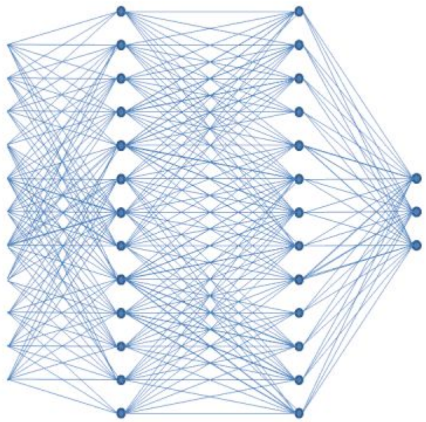
TEXAS
The University of Texas at Austin

In-process Diagnostic methods for Entity Representation Learning on Sequential Data at Scale

PhD Defense Presentation: Diego Garcia-Olano
Advisor: Dr. Joydeep Ghosh

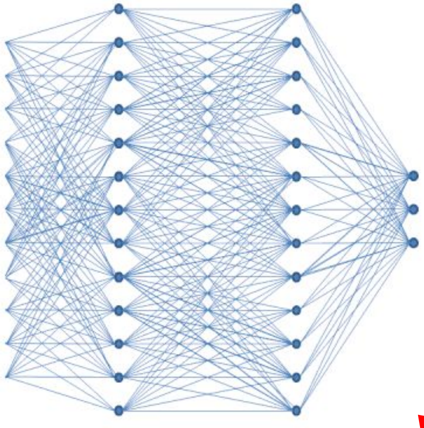
July 22, 2022

Explainable AI for Sequential Data



For image, text and time series data tasks, deep learning neural nets have become the default modeling choice.

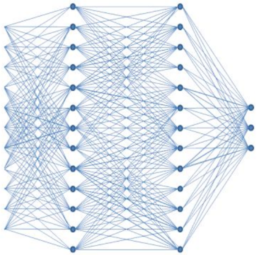
Explainable AI for Sequential Data



For image, text and time series data tasks, deep learning neural nets have become the default modeling choice.

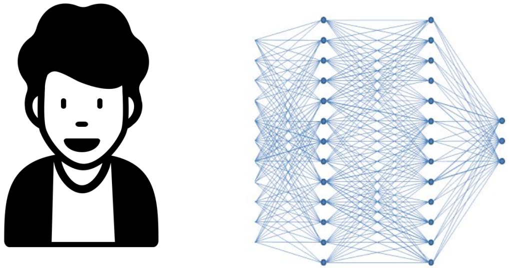
Their ubiquity necessitates **transparency** into how such models arrive at the predictions they make in order that they be deemed **trustworthy** for use in critical domains.

- **Who** are we explaining to:
  End user?   Expert/Researcher?
  Model developers?   Other Models?

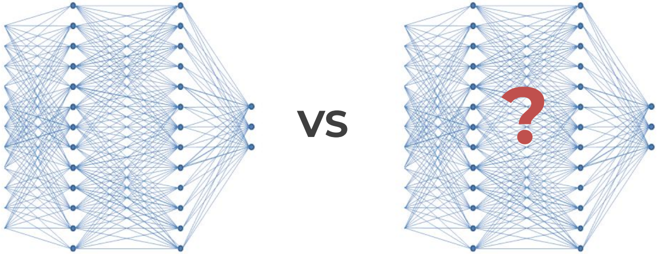


Explainable AI for Sequential Data

- **Who** are we explaining to:
  End user? Expert/Researcher?
  Model developers? Other Models?



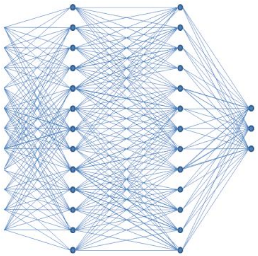
- **White Box** vs **Black Box**:
  Do we have access to the model internals?
  The data it was trained on?

