Advanced Deep Learning: Representations in Practice

Anders Søgaard

coastal

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
4	Phillip/Anders	May 16 (P): Representation and Adversarial Learning May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018) ; Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Architectures

Framework

Fairness /

Explainable AI

Methodology

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
		May 16 (P): Representation and Adversarial Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
4	Phillip/Anders	May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning		
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Today

a) Feature spaces
b) Shoehorning text into vector spaces
c) How to compare vector spaces
d) How to align vector spaces
e) Applications
    i) fMRI analysis
    ii) Cross-lingual learning
    iii) Multi-modal analysis
    iv) Bias detection

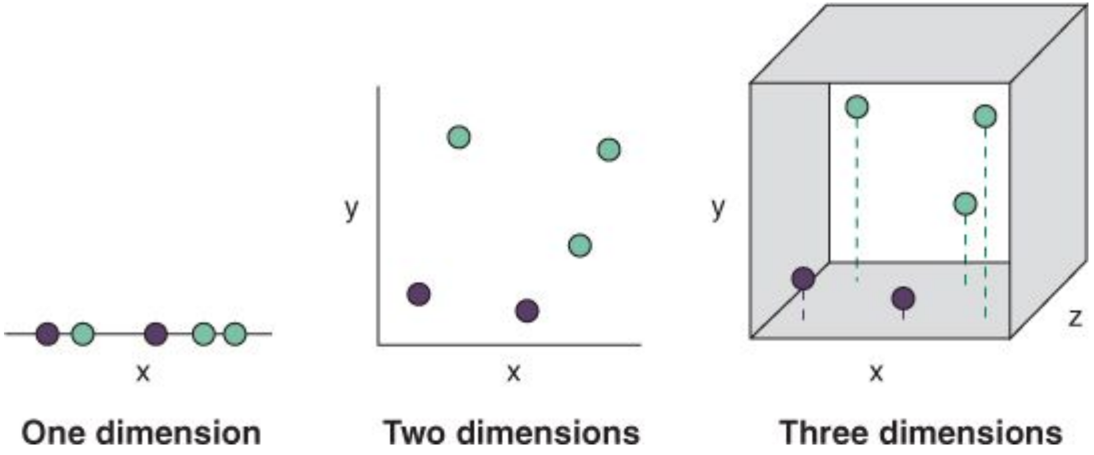
Feature spaces

Feature spaces

- Feature spaces are vector spaces.
- A vector space is anything that supports addition, subtraction and scalar multiplication of vectors and has a zero vector.
- Imagine a feature space with height, weight, and age. This is a three-dimensional vector space with vectors of the form <172, 58, 29>.

Dimensionality trade-off

In standard machine learning, dimensionality was determined by considerations of statistical support and separability. As we saw last time, scaling laws for DNNs seem to also reflect training dynamics.



Text vector spaces

Vector spaces

Old idea.

Salton, 1983

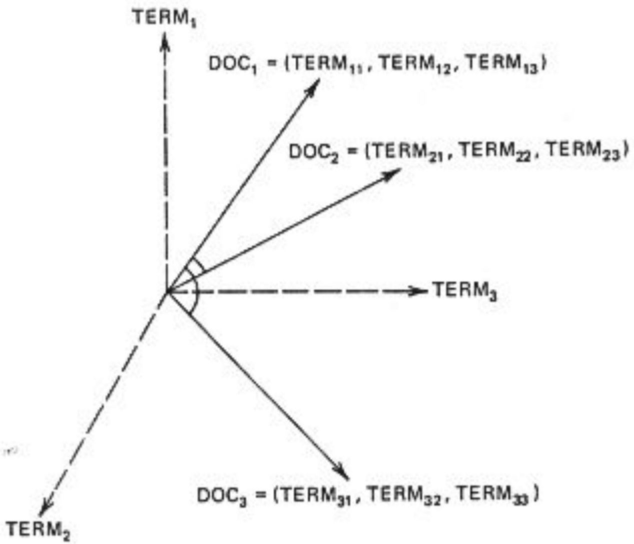
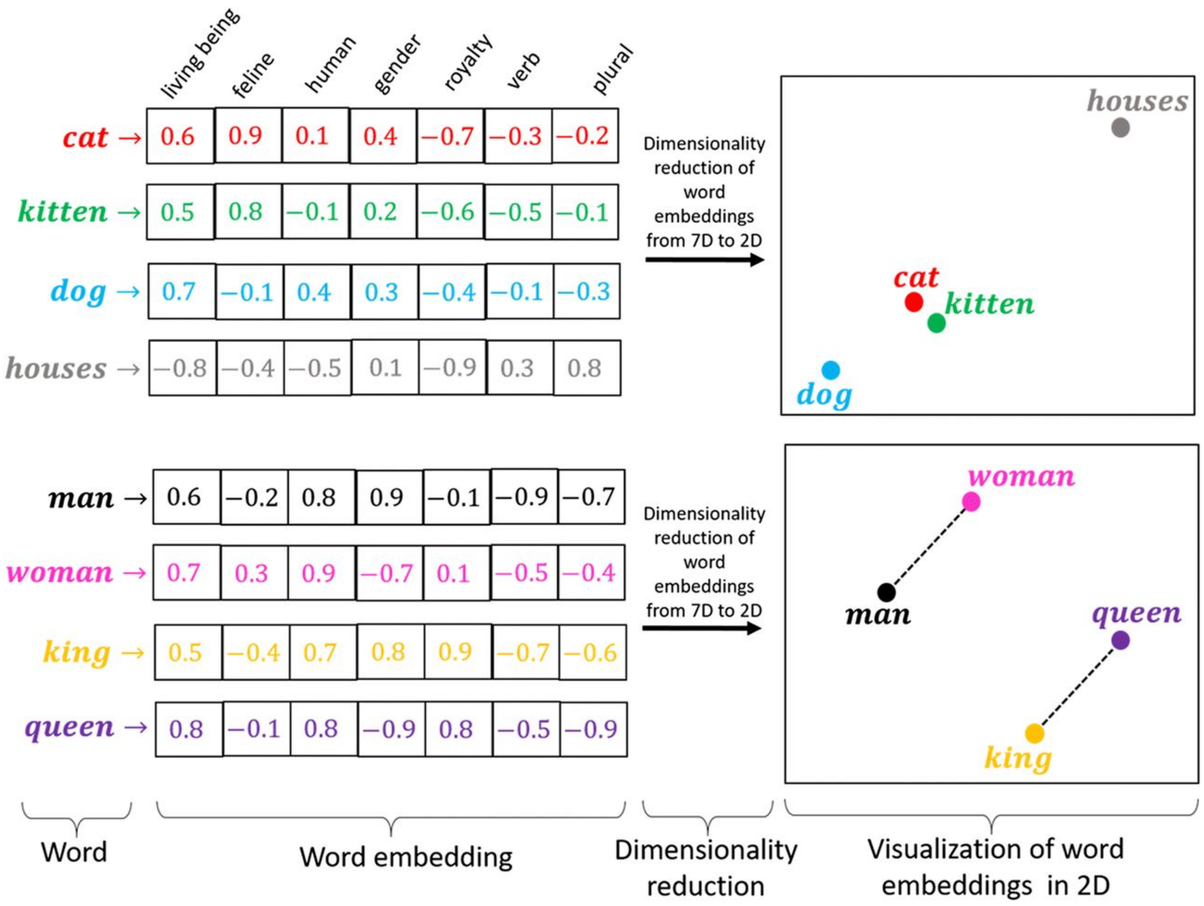


Figure 4-2 Vector representation of document space.

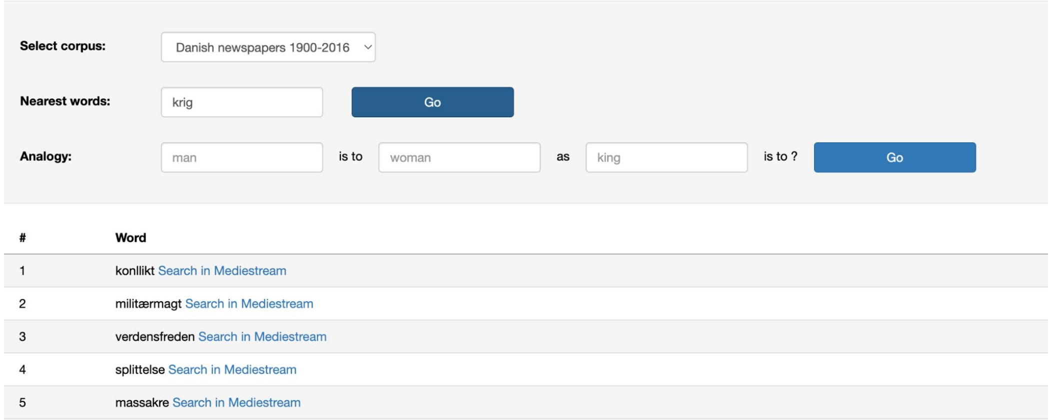
Word Embeddings: Idea

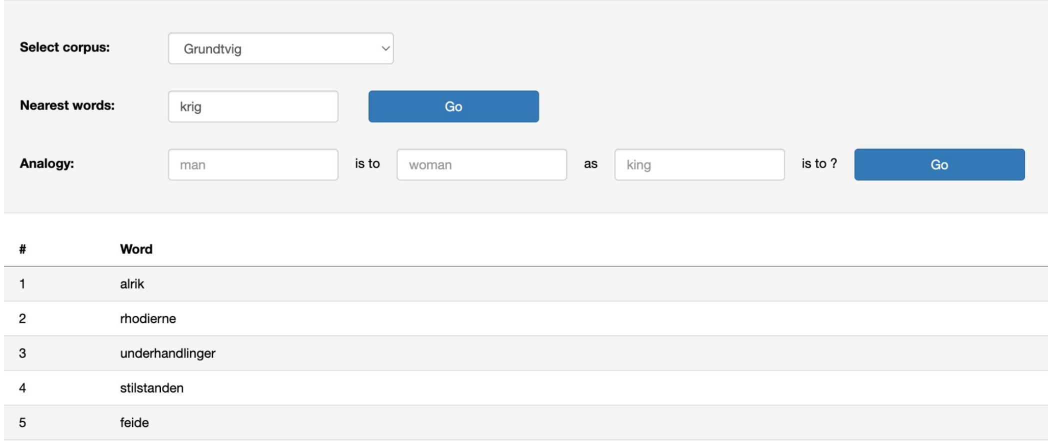
Vector representations of words that enable us to reason about word similarity.



