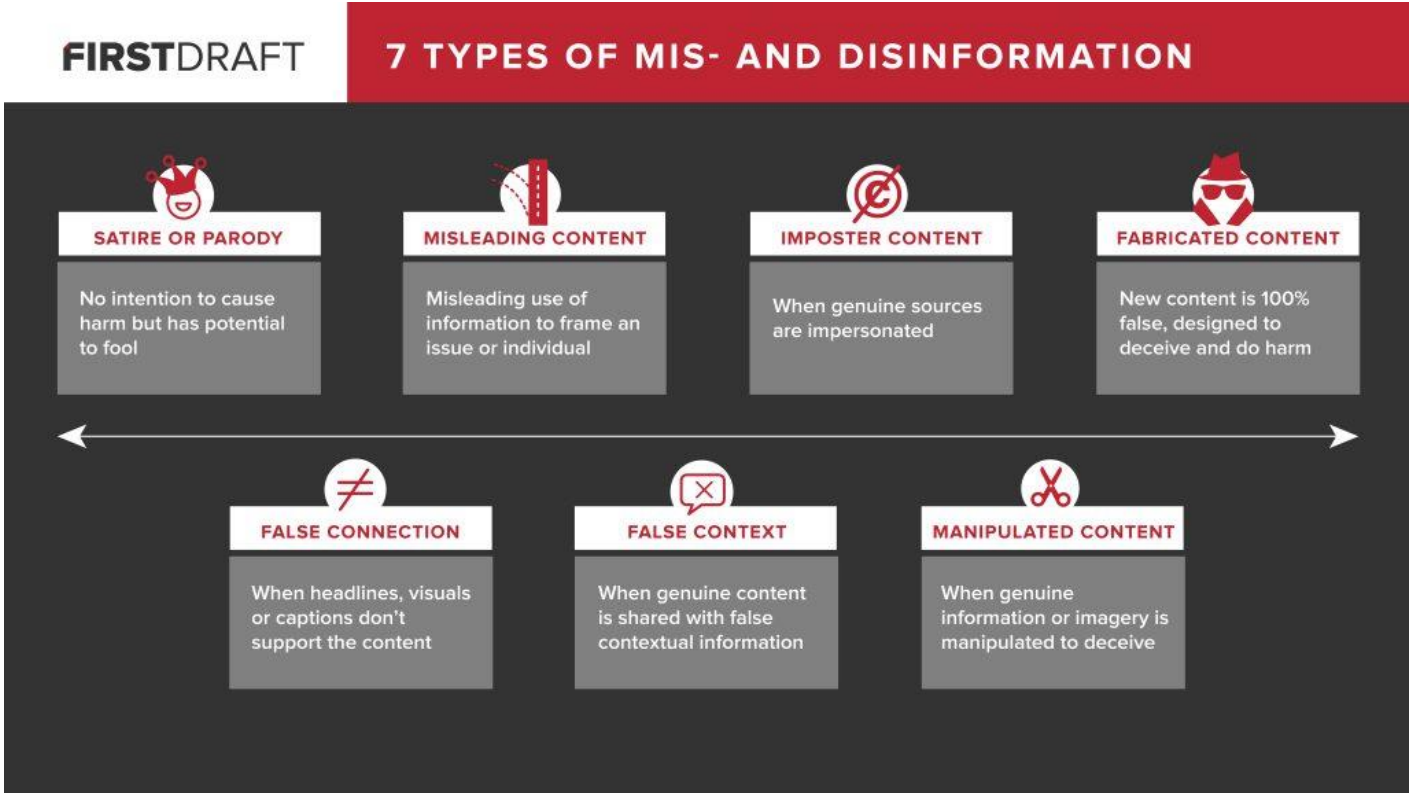
Meme analysis challenges and future

- Current SotA focuses on analysing only one type of memes (image macros) leaving a huge amount of online content unnoticed
- Detection of online memes is also an understudied problem in comparison to meme classification, although it is a prerequisite
- The existing solutions while seemingly well-performing, provide suboptimal results when evaluated in-the-wild
- Meme databases are small-sized and noisy putting an upper bound to the comprehension of online memes by machines, mainly relying on pre-acquired model knowledge
- Multimodal LLMs with their global understanding have the potential to increase the level of automatic meme content interpretation