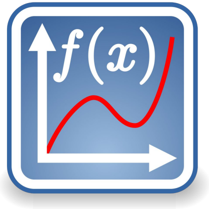


- **Explaining from what point in model process**:
  Pre-model, In-Process or Post Hoc

Secondary g(f(x))

data ->

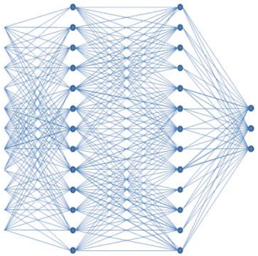




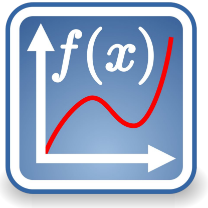
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (Rudin, et al, 2019 Nature)

- **Explaining from what point in model process**:
  Pre-model, In-Process or Post Hoc

data ->



Secondary g(f(x))



Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (Rudin, et al, 2019 Nature)

- **Global** model vs **Individual** instance based explanations

Types of Explainable AI

Post Hoc explanations

Train a secondary model to explain a primary model of interest

Examples

Feature Attribution: (IG, SHAP, etc) pixels/words that lead to model decision
Influential examples: which training data most influenced a model’s output
BERT probing: assess how well a LM encodes properties of language

Types of Explainable AI

Post Hoc explanations

Train a secondary model to explain a primary model of interest

Examples

Feature Attribution: (IG, SHAP, etc) pixels/words that lead to model decision
Influential examples: which training data most influenced a model's output
BERT probing: assess how well a LM encodes properties of language

Issues with Post Hoc secondary model explainers

- Feature Importance independent of task
- Do local or linear approximations give faithful explanations of a primary, possibly very non-linear model ?

Explaining a network's behavior in a way that it wasn't expressly trained for can be problematic & makes assumptions that often do not hold (Chen, Rudin '20)

In-Process methods are designed with explainability in mind

Examples

Prototypes: learn “prototypical” representations
Deep k-NN models: utilize layer representations as additional “clustering” features
Concept based Models: layer specific additional task loss with supervision
Retrieval as Explanation: for tasks involving entity retrieval as an intermediate step

Require access and modifications to the underlying model ….

In-Process methods are designed with explainability in mind

Examples

Prototypes: learn "prototypical" representations

Deep k-NN models: utilize layer representations as additional "clustering" features

Concept based Models: layer specific additional task loss with supervision

Retrieval as Explanation: for tasks involving entity retrieval as an intermediate step

Require access and modifications to the underlying model ….

which is fine for critical applications!

Objectives

In-process explainable models for Sequential Data

- **are an Useful & Under-explored area for sequential data modeling**
- **provide Interpretable and Faithful explanations of model decisions**
- **allow for model “diagnosis” and intervention at inference time.**

Objectives

In-process explainable models for Sequential Data

- **are an Useful & Under-explored area for sequential data modeling**
- **provide Interpretable and Faithful explanations of model decisions**
- **allow for model "diagnosis" and intervention at inference time.**

Entity Representation learning allows for an additional interesting and underexplored explainability aspect that grounds models.

Scalability is vital to the adoption of models in practice
Both play a central role in this work.

Completed Work

Pre-Proposal Works

- Learning Dense Representations for Entity Retrieval. (CoNLL 2019)
- Deep Classification of Time-Series Data with Learned Prototype Explanations. (ICML time series workshop 2019 *joint work with Alan Gee*)
- Biomedical Interpretable Entity Representations. (ACL-IJCNLP 2021)

Post Proposal Works

- Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection (WWW 22. Multimodal Understanding for the Web and Social Media workshop)
- Intermediate Entity-based Sparse Interpretable Representation Learning. *under submission*

Completed Work (Pre-Proposal)

Learning Dense Representations for Entity Retrieval. (CoNLL 2019)	Constructed a **dual mention-entity encoder** that learns dense representations for efficient neural **Entity Retrieval** with an **in-process, iterative hard negatives procedure** for **model learning and inference time inspection**.
Deep Classification of Time-Series Data with Learned Prototype Explanations. (ICML 19)	Adapted a **prototypical autoencoder** classifier to be compatible with **time series data** and allow for **tunable prototype diversity** leading to improved accuracy and **global and instance level explanations**.
Biomedical Interpretable Entity Representations. (ACL-IJCNLP 2021)	Learned a distantly supervised entity type system and data set for use in training a **Biomedical Interpretable Entity model** whose representations exist in a **semantically meaningful vector space** & whose **predictions may be interpreted and diagnosed** with an oracle method.