Demo screenshots

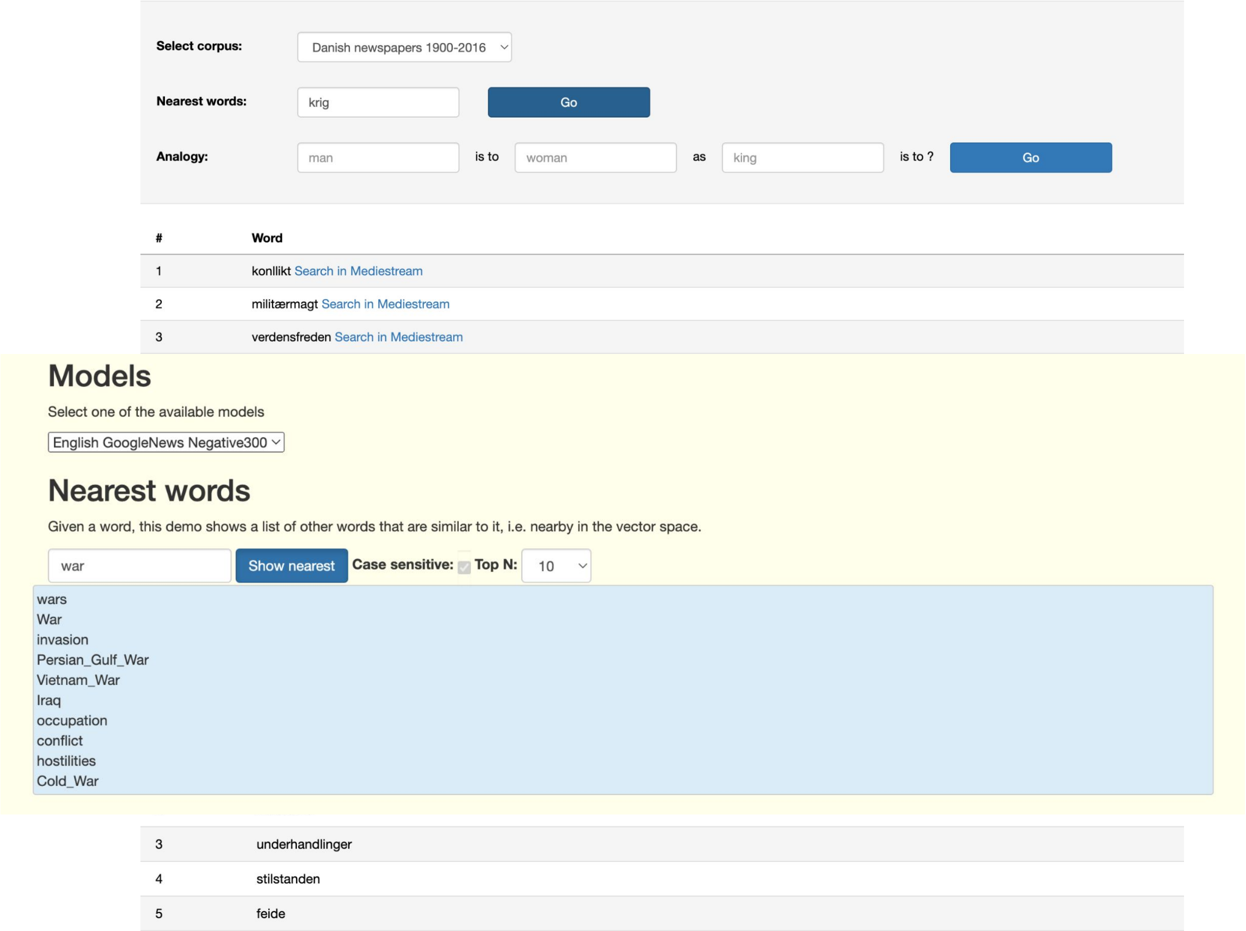
You can use word embeddings to explore associations in different corpora.





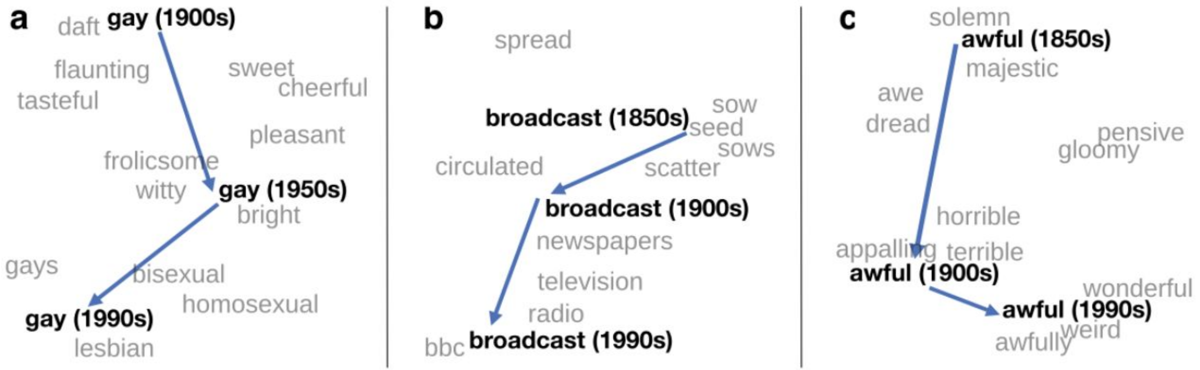
Demo screenshots

You can use word embeddings to explore associations in different corpora.



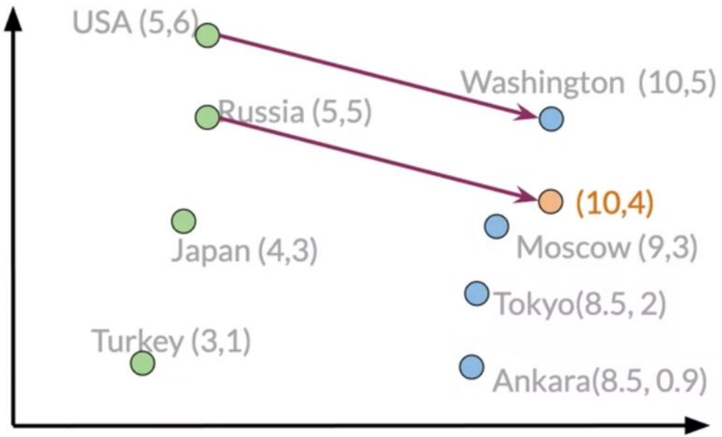
Temporal drift

We can also use word embeddings to monitor changes in word associations over time.



From analogies to isomorphism

If two spaces respect the same analogies, covering all words, they're isomorphic.

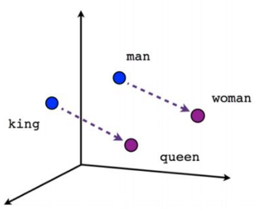


Washington - USA = $\begin{bmatrix} 5 & -1 \end{bmatrix}$

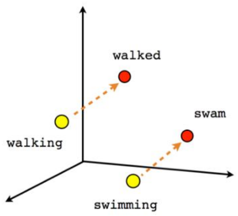
Russia + $\begin{bmatrix} 5 & -1 \end{bmatrix}$ = $\begin{bmatrix} 10 & 4 \end{bmatrix}$

How to evaluate embeddings?

