Large Language Model Applications

Embeddings

Daphne Ippolito and Fernando Diaz

January 27, 2026

logistics

Use of Language Models

Scribe notes

Using a language model to generate any part of a homework answer, scribe notes, and project reports will be considered a violation of academic integrity. To reiterate, all written words in your homework, scribe notes, and project reports should be written by yourself, without the use of AI (unless you are quoting AI outputs as part of your answer).

You may however use AI to learn more about the subject material of the class and to help you write code. If you do use AI systems to help with the homework, please fill out the "Use of AI" question on the homework to help the instructors understand your usage. Use of AI without making an honest attempt at answering this question will be considered a violation of academic integrity.

If you have submitted scribe notes and were confused by the AI use policy and would like to prepare new scribe notes, you have until January 30 to submit a new set of notes.

poll: an embedding is...

Embeddings

Local representations

- A neural network implements a parameterized function defined as a composition of transformations between vector spaces (or tensor spaces).
  - During inferences, we are repeatedly projecting vectors between spaces of different dimensionalities.
- Local representation: When developing a system, we often have elements that are represented in raw, externally defined coordinate systems that do not directly reflect their semantic or latent structure.

elements	representation
words	one-hot vector in $\|V\|$-dimensional space
documents	sequence of words
users	one-hot vector in $\|U\|$-dimensional space
sessions	sequence of turns
images	pixel values

Geoffrey E. Hinton. Distributed representations. Technical report CMU-CS-84-157, Carnegie Mellon University, 1984.

Embeddings

Distributed representations

- A distributed representation is a mapping from a local, symbolic, or discrete representation to a continuous internal representation that supports generalization and efficient computation.
- One-hot projection (words, user ids): A learned mapping from a sparse, high-dimensional indicator space to a dense, typically lower-dimensional continuous space.
- Sequence projection (strings, sessions): A function that maps an ordered sequence of discrete elements to a fixed-dimensional vector, typically via composition operators such as recurrence, attention, pooling, or convolution, while encoding order and context.
- Graph projection (social networks): A mapping from nodes (and their neighborhoods or roles) in a graph to vectors, where geometric proximity reflects structural similarity, connectivity, or shared context in the graph.
- Spatial projection (images): A mapping from structured spatial inputs to vectors that preserve local spatial correlations and hierarchical features, typically using convolutional or patch-based architectures.

Geoffrey E. Hinton. Distributed representations. Technical report CMU-CS-84-157, Carnegie Mellon University, 1984.

poll: embeddings can be used for...

Embeddings

Uses

- Generalization and transfer: embeddings act as reusable intermediate representations that enable transfer across languages, tasks, domains, and modalities, often serving as inputs to downstream models with minimal task-specific supervision.
- Retrieval: embedding externally-defined elements into a single, denser vector space, allows their indexing and retrieval. Retrieval backends often use dense vectors to represent documents, passages, sentences, images.
- Evaluation: embeddings capture semantic similarity not supported by lexical evaluation metrics. Soft metrics (e.g., Bertscore) use dense vectors to represent words.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on machine learning, 2008.
Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd acm international conference on information & knowledge management, 2013.
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: evaluating text generation with bert. In International conference on learning representations, 2020.

poll: embeddings were invented in...

Embeddings

History

- [Salton 1963] Vector Space Model. Documents embedded into |V|-dimensional space as tf.idf-weighted vectors. Sequence information not preserved.
- [Switzer 1964] Dimensionality Reduction. Using factor analysis to capture semantic similarity between words.
- [Hinton 1984] Distributed Representations. Conceptual foundation for learned, distributed meaning in neural networks.
- [Deerwester et al., 1990] Latent Semantic Indexing. Dense latent representations via matrix reconstruction.
- [Hofmann 1999] Probabilistic Latent Semantic Indexing. Dense linear representation via latent variable modeling.

Gerard Salton. Associative document retrieval techniques using bibliographic information. Journal of the ACM, 10(4):440--457, 1963.
Paul Switzer. Vector images in document retrieval. In Proceedings of the symposium on statistical association methods for mechanical documentation, 1964.
Geoffrey E. Hinton. Distributed representations. Technical report CMU-CS-84-157, Carnegie Mellon University, 1984.
Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391-407, 1990.
Thomas Hofmann. Probabilistic latent semantic indexing. In Sigir '99: proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval, 50--57, New York, NY, USA, 1999.

Embeddings

History

- [Bengio et al., 2003] Neural Language Models. Learned embeddings for generalization in NLP.
- [Collobert and Weston, 2008] Multitask representation learning. Word embeddings based on performance across multiple tasks.
- [Mikolov et al., 2013; Pennington et al., 2014] Word embeddings (word2vec, GloVe). Scalable neural and matrix-factorization-based methods for learning from co-occurrence statistics.

Up to this point, word embeddings were *static,* a representation did not change as a function of the words around it.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. J. Mach. Learn. Res., 3:1137--1155, March 2003.
Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on machine learning, ICML '08, 160–167, New York, NY, USA, 2008.
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Workshop proceedings of the 2013 international conference on learning representations, 2013.
Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: global vectors for word representation. In Empirical methods in natural language processing (emnlp), 1532–1543, 2014.

poll: static embeddings are bad because...

The problem with static embeddings

- Ambiguity
  - A word can mean multiple different things, based on the topical context.
  - A word can mean multiple different things, based on its grammatical role.
- Intuition for a solution
  - Knowing the context of a word (topical, grammatical), can help disambiguate its meaning.
  - “One sense per discourse”: An ambiguous word will have a single sense within a document (or dialogue) [Gale et al., 1992].

We conclude that with probability about 94% (51/54), two polysemous nouns drawn from the same article will have the same sense. In fact, the experiment tested a particularly difficult case, since it did not include any unambiguous words. If we assume a mixture of 60% unambiguous words and 40% polysemous words, then the probability moves from 94% to $100\% \times .60 + 94\% \times .40 \approx 98\%$. In other words, there is a very strong tendency (98%) for multiple uses of a word to share the same sense in well-written coherent discourse.

William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In Proceedings of the workshop on speech and natural language, HLT '91, 233--237, Stroudsburg, PA, USA, 1992. , Association for Computational Linguistics.

Topic-specific word embeddings

- "one sense per discourse" suggests that knowing the topic can inform a word's representation.
- given a string x, how do we determine the right topic?
- information retrieval provides a way to, given x, retrieve topically-relevant documents.
- can use these documents to train word embeddings using static methods.
  - although the embedding is static, it changes based on x.