mever

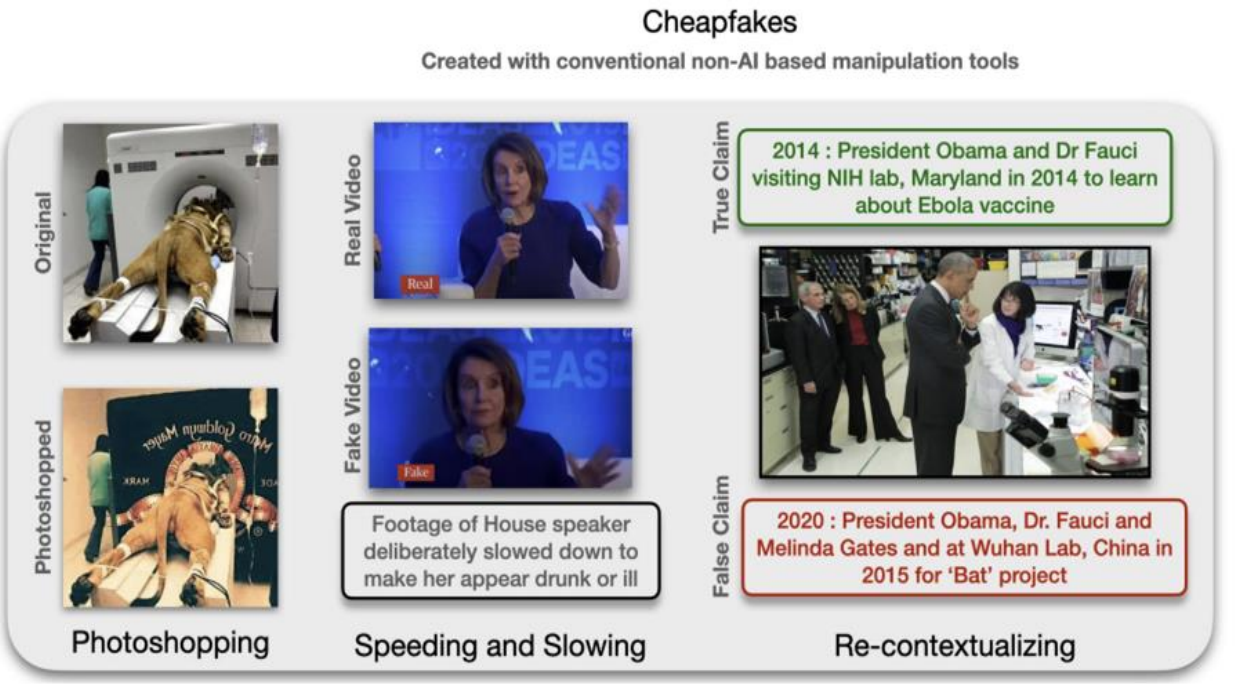
Multimodal misinformation detection

A broader issue: information disorder



https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/

Multimodal misinformation



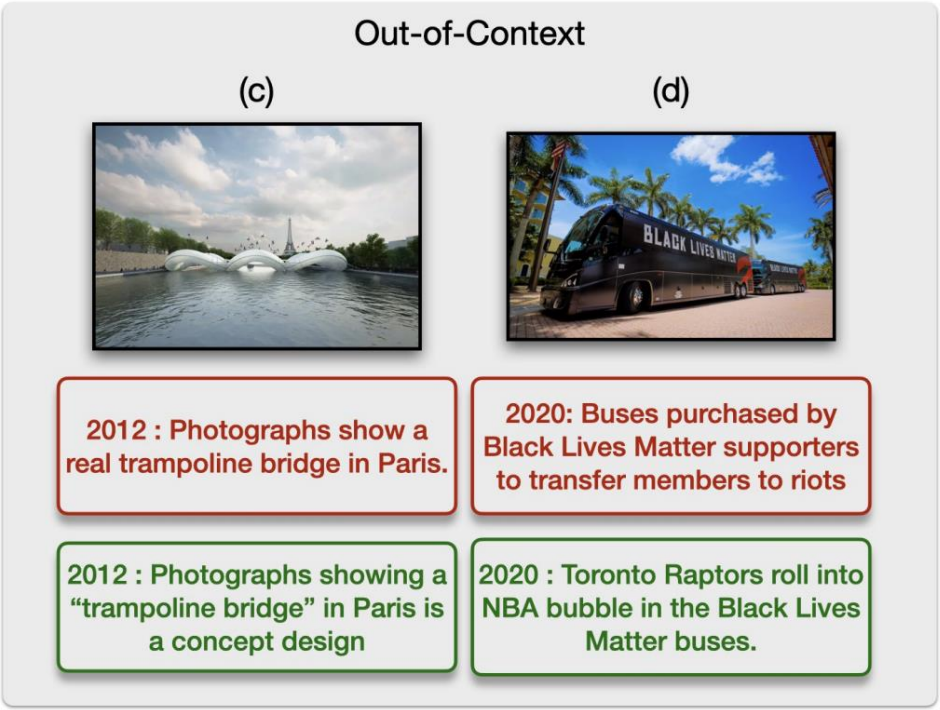
A low-cost but effective form of misinformation

Aneja, S., Midoglu, C., Dang-Nguyen, D. T., Khan, S. A., Riegler, M., Halvorsen, P., ... & Adsumilli, B. (2022). ACM multimedia grand challenge on detecting cheapfakes. *arXiv preprint arXiv:2207.14534*.

Out-of-context misinformation

“Repurposed images used to support misleading narratives.”