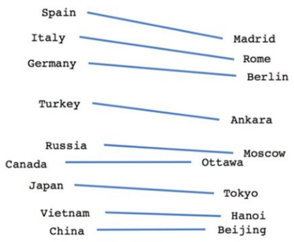
- Word associations
- Word analogies
- Alignment with lexical databases
- Downstream applications



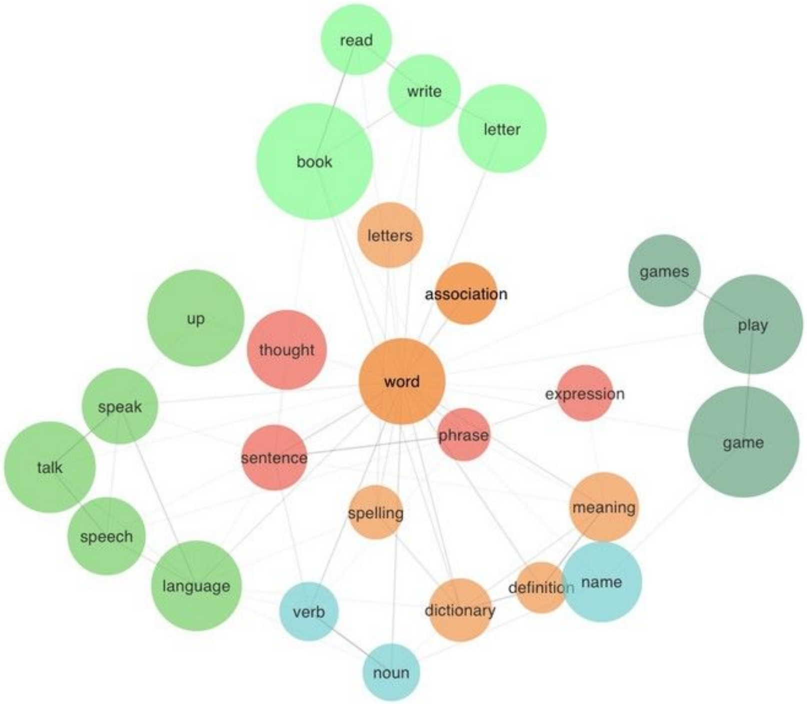
Male-Female



Verb tense

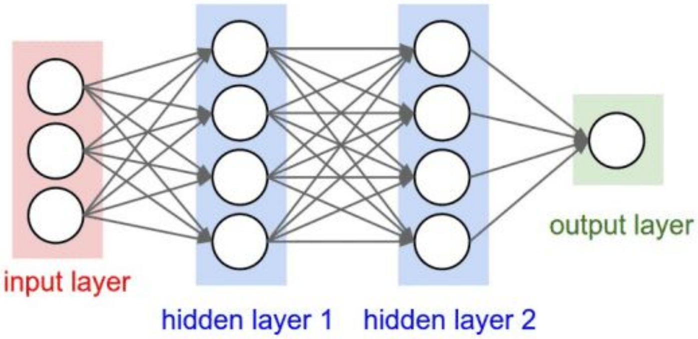


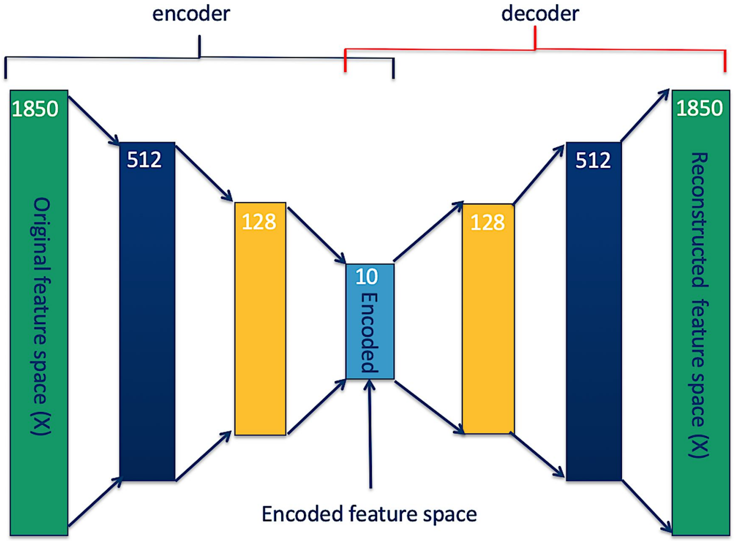
Country-Capital



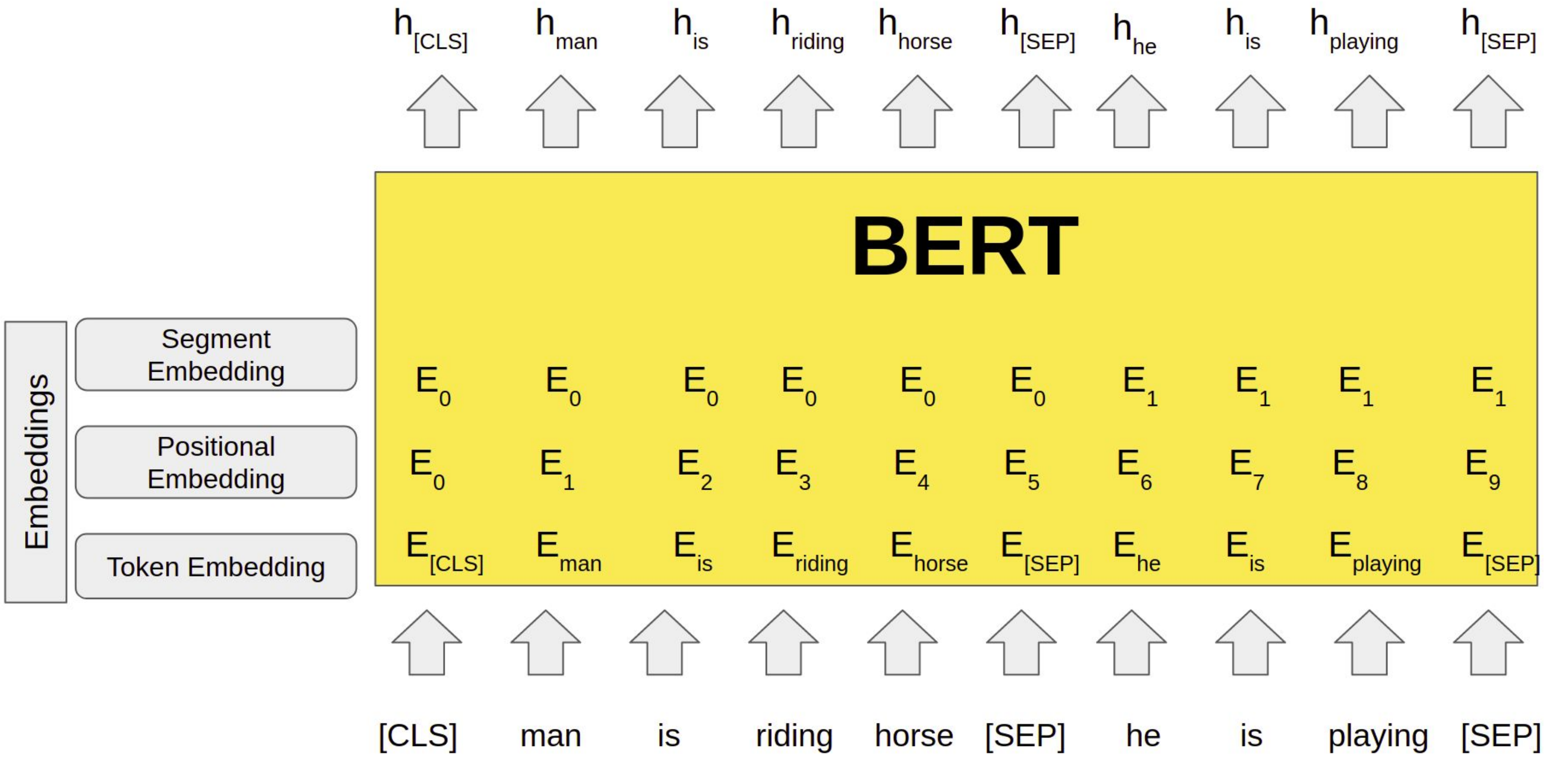
Deep networks as representation learners

Pretrained language models – such as BERT and GPT – are also text vectorizers. DNNs are generally representation learners and can be used as such (frozen); see e.g. Bert-as-a-service.





Challenges evaluating LM embeddings



Strategies

- Token embeddings
- Output token embeddings
- Sentence or output token embeddings in simple contexts, e.g., 'This is a ___'
- Averaging over simple contexts
- Averaging over randomly sampled contexts

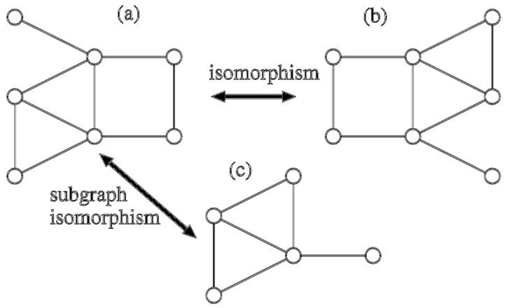
How to compare

Representational similarity

- Isometry

Challenge:

Vector spaces of different models or modalities are of course never *completely* isometric, but isomorphic?



Isomorphisms of Vector Spaces

Definition: An *isomorphism* (of V with W) is a bijective linear map $T : V \to W$.

Two vector spaces V and W are *isomorphic* if there exists an isomorphism $T : V \to W$. If this is the case, then we write $V \simeq W$.

Theorem 25: Let $T : V \to W$ be any map. Then T is bijective if and only if it has an inverse.

Theorem 26: Let $T : V \to W$ be a linear map and suppose that $\dim(V) = \dim(W) < \infty$. The following statements are equivalent.

(a) T is an isomorphism.

(b) T is invertible (i.e., T has an inverse).

(c) T is injective.

(c) T is surjective.