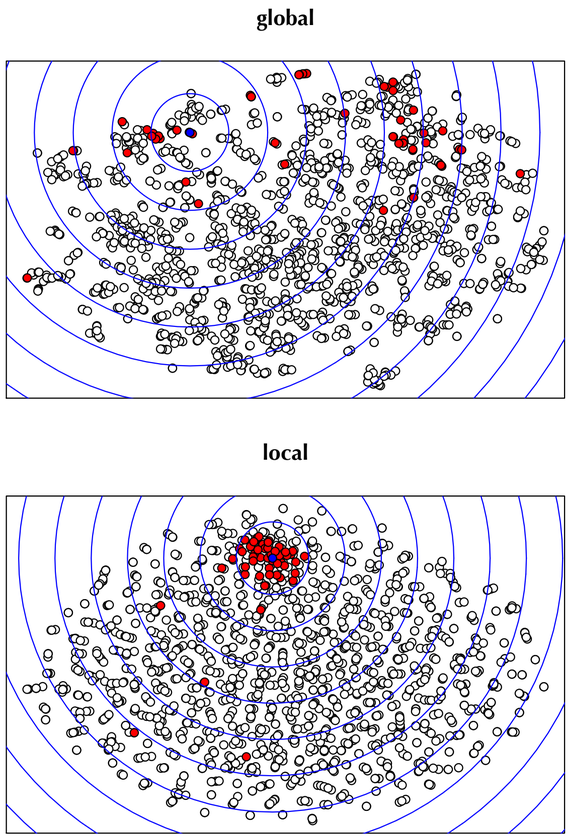
global	local
cutting	tax
squeeze	deficit
reduce	vote
slash	budget
reduction	reduction
spend	house
lower	bill
halve	plan
soften	spend
freeze	billion

Figure 3: Terms similar to 'cut' for a word2vec model trained on a general news corpus and another trained only on documents related to 'gasoline tax'.

Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. In Proceedings of the 54th annual meeting of the association for computational linguistics, 2016.

Topic-specific word embeddings

- "one sense per discourse" suggests that knowing the topic can inform a word's representation.
- given a string x, how do we determine the right topic?
- information retrieval provides a way to, given x, retrieve topically-relevant documents.
- can use these documents to train word embeddings using static methods.
  - although the embedding is static, it changes based on x.



Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. In Proceedings of the 54th annual meeting of the association for computational linguistics, 2016.

The problems with topic-specific embeddings

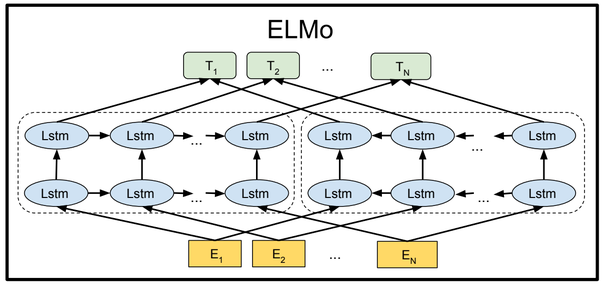
- Topic-specific word embeddings are bad because...
  - limited by retrieval system
  - need to learn a new embedding for every sentence
  - does not capture local context
- “One sense per collocation”: An ambiguous word will have a single sense within a local context (e.g., phrase, sentence) [Yarowsky et al., 1993].
  - “You shall know a word by the company it keeps!” [Firth 1957]

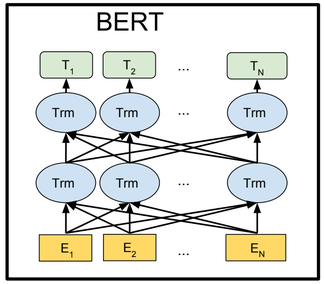
Previous work [Gale, Church and Yarowsky, 1992] showed that with high probability a polysemous word has one sense per discourse. In this paper we show that for certain definitions of collocation, a polysemous word exhibits essentially only one sense per collocation. We test this empirical hypothesis for several definitions of sense and collocation, and discover that it holds with 90-99% accuracy for binary ambiguities. We utilize this property in a disambiguation algorithm that achieves precision of 92% using combined models of very local context.

David Yarowsky. One sense per collocation. In Proceedings of the workshop on human language technology, HLT '93, 266--271, Stroudsburg, PA, USA, 1993. , Association for Computational Linguistics.
J.R. Firth. Studies in linguistic analysis: special volume of the philosogical society. Special Volume of the Philological Society. Blackwell, 1957.

Contextual Representations

- Contextual representations adapt a word's representation based on neighboring words.
- ELMo [Peters et al., 2018] uses bidirectional LSTMs to capture context from neighboring words.
- BERT [Devlin et al., 2019] uses bidirectional transformers to capture context from neighboring words.





Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), 2018.
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019.