Completed Work (Pre-Proposal)

Learning Dense Representations for Entity Retrieval. (CoNLL 2019)	Constructed a **dual mention-entity encoder** that learns dense representations for efficient neural **Entity Retrieval** with an **in-process, iterative hard negatives procedure** for **model learning and inference time inspection**.
Deep Classification of Time-Series Data with Learned Prototype Explanations. (ICML 19)	Adapted a **prototypical autoencoder** classifier to be compatible with **time series data** and allow for **tunable prototype diversity** leading to improved accuracy and **global and instance level explanations**.
Biomedical Interpretable Entity Representations. (ACL-IJCNLP 2021)	Learned a distantly supervised entity type system and data set for use in training a **Biomedical Interpretable Entity model** whose representations exist in a **semantically meaningful vector space** & whose **predictions may be interpreted and diagnosed** with an oracle method.

Learning Dense Representations for Entity Retrieval

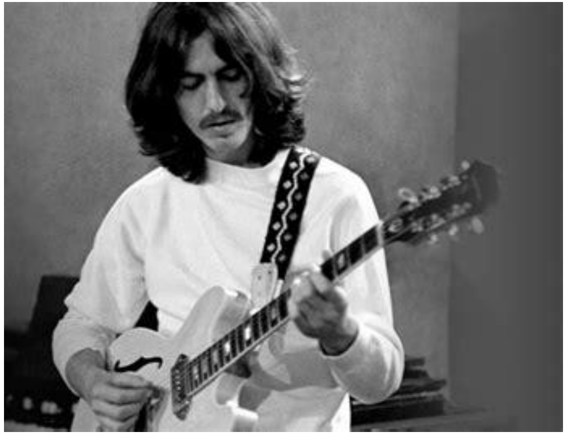
Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, Eugene., Garcia-Olano, D. "Learning Dense Representations for Entity Retrieval". Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 2019.

Entity Resolution

Example Query: What is George Harrison's favorite Nintendo game?

Beatles Guitarist

Highest Popular Prior



Former Senior VP of Marketing

at Nintendo of America.