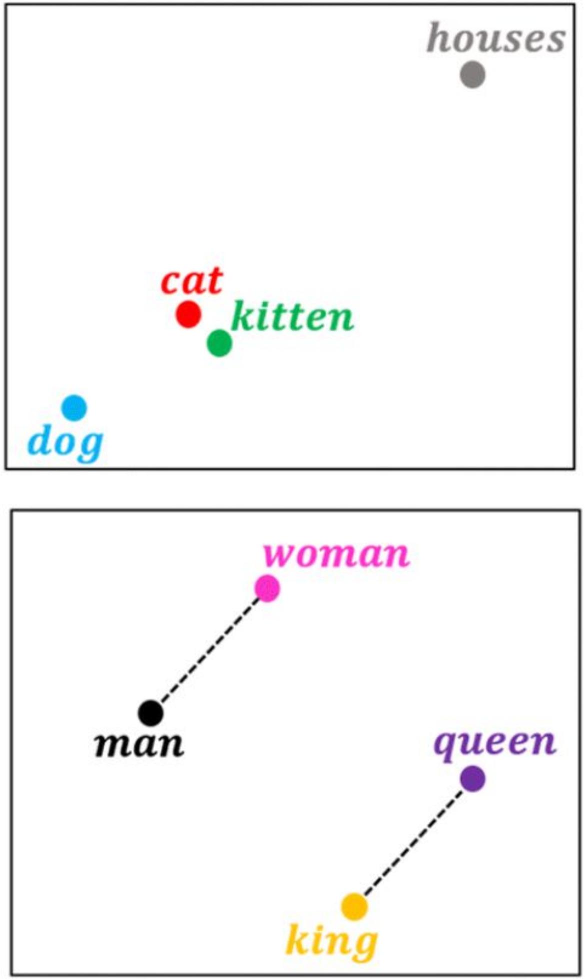
Remark: This theorem generalizes the Fundamental Theorem of Invertible Matrices. Indeed, if we take

$$V = W = \mathbb{R}^n \quad \text{and} \quad T = T_A,$$

where A is an $n \times n$ matrix, then we obtain the Fundamental Theorem.

Nearest neighbor graphs/analogy

- Two embedding spaces are nearest neighbor graph isomorphic if they satisfy the same nearest neighbor relationships.
- Two embedding spaces are graph isomorphic if they satisfy the same analogies.

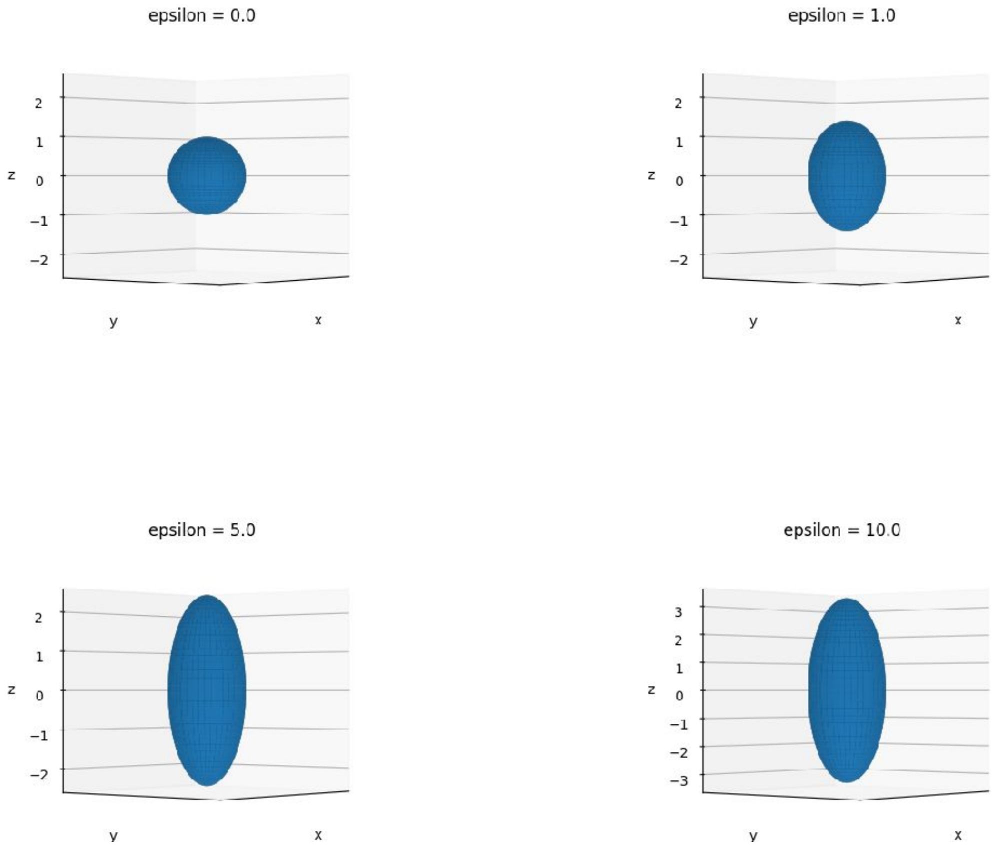


Epsilon isometry

Given a positive real number ε, an **ε-isometry** or **almost isometry** (also called a **Hausdorff approximation**) is a map f:X->Y such that

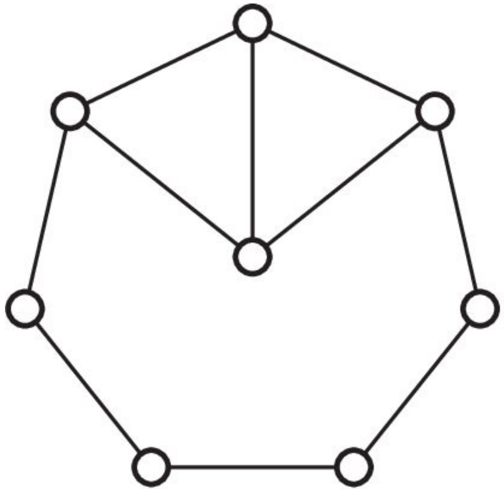
$$|d_Y(f(x), f(x')) - d_X(x, x')| < \varepsilon$$
$$d_Y(y, f(x)) < \varepsilon$$

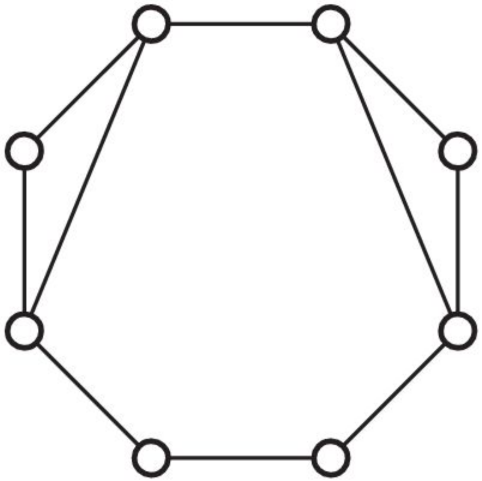
That is, an ε-isometry preserves distances to within ε and leaves no element of the co-domain further than ε away from the image of an element of the domain.



Isospectrality

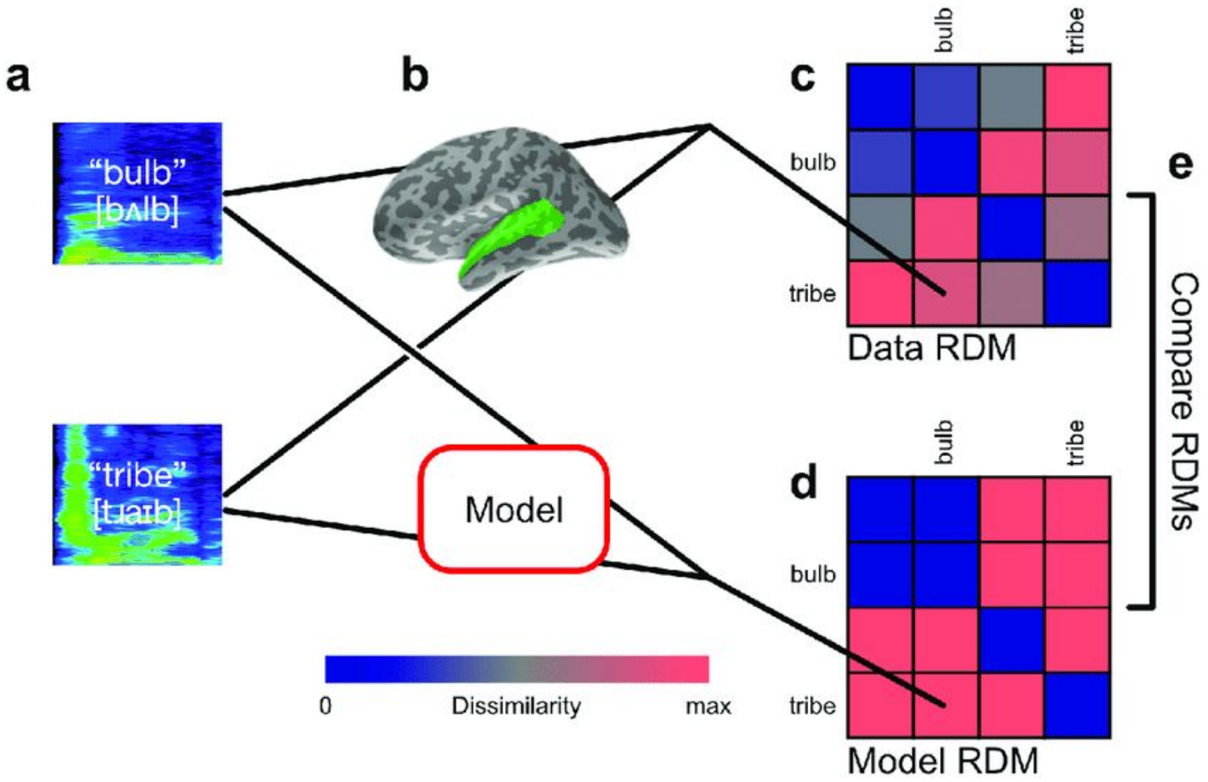
Two graphs are called cospectral or isospectral if the adjacency matrices of the graphs are isospectral, that is, **if the adjacency matrices have equal multisets of eigenvalues**.