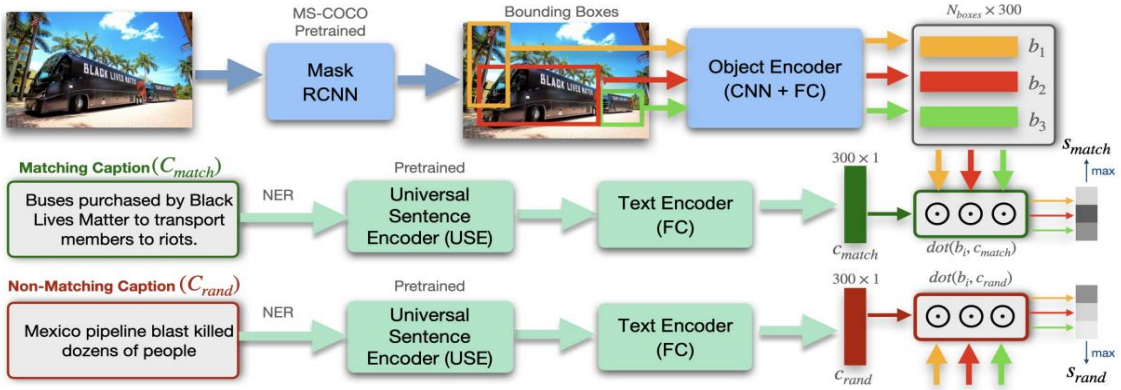
- Misleading narratives
- Amplification of bias
- Sensationalism and clickbait
- “CheapFakes”



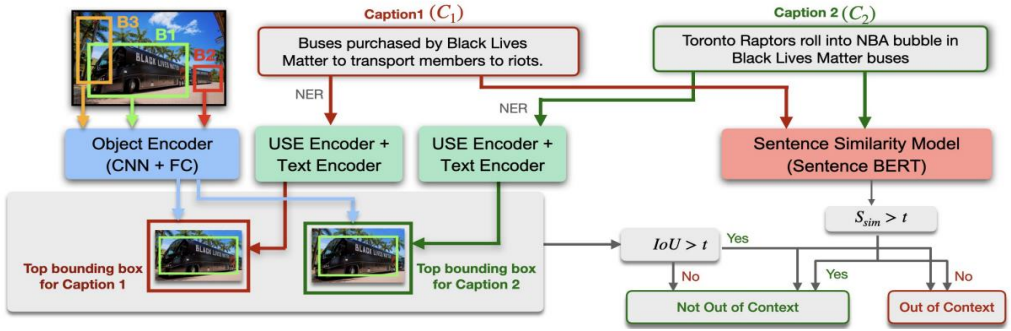
Aneja, S., Bregler, C., & Nießner, M. (2023, June). COSMOS: Catching Out-of-Context Image Misuse Using Self-Supervised Learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 14084-14092).

Out-of-context misinformation

Training



Inference



Aneja, S., Bregler, C., & Nießner, M. (2023, June). COSMOS: Catching Out-of-Context Image Misuse Using Self-Supervised Learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 14084-14092).

Out-of-context misinformation: NewsCLIPings

(1) Query Caption: Angela Merkel speaks to the German parliament.


Pristine


Semantics / CLIP Text-Image


Semantics / CLIP Text-Text


Person / SBERT-WK Text-Text

Scene / ResNet Place

(2) Query Caption: Fukushima Daiichi nuclear power plant after Japan s earthquake and tsunami in March.


Pristine


Semantics / CLIP Text-Image


Semantics / CLIP Text-Text

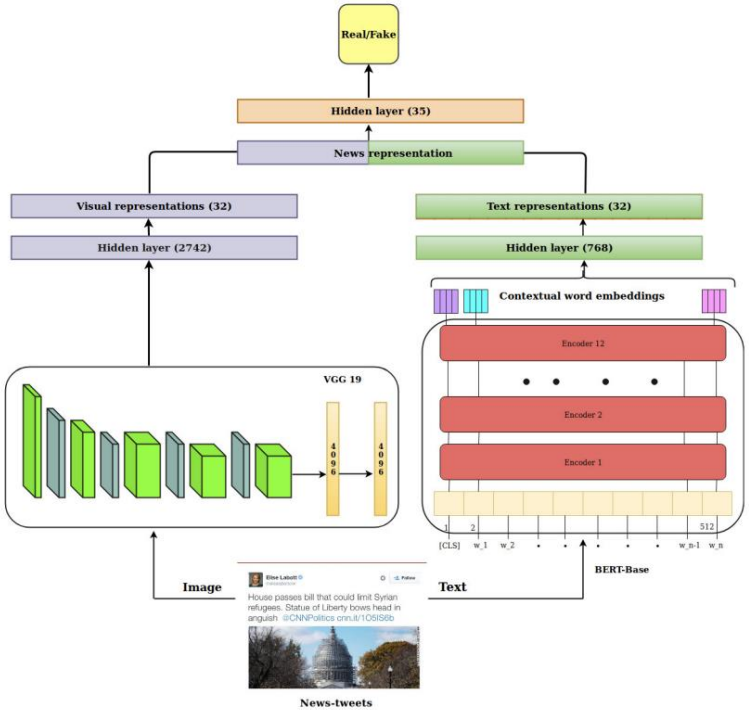
Person / SBERT-WK Text-Text


Scene / ResNet Place

CLIP + Person + Scene matching -> “hard negative samples” -> more effective detection models.