Contextual Representations

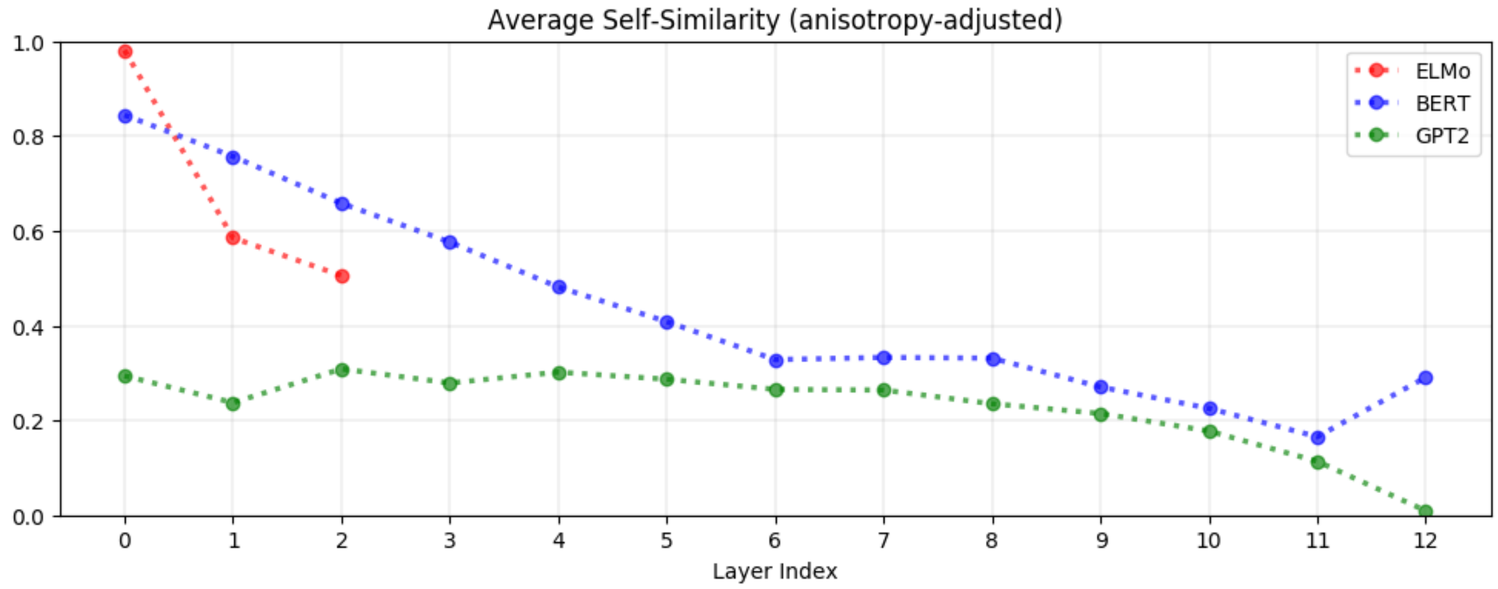


Figure 2: The average cosine similarity between representations of the same word in different contexts is called the word's *self-similarity* (see Definition 1). Above, we plot the average self-similarity of uniformly randomly sampled words after adjusting for anisotropy (see section 3.4). In all three models, the higher the layer, the lower the self-similarity, suggesting that contextualized word representations are more context-specific in higher layers.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp), 2019.

poll: which of these change word embeddings?

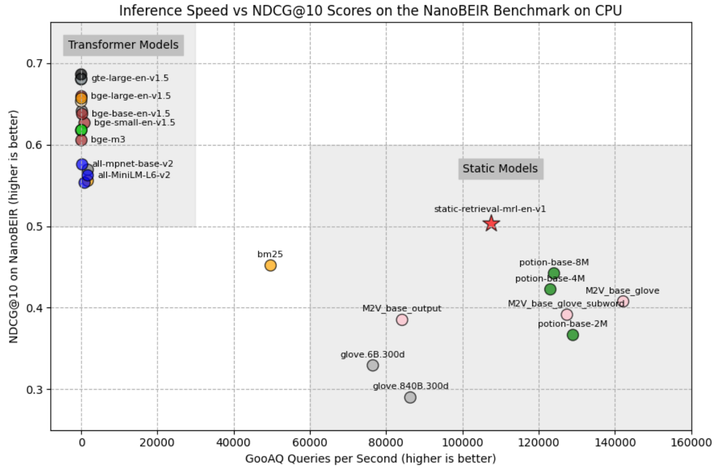
Finetuning, Prompting, LoRA

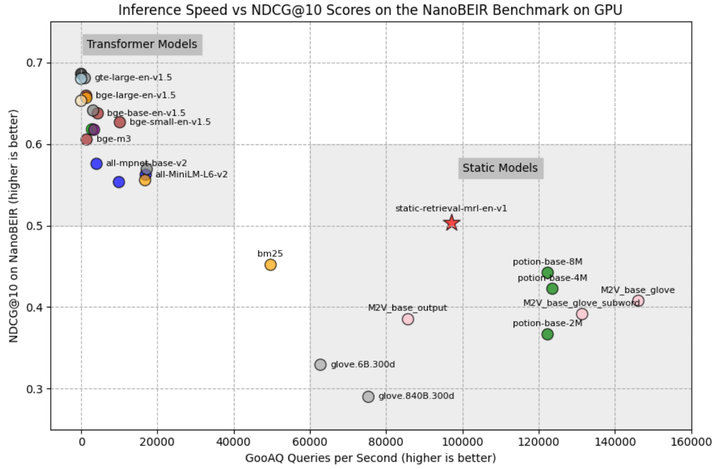
- finetuning *can* change input embeddings...but rarely does
- prompt/prefix tuning does not change input embeddings...but does change contextual embeddings
- LoRA...

Static vs Contextual Embeddings

Effectiveness

- Contextual embeddings are often more effective (e.g., higher precision),
  - 1.4 x (NDCG) effectiveness improvement from most effective static and most effective contextual
  - 1.4 x (NDCG) effectiveness improvement from fastest static and fastest contextual



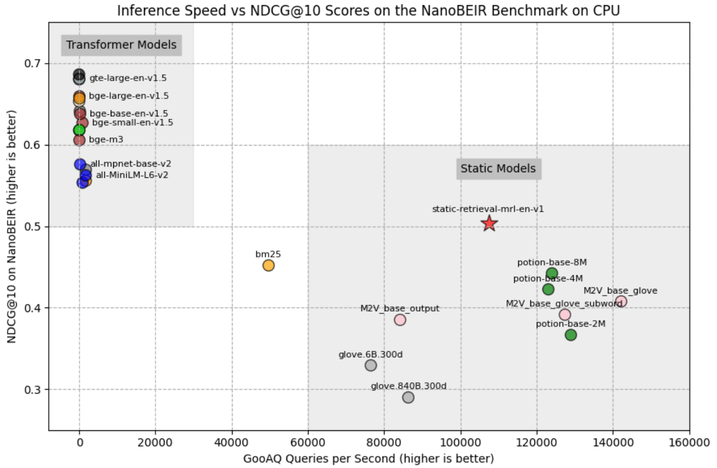


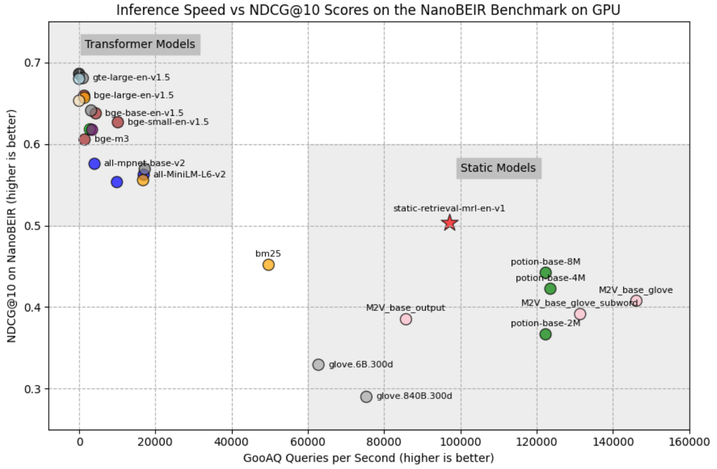
Tom Aarsen, "Train 400x faster Static Embedding Models with Sentence Transformers", January 15, 2025.