RSA

a) Consider all pairs in a dataset, *b* and *t*.
b) Obtain representations in two models, e.g., **v**=*vector(b)* and **w**=*vector(t)*.
c) Construct the representational disimilarity matrix (RDM) for each model, in which *distance*(**v**,**w**) is a cell value.
d) Compute the correlation (e.g., Spearman's rho) across the cell values of the two RDMs.



How to align

Exercise

Do you have any techniques in your toolbox for aligning vector spaces?

Hint: You can think of vector spaces as datasets, matrices, graphs, or metric spaces, as you see fit.

Two ways

Supervised

- Multinomial regression
- Autoencoding
- Procrustes alignment
- CCA

Unsupervised

- Vanilla GANs
- Wasserstein GANs
- Other variations of GANs
- Point-set registration (ICP)

Supervised

Supervision

Supervision takes the form of a seed of paired vectors:

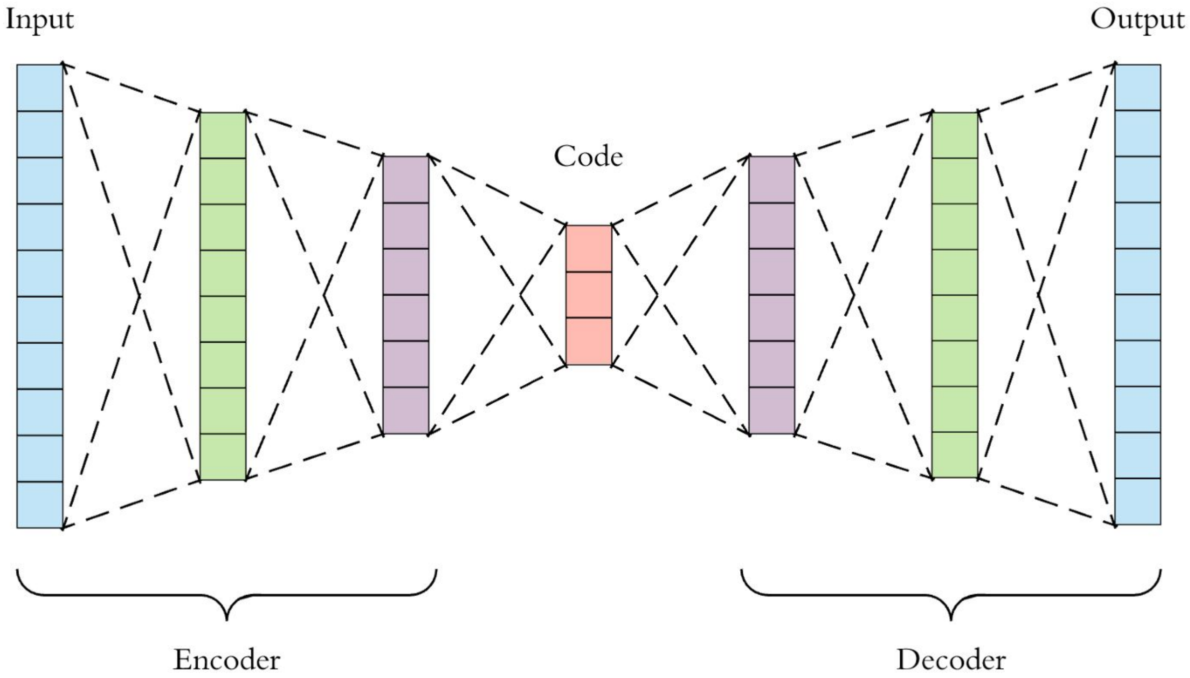
$$\mathbf{v}_S \sim \mathbf{v}_T$$

A simple regression model

For each vector $\mathbf{v_S}$ in vector space **S**, predict $\mathbf{v_T[0]}$ for the corresponding vector $\mathbf{v_T}$ in vector space **T**. For *d*-dimensional spaces, this amounts to training *d* (linear) regression models.

Autoencoder

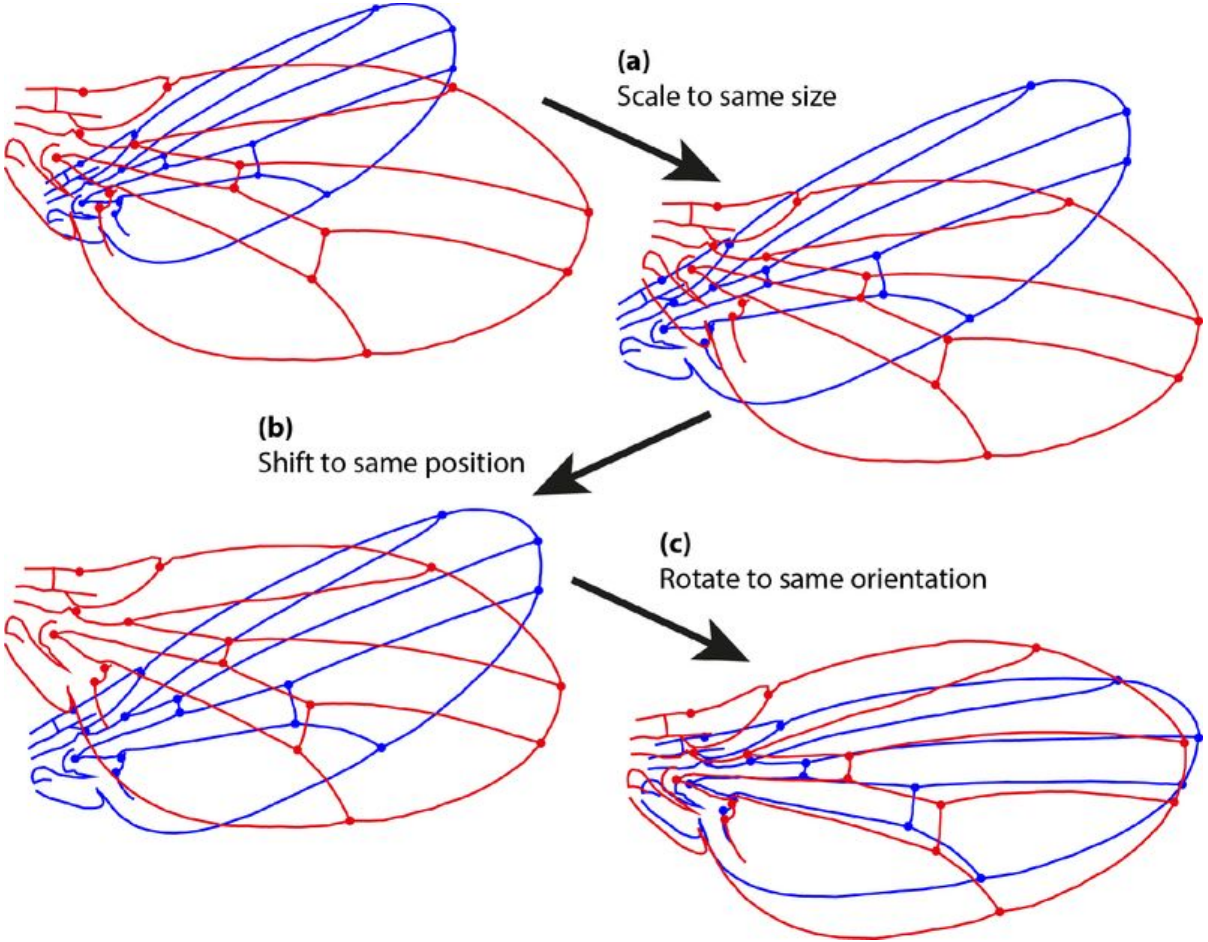
We can also use an autoencoder for the same purpose. For each seed vector pair, simply reconstruct one from the other. This learns a mapping from **S** into **T**.



Procrustes alignment

For two identically sized matrices $\mathbf{M_S}$ and $\mathbf{M_T}$, we apply a transform to the other to minimize the sum of distances between the representations of our seed pairs.

Methods differ depending on what you take the transformation matrices to be, e.g., orthogonal, rotation, symmetric or permutation matrices.