Luo, G., Darrell, T., & Rohrbach, A. (2021). NewsCLIPpings: Automatic generation of out-of-context multimodal media. arXiv preprint arXiv:2104.05893.

Multimodal misinformation: SpotFake



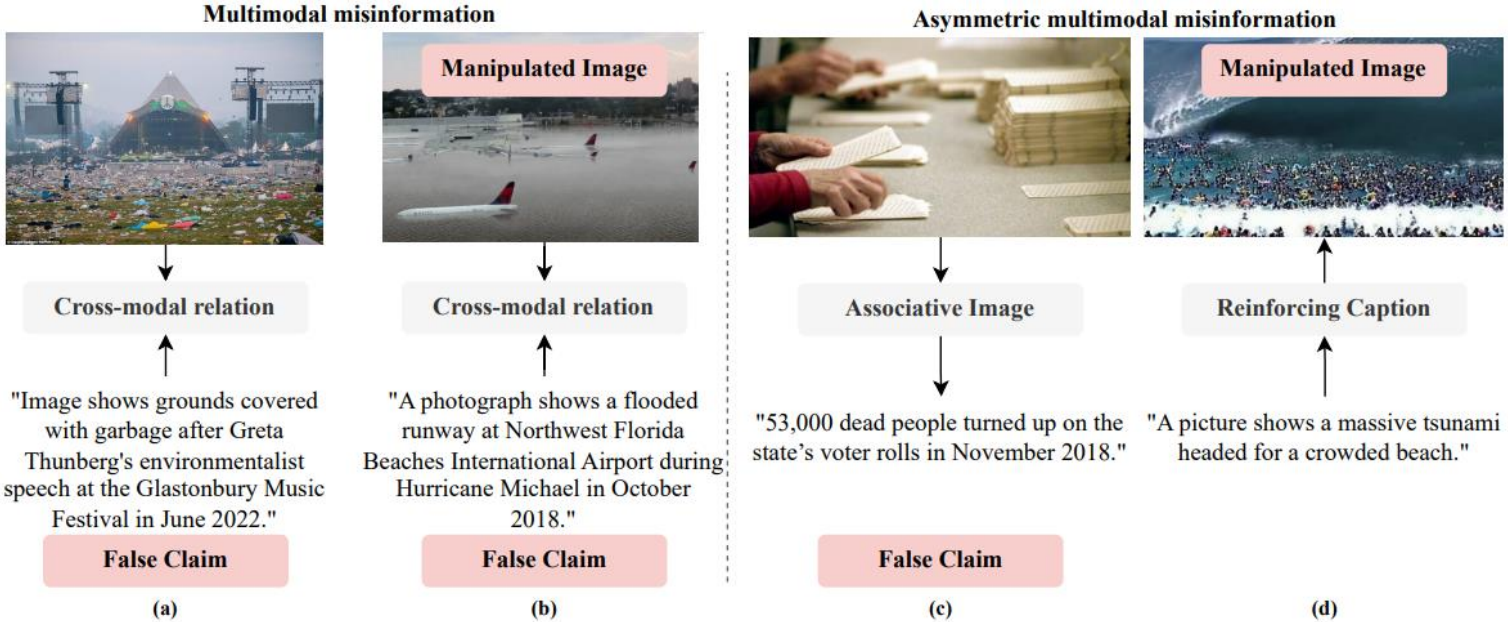
SpotFake:

- Visual encoder (VGG)
- Textual encoder (BERT)
- Classifier: Real or Fake

78% acc. on Twitter VMU dataset and 89% on Weibo dataset

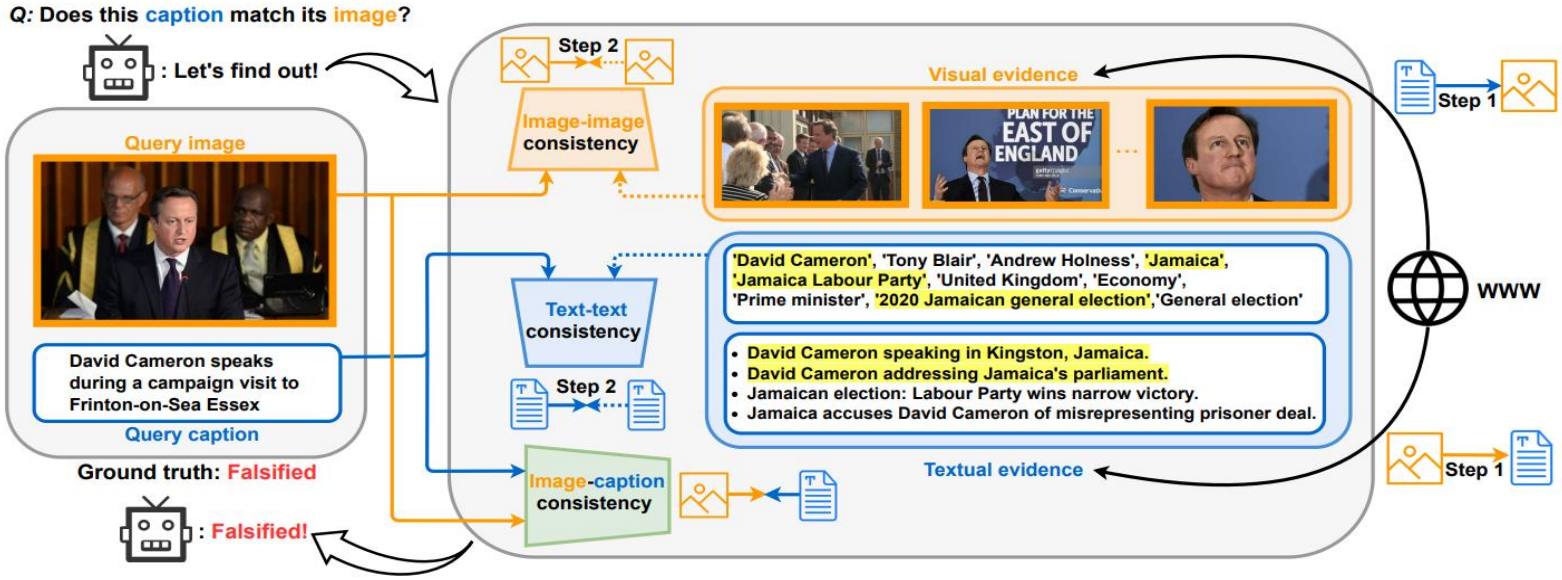
Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. I. (2019, September). Spotfake: A multi-modal framework for fake news detection. In 2019 IEEE fifth international conference on multimedia big data (BigMM) (pp. 39-47). IEEE.

Assessing multimodal misinformation detection



Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2023). VERITE: A Robust Benchmark for Multimodal Misinformation Detection Accounting for Unimodal Bias. arXiv preprint arXiv:2304.14133 (accepted in IJMIR).

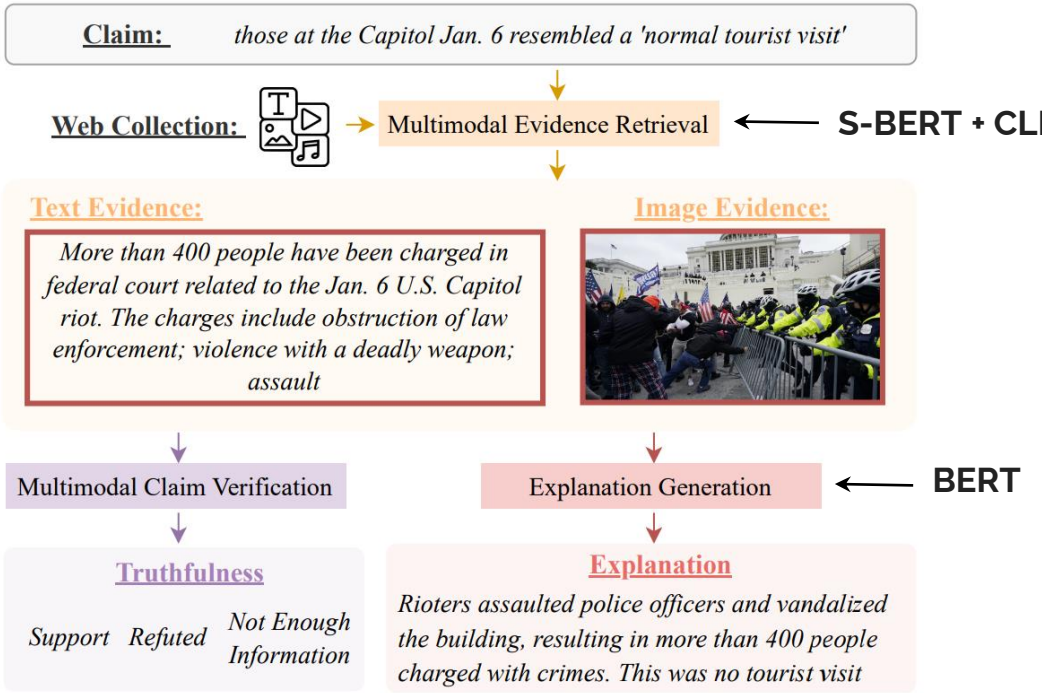
Multimodal fact-checking



Consistency-Checking Network (CCN): collects external information from the web and examines its consistency with the claim's image and text.

Abdelnabi, S., Hasan, R., & Fritz, M. (2022). Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14940-14949).

End-to-end fact-checking



"End-to-End Fact-checking":

1. Multimodal Evidence Retrieval
2. Claim Verification
3. Explanation Generation

Yao, B. M., Shah, A., Sun, L., Cho, J. H., & Huang, L. (2023, July). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In Proc. of 46th International ACM SIGIR Conference (pp. 2733-2743).