Static vs Contextual Embeddings

Efficiency

- Although contextual embeddings are often more effective, static embeddings involve a simple lookup.
  - 81 x (CPU)/8.6 x (GPU) speed improvement between fastest static and fastest contextual
  - 1918 x (CPU)/101 x (GPU) speed improvement between most effective static and most effective contextual
- Understand the constraints of your task (effectiveness, efficiency) when you select embeddings.





Tom Aarsen, "Train 400x faster Static Embedding Models with Sentence Transformers", January 15, 2025.

Multi-task training

- In addition to contextual representations, BERT leveraged the concept of multi-task learning (MTL).
- Originally developed in the machine learning community, MTL jointly optimizes model parameters to perform well across a variety of tasks [Caruana 1997].
- Often leads to strong improvements in effectiveness by leveraging relevant representations between tasks.
- Collobert and Weston [2008] applied to NLP.

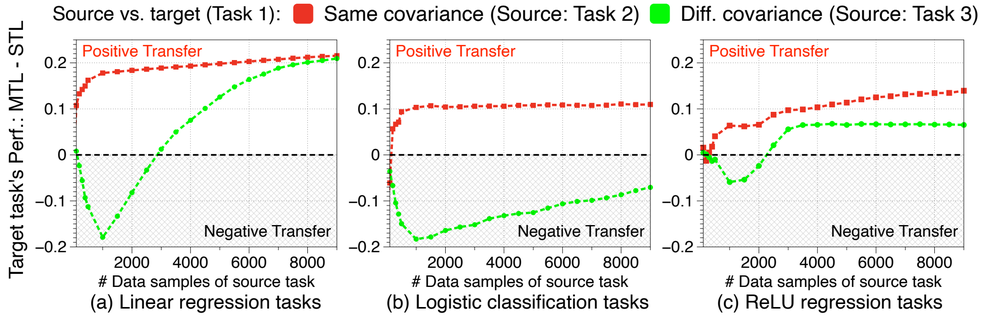
	ROOT-MEAN SQUARED ERROR ON TEST SET						
TASK	**Single Task Backprop (STL)**				**MTL**	Change MTL to Best STL	Change MTL to Mean STL
	2HU	4HU	8HU	16HU	16HU		
1 or 2 Lanes	.201	.209	.207	**.178**	**.156**	-12.4% *	-21.5% *
Left Edge	**.069**	.071	.073	.073	**.062**	-10.1% *	-13.3% *
Right Edge	.076	.062	.058	**.056**	**.051**	-8.9% *	-19.0% *
Line Center	.153	**.152**	.152	.152	**.151**	-0.7%	-0.8%
Road Center	.038	**.037**	.039	.042	**.034**	-8.1% *	-12.8% *
Road Greylevel	**.054**	.055	.055	.054	**.038**	-29.6% *	-30.3% *
Edge Greylevel	**.037**	.038	.039	.038	**.038**	2.7%	0.0%
Line Greylevel	.054	.054	**.054**	.054	**.054**	0.0%	0.0%
Steering	.093	**.069**	.087	.072	**.058**	-15.9% *	-27.7% *

Rich Caruana. Multitask learning. Machine Learning, 28:41---75, 1997.
Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on machine learning, 2008.

poll: when will multi-task learning hurt?

Multi-task training

- although MTL can be good in general, it can also lead to under-performance when,
  - task inconsistency: when tasks contradict in the appropriate representations, gradients can compete
  - task dominance: when some tasks have more data, those representations can dominate the learned representations
- tends not to be as much of an issue in NLP because,
  - fine-tuning can mitigate task inconsistency and dominance
  - language tasks have significant overlap



Performance improvement of a target task (Task 1) by MTL with a source task vs. STL. Red: positive transfer when the source is Task 2, which has the same covariance matrix with target. Green: negative (to positive) transfer when the source is Task 3, which has a different covariance from the target, as its # of samples increases.

If your application involves learning representations with multi-task learning, pay attention to task correlations and distribution

Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In International conference on learning representations, 2020.

Representation bias

- Task dominance is one way that representations can be influenced in unexpected ways by training.
- Learned representations can be influenced by,
  - architecture
  - labels
  - data
- Social biases reflected in training data have been well-documented [Kurita et al., 2019].
- More generally, training data representation can significantly impact representations across a variety of dimensions [Rolf et al., 2021]

Category	Ours on BERT *Log Probability Bias Score*
Pleasant/Unpleasant (Insects/Flowers)	0.8744*
Pleasant/Unpleasant (EA/AA)	0.8864*
Career/Family (Male/Female)	1.126*
Math/Arts (Male/Female)	0.8495*
Science/Arts (Male/Female)	0.9572*

If your application involves using pre-trained representations, make sure you understand the implicit bias.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, Proceedings of the first workshop on gender bias in natural language processing, 2019.
Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. Representation matters: assessing the importance of subgroup allocations in training data. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th international conference on machine learning, volume 139 of Proceedings of Machine Learning Research, 9040--9051, , PMLR, 18--24 Jul 2021.

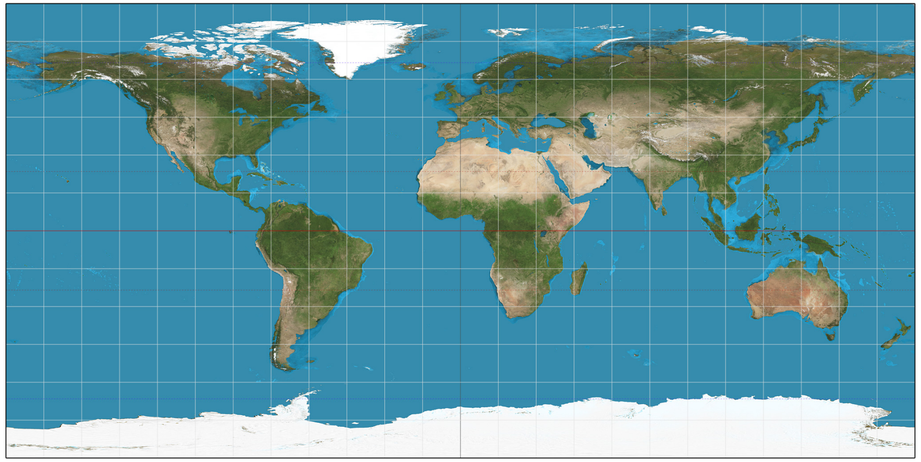
Geometry of embeddings

- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.

consider the earth...