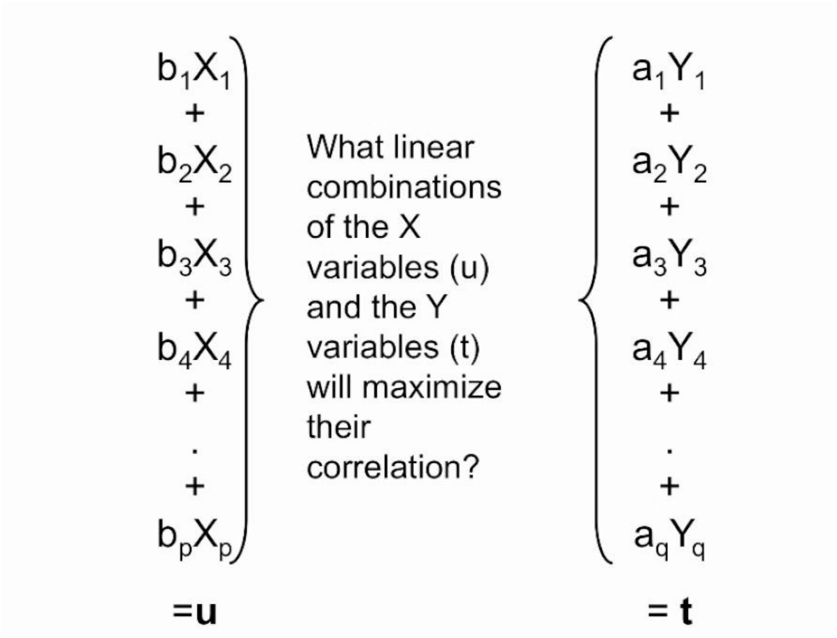


CCA

Estimate a and b to minimize

$$\rho = \operatorname{corr}(a^T X, b^T Y)$$

Note: CCA is symmetric. There's also a symmetric version in which you do two-sided Procrustes into the averaged space of **S** (*X*) and **T** (*Y*).



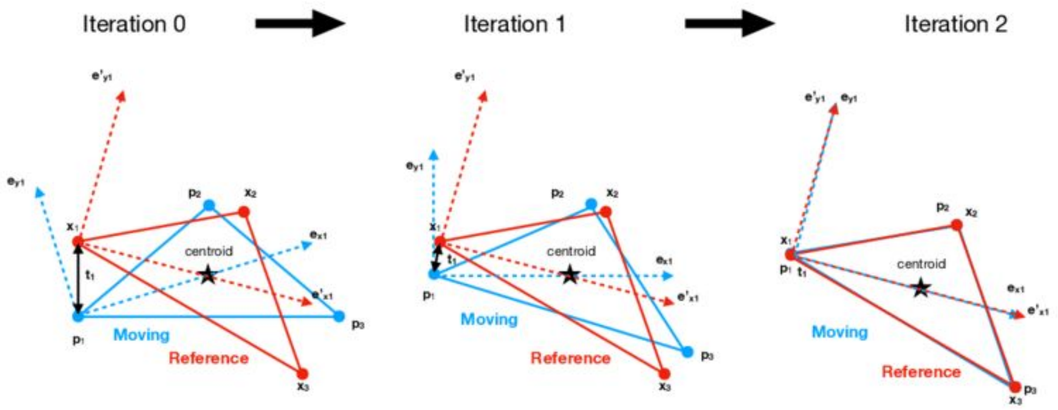
Unsupervised

Iterative Closest Point

For each point, match the closest point in the reference point cloud.

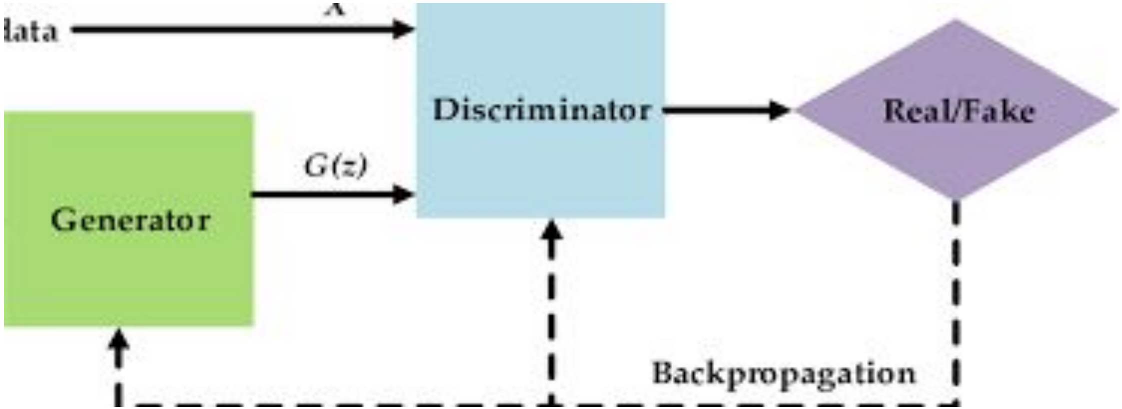
Estimate the rotation and translation matrix minimizing the pairwise distance.

Iterate (re-associate the points, etc).



Generative Adversarial Networks

Make vectors from **S** *fake*, vectors from **T** *real*. The Generator is now a linear projection.



Applications

fMRI analysis

The fMRI toolbox

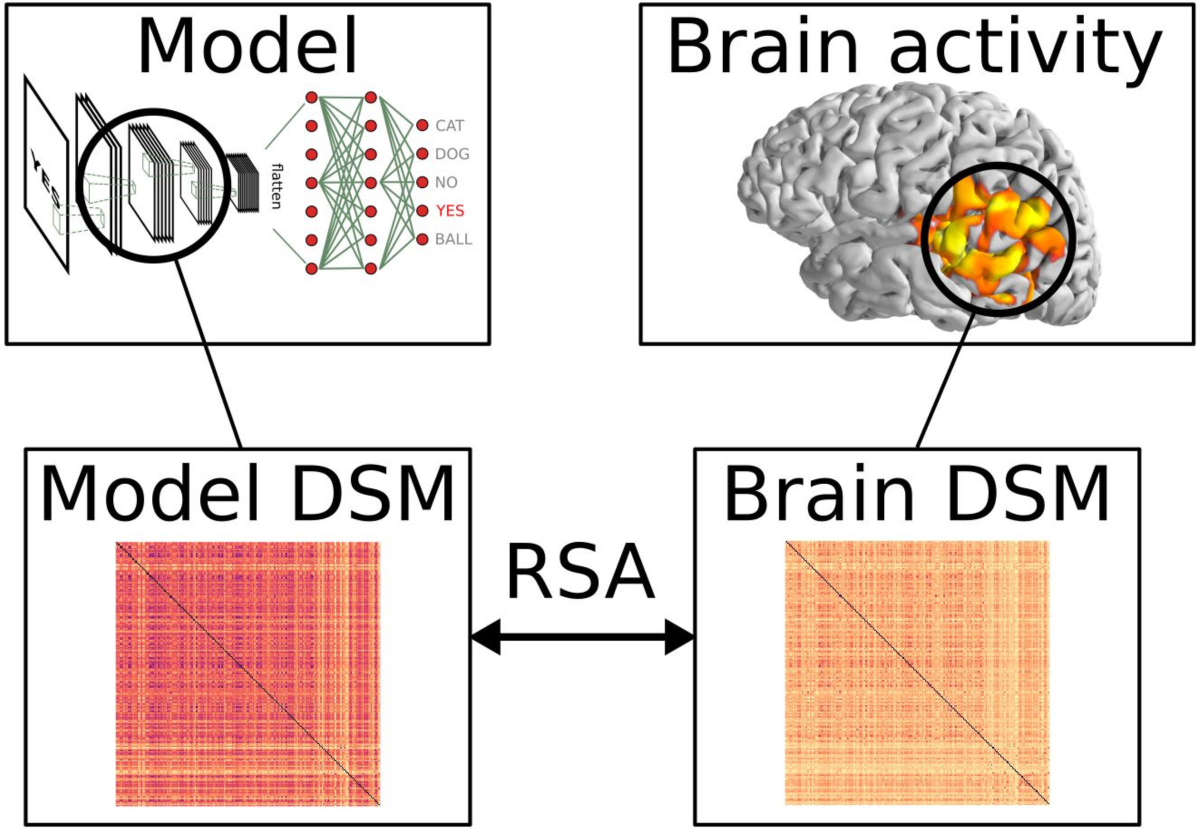
Isometry

- RSA to evaluate cognitive models

Alignment

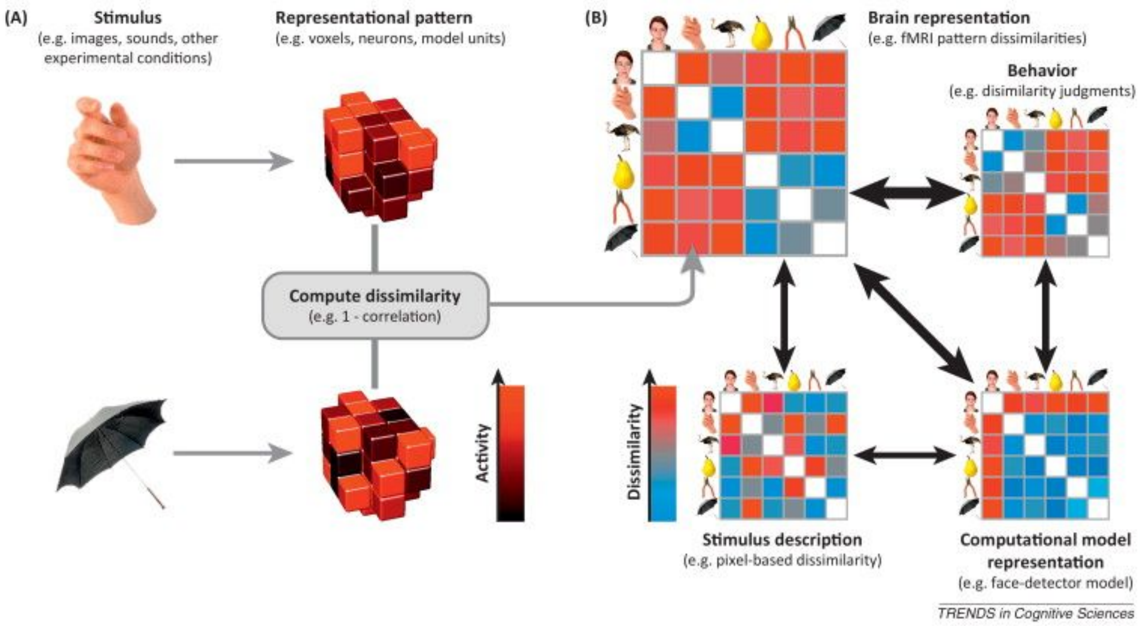
- Procrustes to align data from multiple subjects
- CCA to align data from multiple subjects

CV and NLP models are used to decode brain images, but to probe the potential for such decoding, we want to quantify the structural similarity of model and brain vector spaces.