Multimodal misinformation: Major challenges

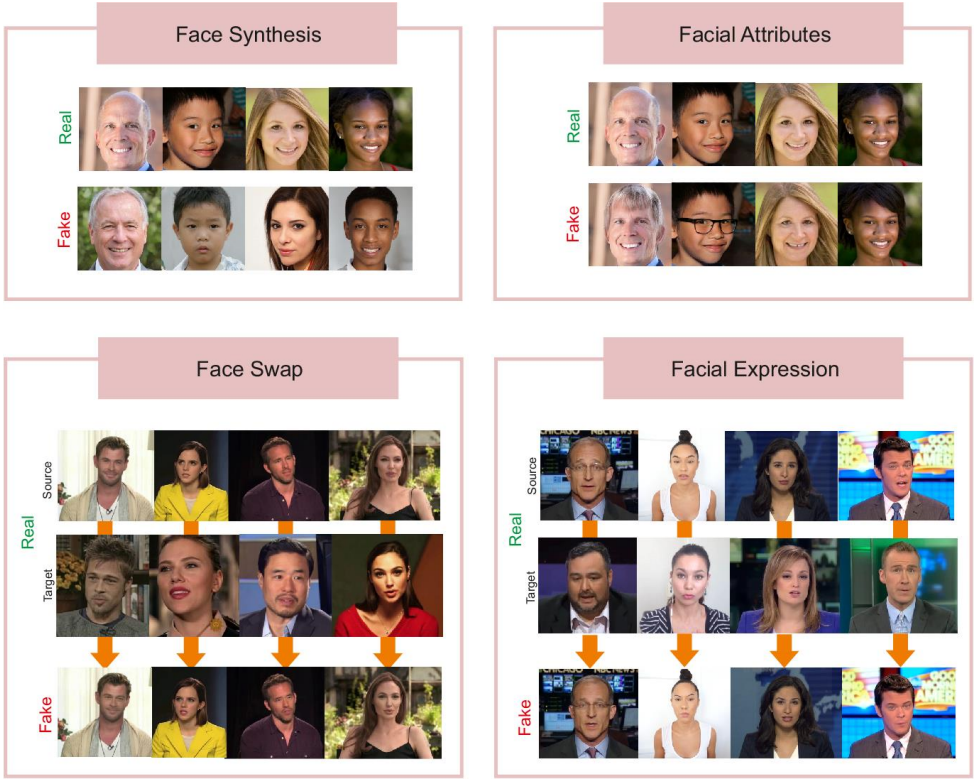
- Contextualization and meta information
- Transparent and accountable models
- Fine-grained detection
- Bias, region, and cultural awareness
- Disinformation on emerging and evolving topics

mever

Multimodal deepfake detection

Multimodal deepfake detection

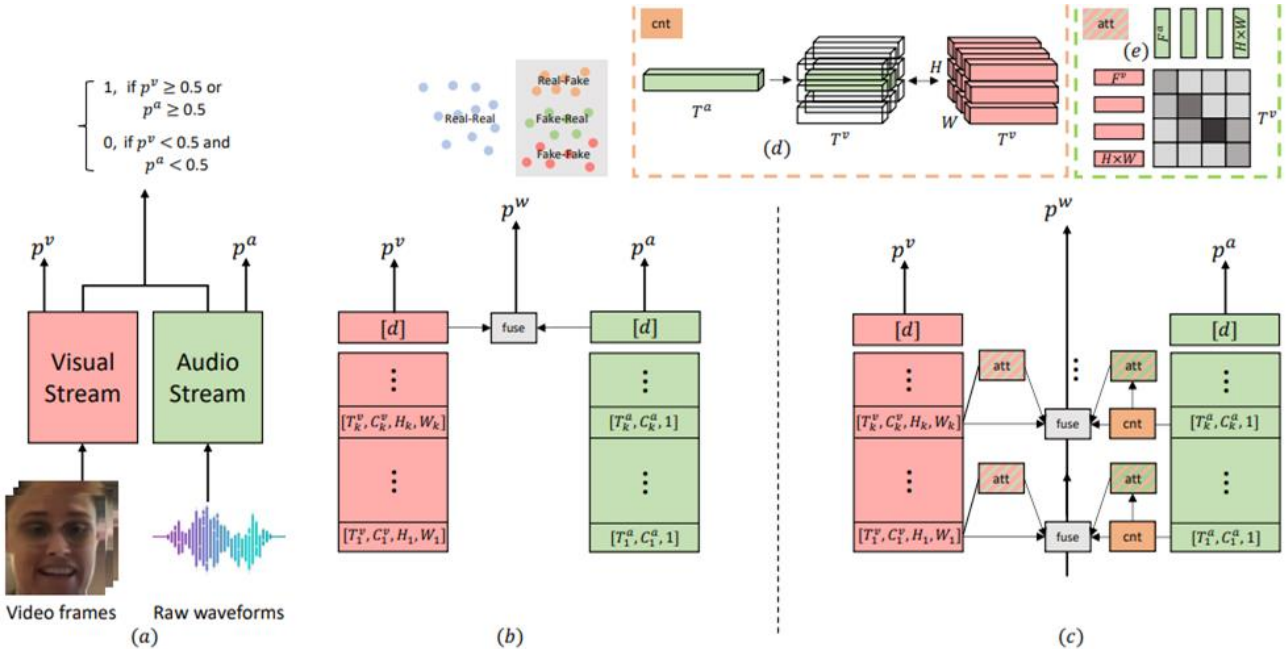
Deepfakes: synthetically generated fake images/videos created by means of deep learning-based architectures with the aim to be indistinguishable from real media to humans.



Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

Multimodal deepfake detection

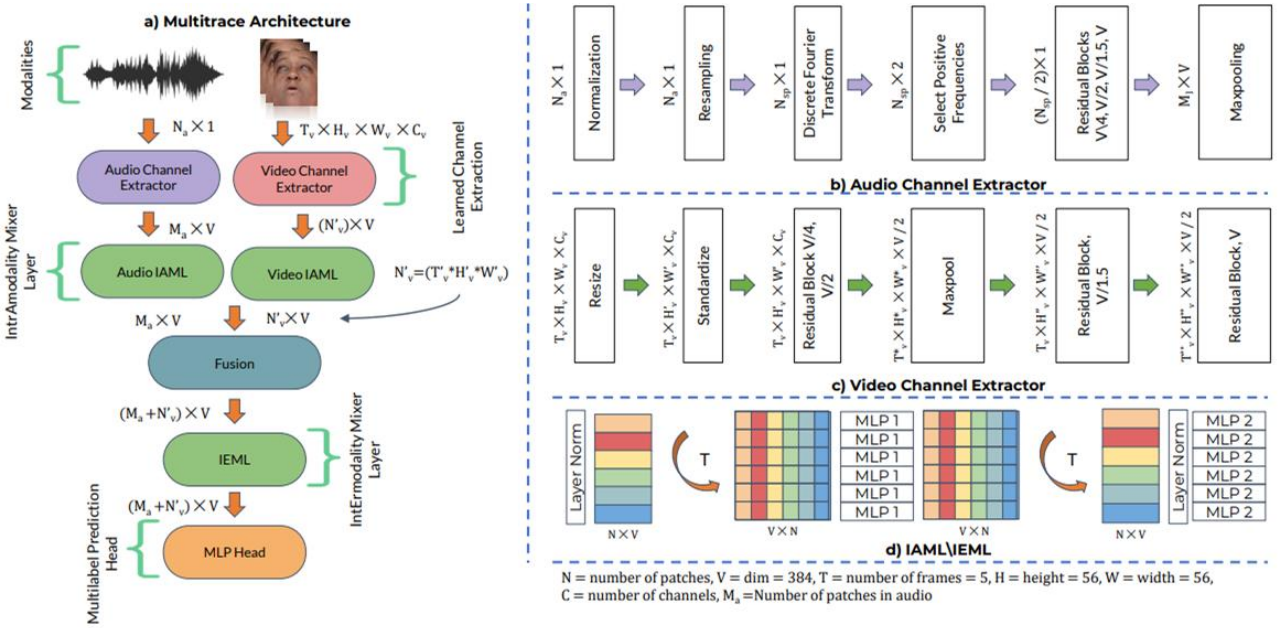
Spatio-temporal convolutional residual blocks & 1D CNN fusion with attention components



Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14800-14809).

Multimodal deepfake detection

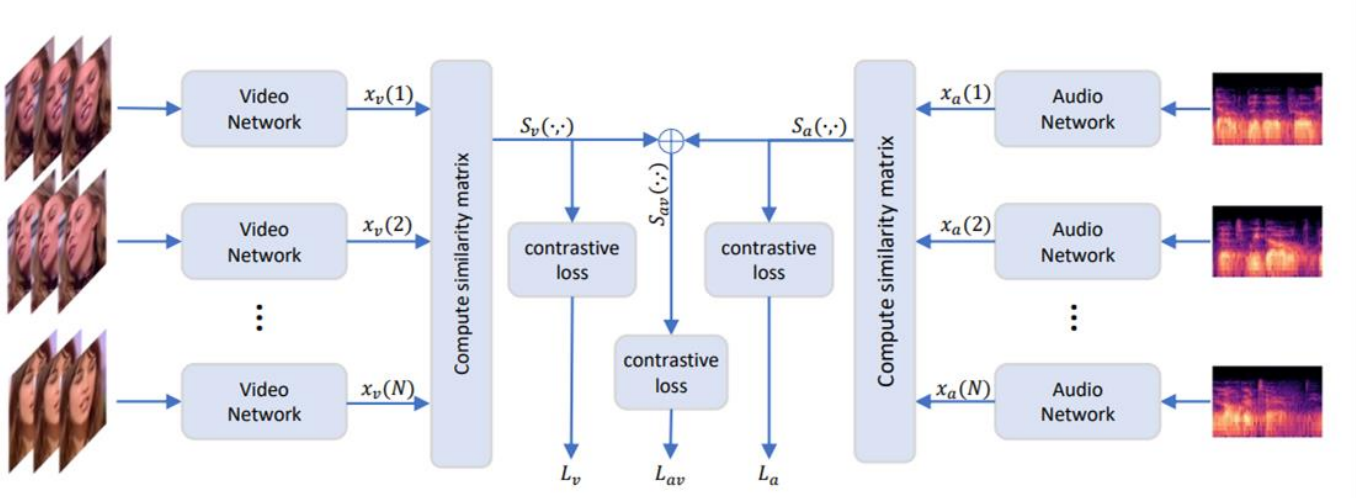
1D and 3D ResNets with separate MLP mixers for intra- and inter-modality information fusion