Geometry of embeddings

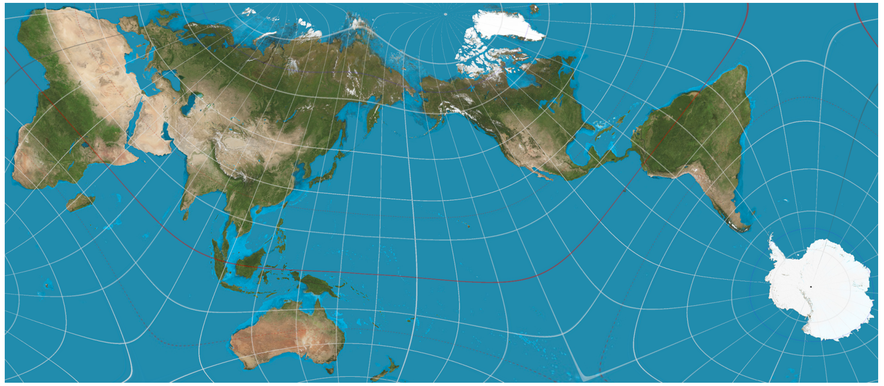
- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.



is this a good embedding into two dimensions?

Geometry of embeddings

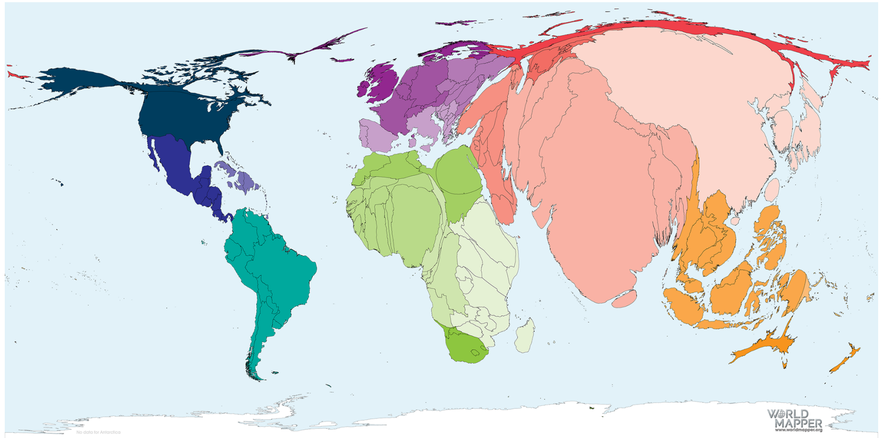
- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.



is this a good embedding into two dimensions?

Geometry of embeddings

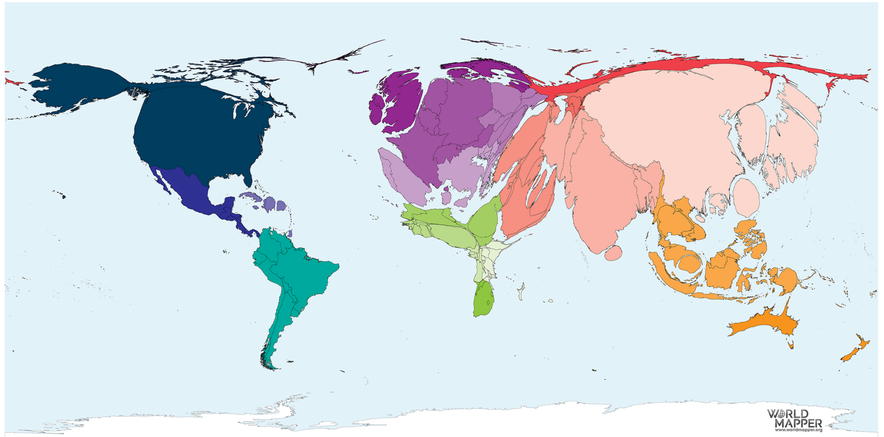
- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.



is this a good embedding into two dimensions?

Geometry of embeddings

- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.

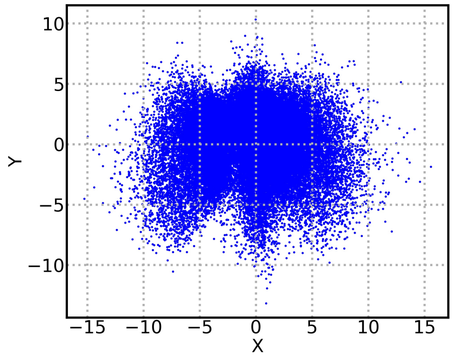


is this a good embedding into two dimensions?

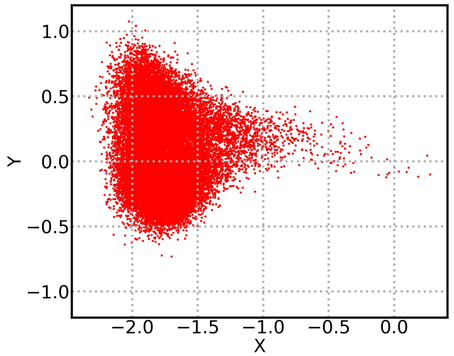
Geometry of embeddings

- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.
- Transformer-based embeddings tend to be anisotropic, isolated to a narrow cone [Gao et al., 2019; Ethayarajh 2019; Meng et al., 2021]

word2vec



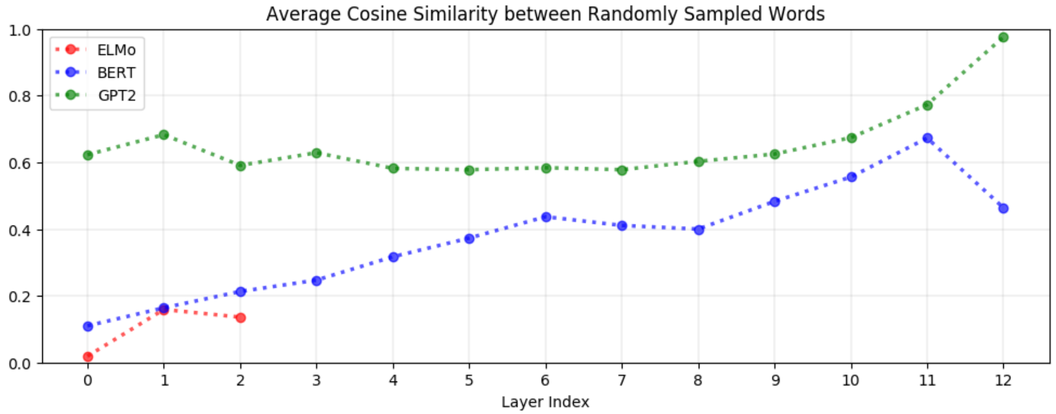
transformer



Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In International conference on learning representations, 2019.
Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp), 2019
Yu Meng, Chenyan Xiong, Payal Baja, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. Coco-lm: correcting and contrasting text sequences for language model pretraining. In Proceedings of the 35th international conference on neural information processing systems, 2021.