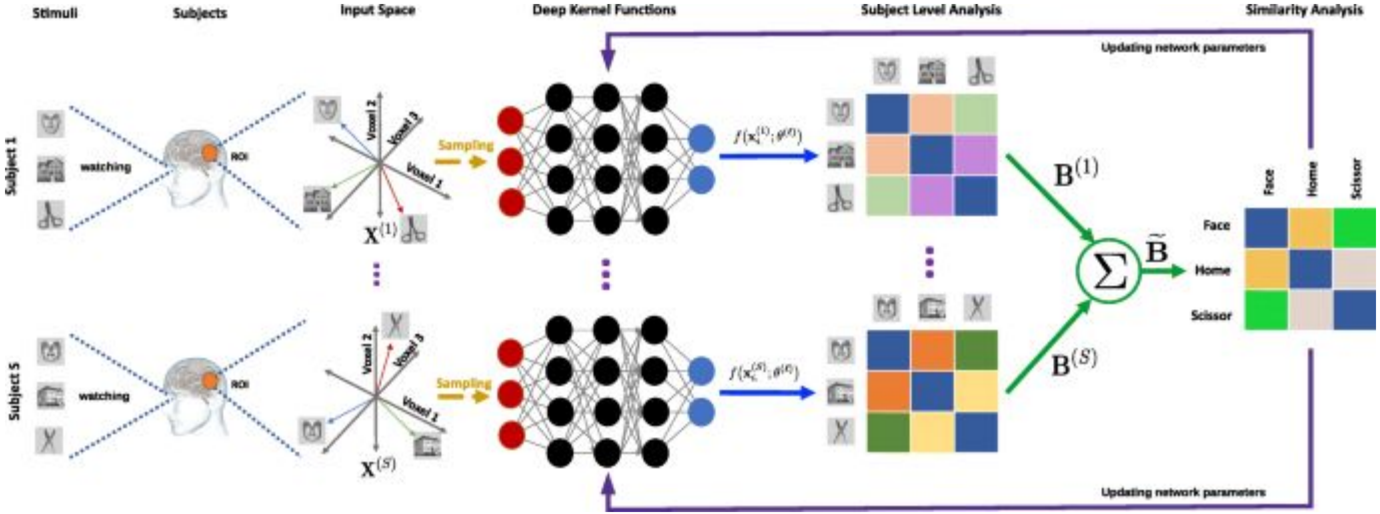


RSA workflow

a) Consider voxel representations of say a hand and an umbrella.
b) Compute *distance*(**v**=*vector(h)*, **w**=*vector(u)*.
c) Construct the representational disimilarity matrix (RDM) for each model, in which *distance*(**v**,**w**) is a cell value.
d) Construct your baseline RDM, e.g., images, model representations, other subjects.
e) Compute the correlation (e.g., Spearman's rho) across the cell values of the two RDMs.



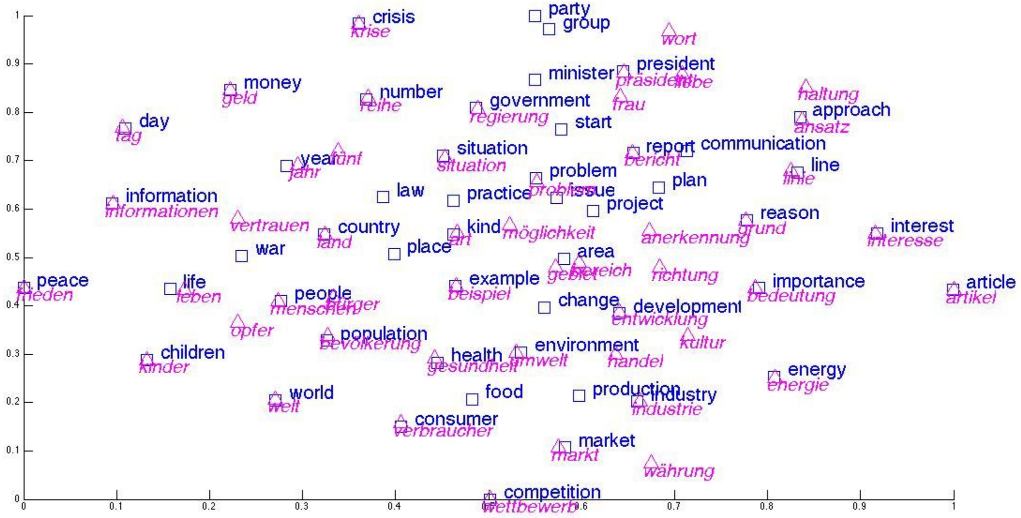
Deep RSA Learning



Cross-lingual

Objective

Evaluating the structural similarity of vector spaces across languages or explicitly aligning them.



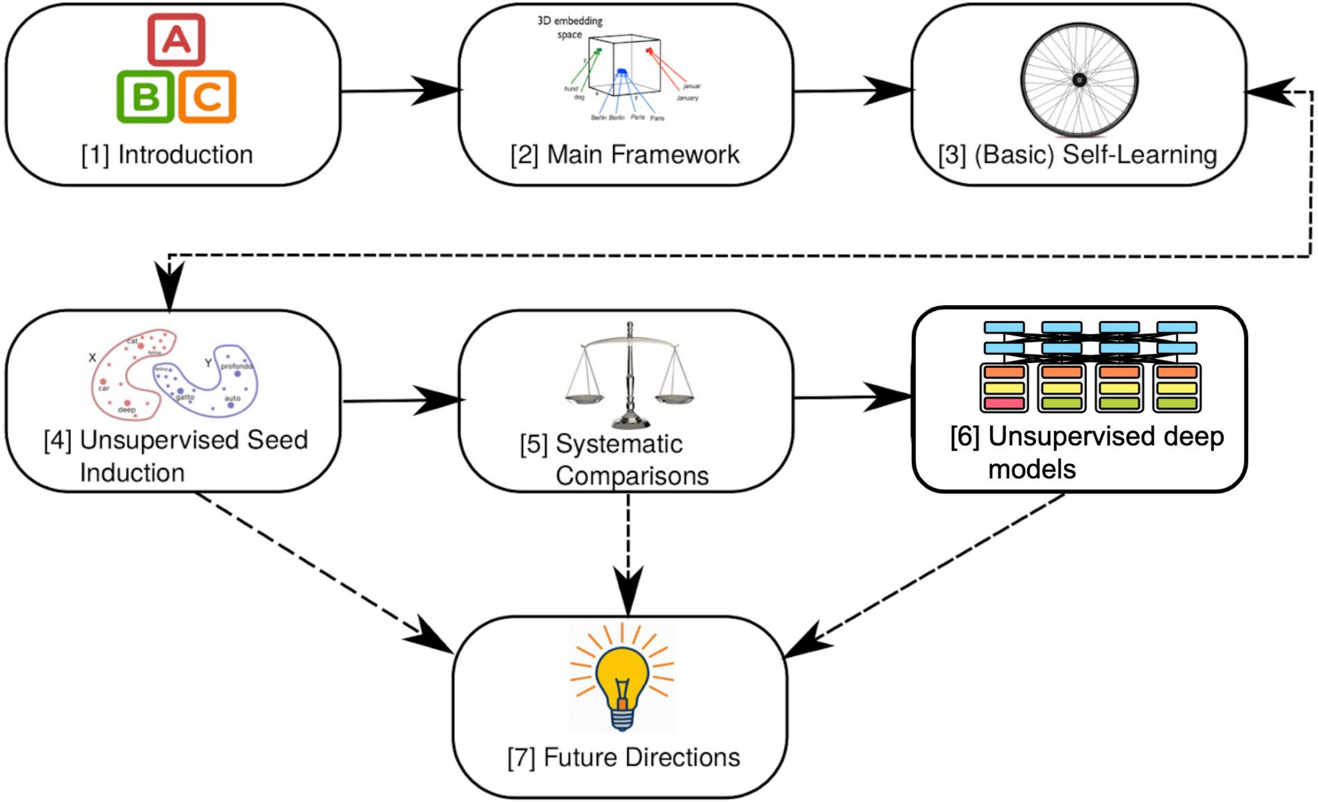
Two ways

Supervised

- Multinomial regression
- Autoencoding
- Procrustes alignment
- CCA

Unsupervised

- Vanilla GANs
- Wasserstein GANs
- Other variations of GANs
- Point-set registration (ICP)



See our tutorial here

Multi-modal

Aligning text and image

If we align ResNet and GPT-J, for example, we get a P@1=0.4 with modest levels of supervision.