Raza, M. A., & Malik, K. M. (2023). Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 993-1000).

Multimodal deepfake detection

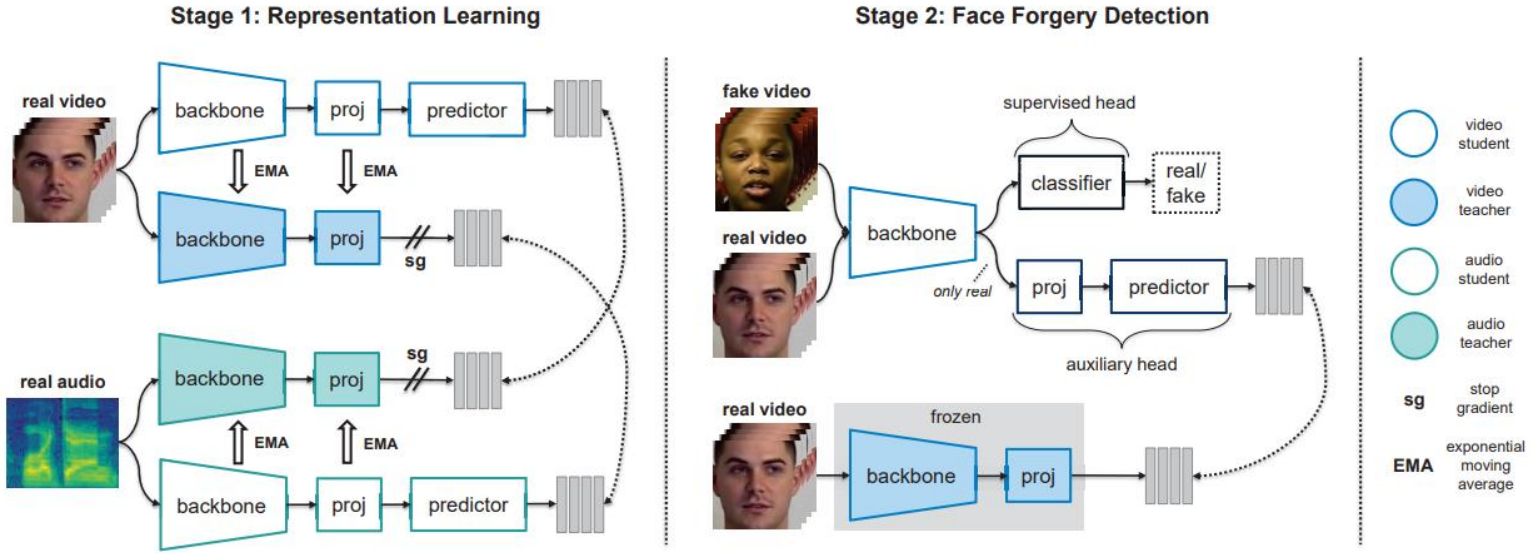
Person Of Interest (POI) Deepfake detection: Is the individual present in the video the person it claims to be?



Cozzolino, D., Pianese, A., Nießner, M., & Verdoliva, L. (2023). Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 943-952).

Multimodal deepfake detection

First: learn temporally dense video representations in a self-supervised way. Second: perform multi-task learning: forgery prediction & embedding reconstruction



Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14950-14962).

Multimodal deepfake detection: Challenges

- Low cross-dataset, cross-manipulation, in-the-wild generalization ability by current state of the art
- Robustness to content corruptions, e.g., compression, is limited
- Hard to explain outputs to journalists, black box predictions
- The arms race between the attackers and the defenders indicates the need for robust solutions

Takeaways

- Rapid developments in deep learning over the last years have given rise to models with strong Vision Language Understanding capabilities
- VLU capabilities match well with Journalistic needs
- Deploying multimodal models in task-specific settings can lead to good performance without too much additional effort

BUT

- still many design choices wrt learning multimodal representations
- hard challenges in terms of trustworthiness: output explanation, false positives, bias, etc.
- potential for misuse and adversarial attacks by malicious actors

Acknowledgements

Mr. George Karantaidis
Project Manager

Dr. Christos Koutlis
Post-doctoral Researcher

Mr. Stefanos Papadopoulos
PhD Candidate

mever

Thank you

Reach me at papadop @ iti.gr
Follow @sympap &
@meverteam