Geometry of embeddings

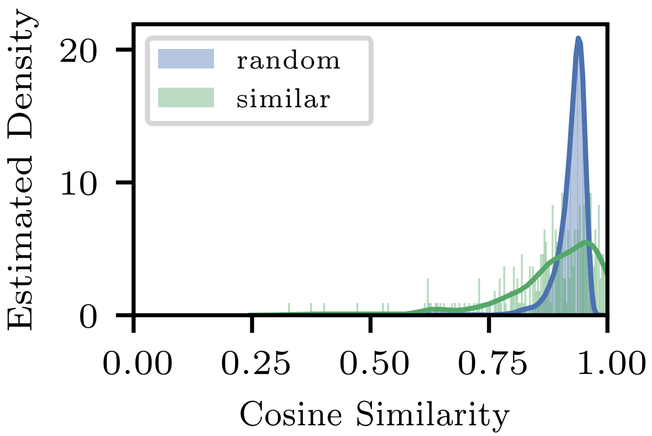
- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.
- Transformer-based embeddings tend to be anisotropic, isolated to a narrow cone [Gao et al., 2019; Ethayarajh 2019; Meng et al., 2021]



Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In International conference on learning representations, 2019.
Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp), 2019
Yu Meng, Chenyan Xiong, Payal Baja, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. Coco-lm: correcting and contrasting text sequences for language model pretraining. In Proceedings of the 35th international conference on neural information processing systems, 2021.

Geometry of embeddings

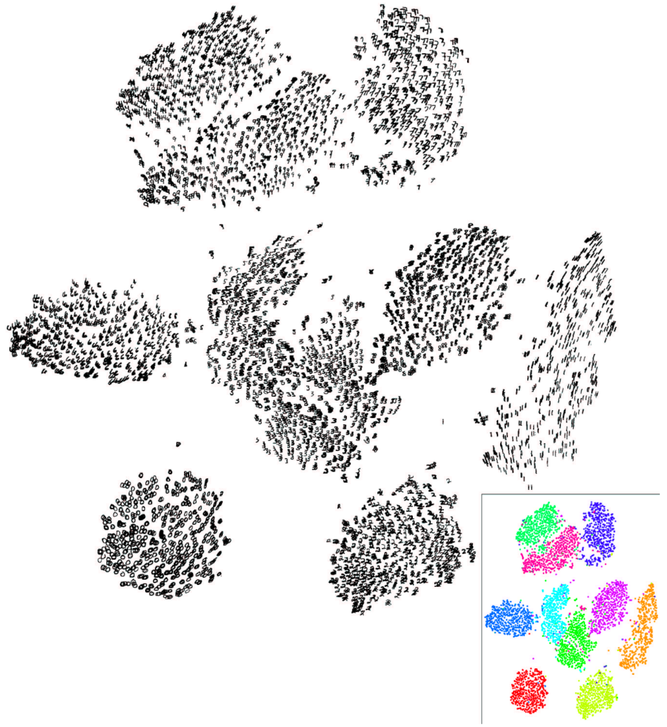
- Projection from one embedding to another involves,
  - compression: may throw away information during embedding; what to throw away determined by the training objective.
  - distortion: will expand or contract representations during embedding; how to distort determined by the training objective.
- Transformer-based embeddings tend to be anisotropic, isolated to a narrow cone [Gao et al., 2019; Ethayarajh 2019; Meng et al., 2021]



Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In International conference on learning representations, 2019.
Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp), 2019
Yu Meng, Chenyan Xiong, Payal Baja, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. Coco-lm: correcting and contrasting text sequences for language model pretraining. In Proceedings of the 35th international conference on neural information processing systems, 2021.

Embedding visualization

- t-SNE is the dominant method of visualizing embeddings.
  - Preserves local neighborhoods by converting high-dimensional distances into probabilities representing how likely points are to be neighbors, then arranges points in 2D so similar items stay close together
- embeddings are already a projection
- t-SNE projects further into two-dimensional space.



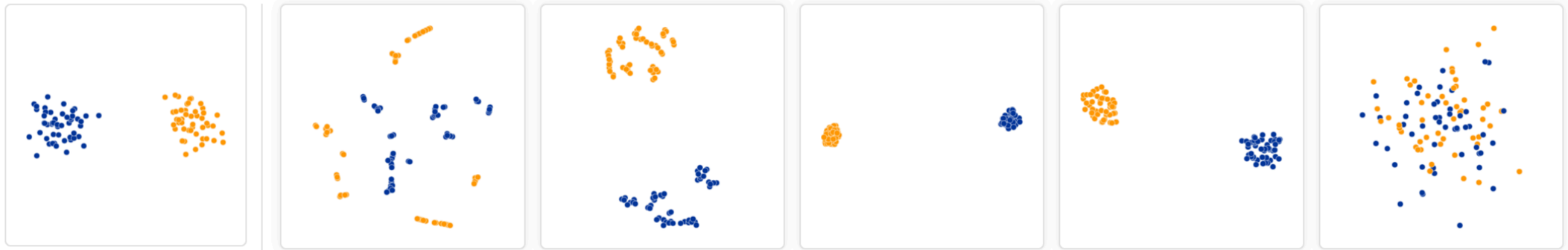
Visualization of 6,000 digits from the MNIST data set produced by the random walk version of t-SNE (employing all 60,000 digit images)

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579--2605, 2008.

poll: t-SNE can be deceptive when...

t-SNE

Hyperparameters matter



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

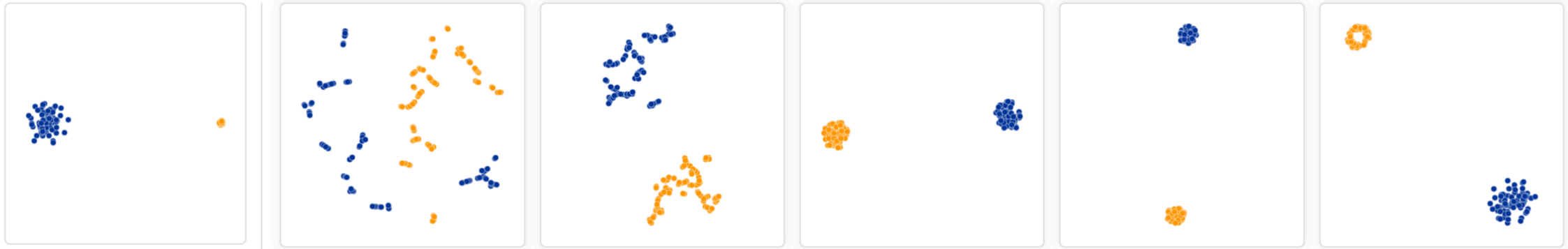
Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016.