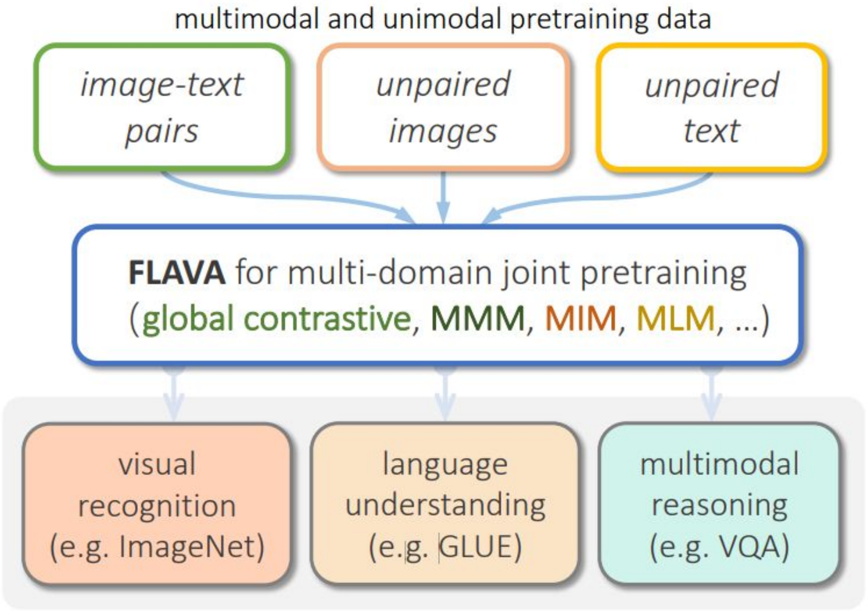
Applications

- Grounding
- Image captioning
- Visual question answering

How many slices of pizza are there?
Is this a vegetarian pizza?

FLAVA

FLAVA is a vision and language transformer by Facebook, presented at [CVPR 2022](). FLAVA is jointly pretrained for vision, text and their alignment.



FLAVA

It does really well on both image, text, and multimodal tasks.

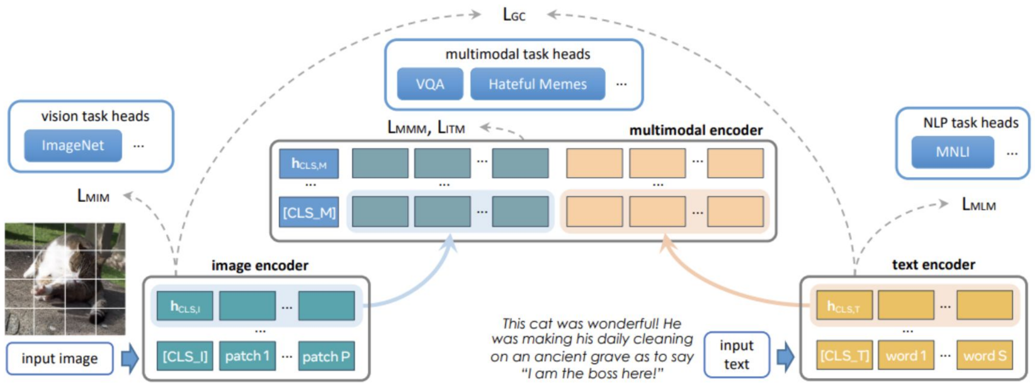
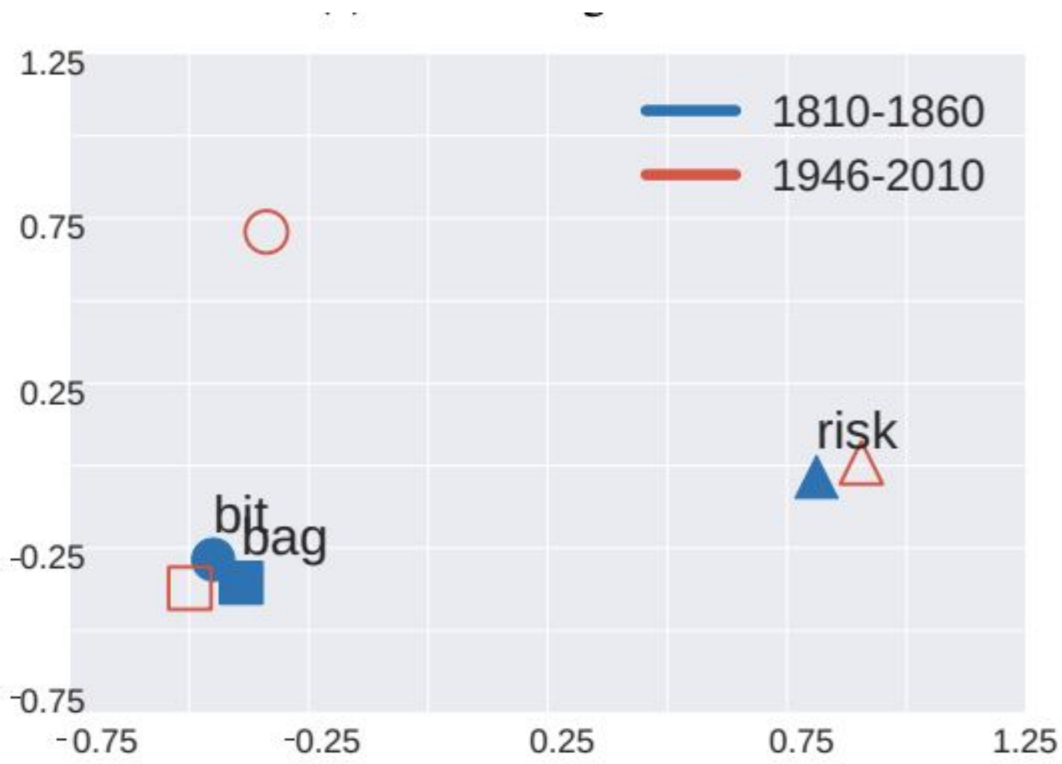


Figure 2. **An overview of our FLAVA model**, with an image encoder transformer to capture unimodal image representations, a text encoder transformer to process unimodal text information, and a multimodal encoder transformer that takes as input the encoded unimodal image and text and integrates their representations for multimodal reasoning. **During pretraining**, masked image modeling (MIM) and mask language modeling (MLM) losses are applied onto the image and text encoders over a single image or a text piece, respectively, while contrastive, masked multimodal modeling (MMM), and image-text matching (ITM) loss are used over paired image-text data. **For downstream tasks**, classification heads are applied on the outputs from the image, text, and multimodal encoders respectively for visual recognition, language understanding, and multimodal reasoning tasks.

Bias detection

Shift detection

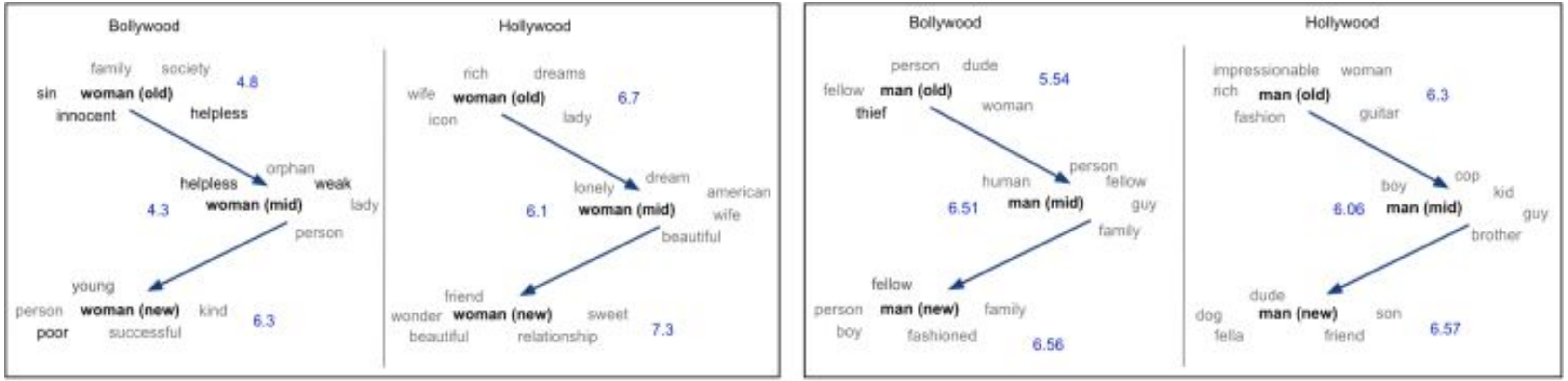
We can detect when some words have shifted meaning.



Exercise

How can you use what we have talked about today to detect bias?

Hint: In many ways.



A Woman over the years

B Man over the years

Shift detection in mentions of men and women in Bollywood/Hollywood