t-SNE

Cluster sizes in a t-SNE plot mean nothing



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016.

t-SNE

Distances between clusters might not mean anything

Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

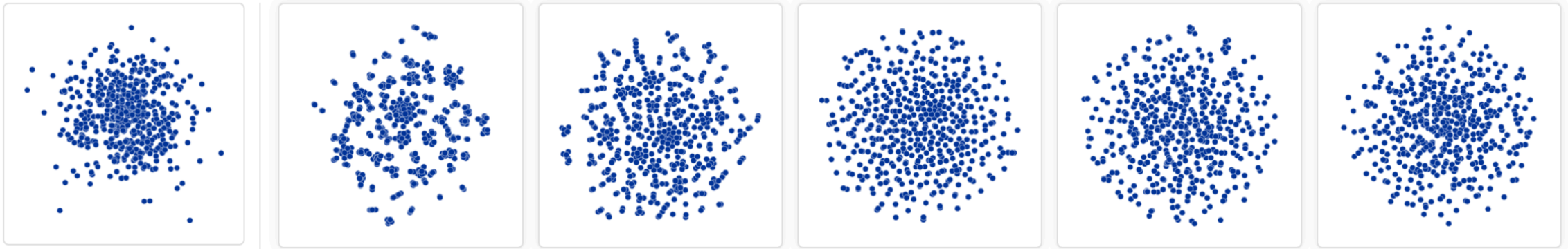
Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016.

t-SNE

Random noise doesn't always look random



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016.

t-SNE

Hallucinating shapes

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016.

Sparsity

- So far, we've been discussing dense embeddings.

poll: we prefer dense representations when...

poll: we prefer sparse representations when...

Sparsity

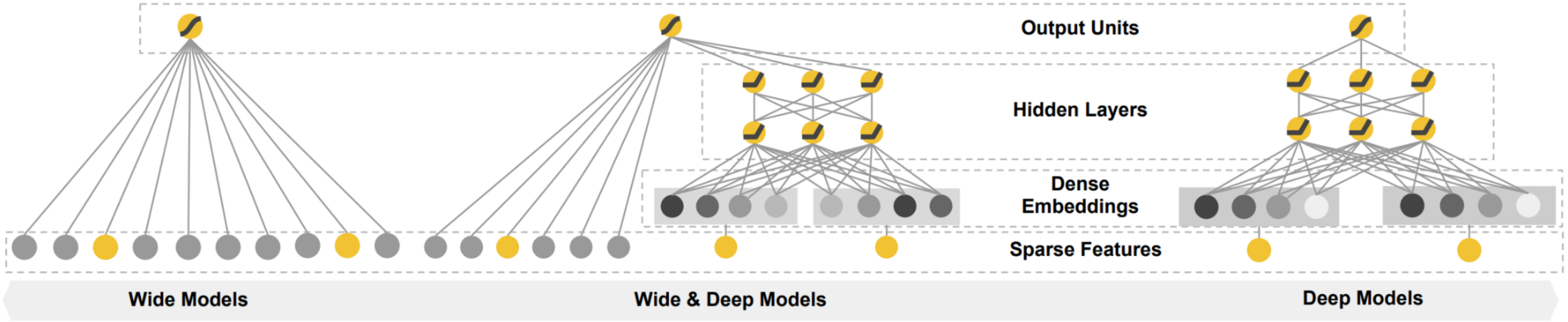
- So far, we've been discussing dense embeddings.

- Dense representations
  - native density: sensor readings
  - learned density: lower-dimensional projections; distributed representations
  - supports generalization; more limited memorization
  - efficiency reduces with dimensionality

- Sparse representations
  - native sparsity: one-hot representations; tf.idf vectors
  - learned sparsity: directly learn or post-process representations to ensure sparsity
  - supports memorization; more limited generalization
  - efficiency reduces with density

Hybrid representations

Wide and deep models

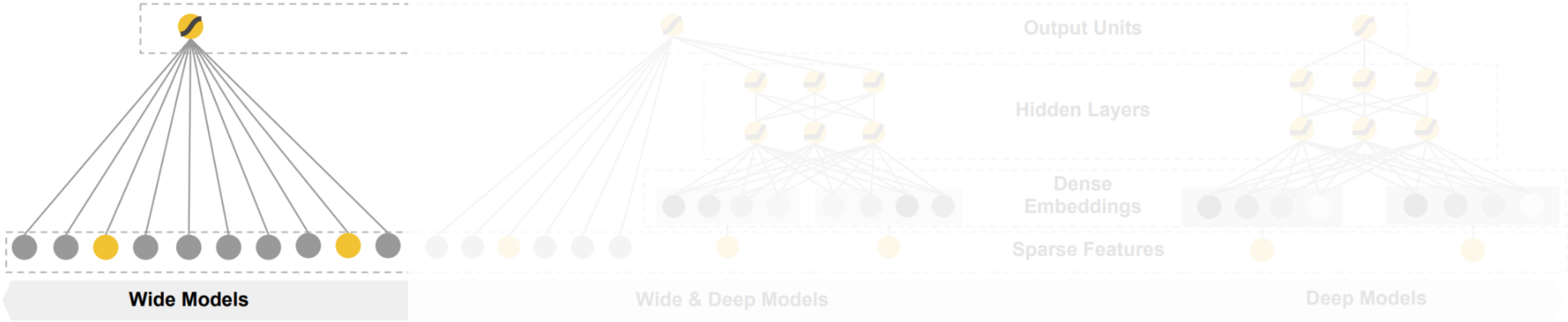


can combine benefits of sparse feature memorization with generalization from deep networks.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

Wide and deep models



memorization

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

Wide network

- wide part is a linear model that captures memorization
  - learns explicit co-occurrence patterns between features (e.g., user-item interactions, popular item biases).
  - memorizes frequent item-user interactions (e.g., "User A likes Action movies").
  - uses feature crosses (e.g., "Users in California like Surfing gear").
  - handles cold-start issues better when engineered features are useful.

- examples
  - user id (~m features), item id (~n features)
  - user id x item id (m x n features)

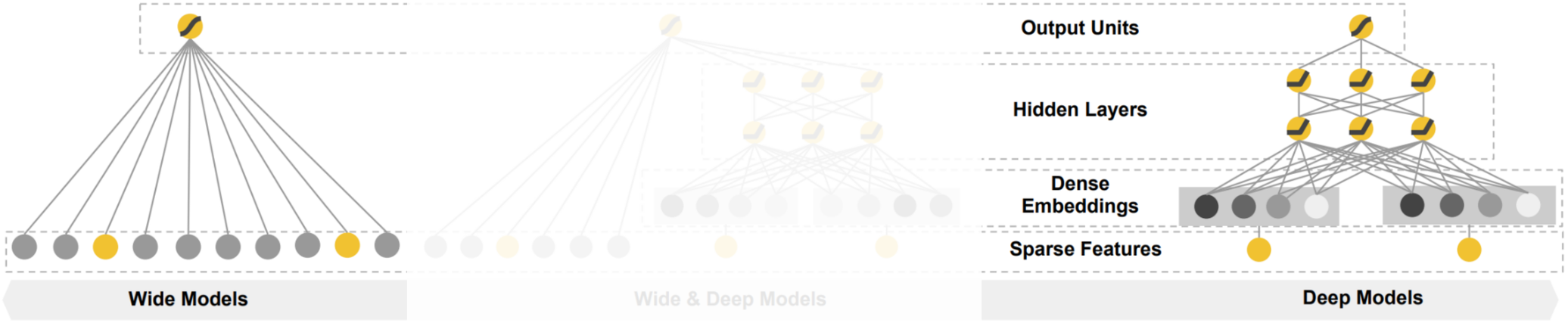
- when it is helpful
  - categorical features with strong co-occurrence patterns.
  - recommendation settings where past interactions dictate behavior (e.g., repeat purchases).

- limitations
  - cannot generalize well to unseen feature combinations (e.g., new items, new users).
  - requires manual feature engineering (e.g., defining which features to cross).

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

Wide and deep models



memorization

generalization

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

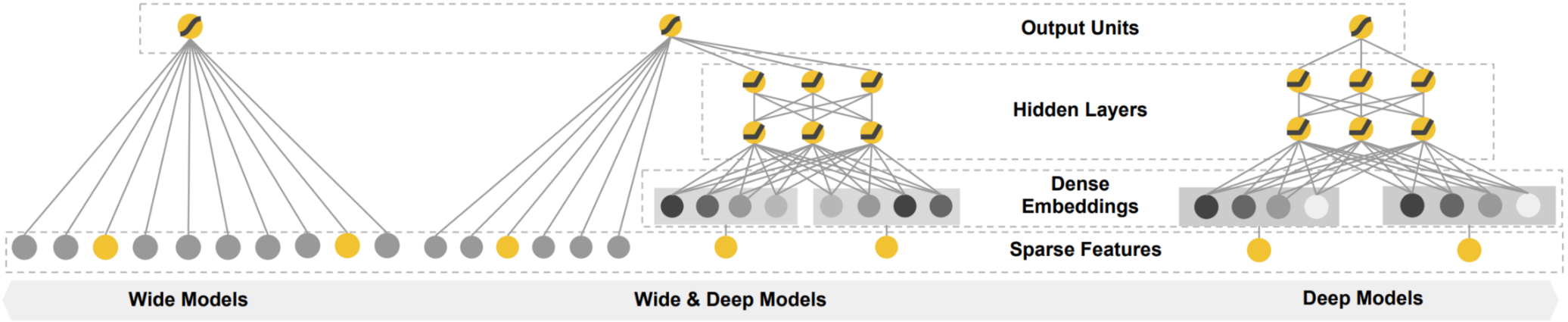
Deep network

- deep part captures generalization by learning complex, nonlinear relationships between features.
- automatically discovers feature interactions that not explicitly engineered in the wide part.
- embedding layer
  - converts sparse features (e.g., individual user/item IDs) into dense vectors.
  - user 123 → [0.1, -0.3, 0.7, ...]
- deep network
  - stack of fully connected layers
- when it is helpful
  - latent relationships between features (e.g., "Users who like X tend to like Y, even if never co-occurring in history").
  - unseen users/items (generalization through embeddings).
  - unknown feature interactions (rather than manually defining feature crosses in wide).
- when it is helpful
  - Requires more data than the wide model to generalize well.
  - Harder to interpret than the wide model.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

Wide and deep models



memorization

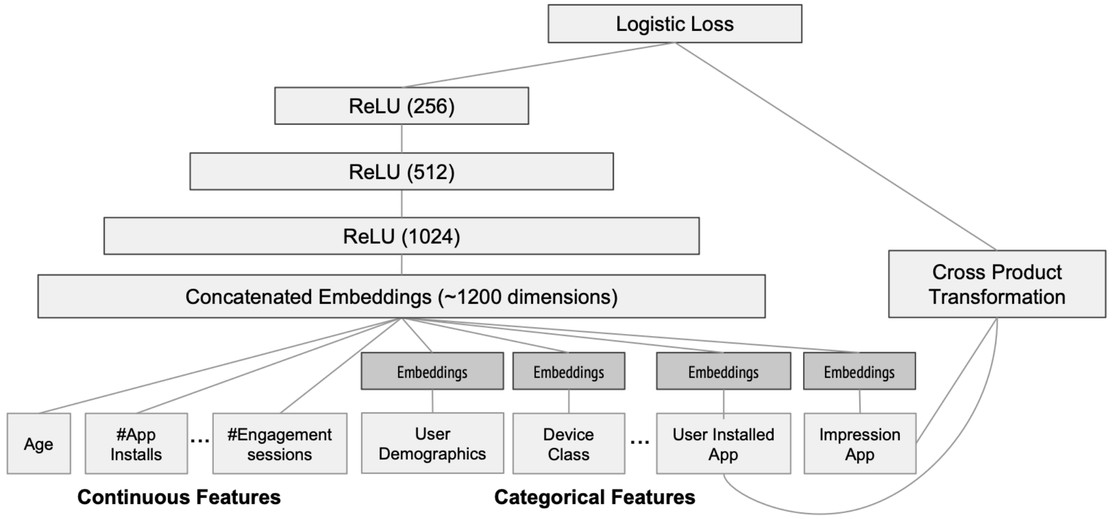
memorization and generalization

generalization

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Hybrid representations

Wide and deep models



Model	Offline AUC	Online Acquisition Gain
Wide (control)	0.726	0%
Deep	0.722	+2.9%
Wide & Deep	0.728	+3.9%

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems, DLRS 2016.

Next time

- moving from embeddings to retrieval...