Wiki Entity IDs Q2643 Q5540278

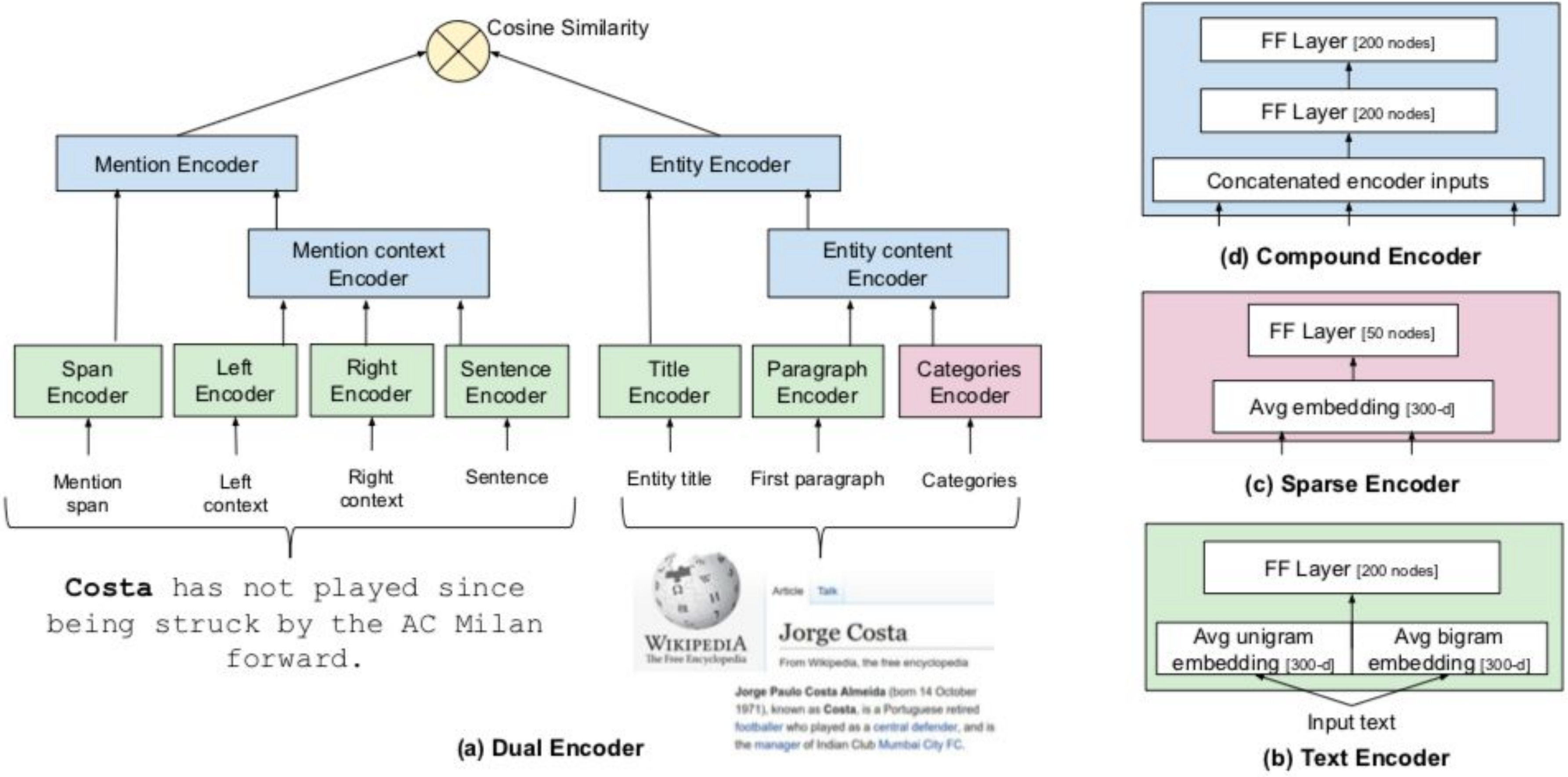
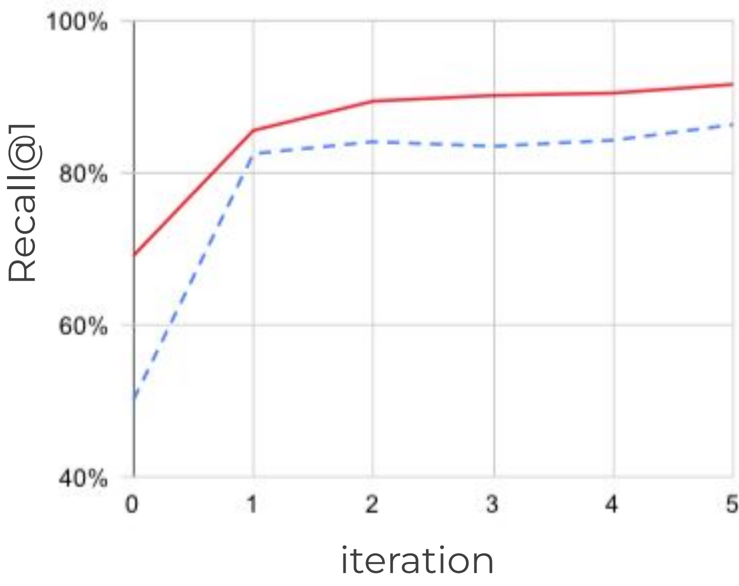


Figure 1: Architecture of the dual encoder model for retrieval (a). Common component architectures are shown for (b) text input, (c) sparse ID input, and (d) compound input joining multiple encoder outputs. Note that all text encoders share a common set of embeddings.

Hard Negative In-Process Explanations

During each iteration of training,
we **identify entities which our model assigns a higher ranking than the true entity** associated with a given mention and context.

These **hard negatives** **can be inspected over time** during training or inference to assess the mention/contexts and entities that are added which are difficult for the model to learn (esp. later iterations)



This **interpretable in-process information about the learning process** could be used to:

- improve **error analysis**,
- identify cases where **additional supervision** could be useful
- gauge **confidence** in inference time predictions

Summary of Contributions

Proposed **first neurally learned**, robust & efficient **approach to Entity Resolution**

Define a **novel dual encoder architecture** for
learning entity and mention **embeddings** suitable **for retrieval**

Describe a fully **unsupervised, hard-negative mining** algorithm
that greatly improves retrieval performance and
can be used **to track and explain model learning.**

Approximate nearest neighbor search yields quality candidate entities efficiently.

Outperform discrete retrieval baselines (alias table, BM25) and
gives results competitive with the best reported accuracy on TACKBP-2010.

Biomedical Interpretable Entity Representations

Garcia-Olano, D., Onoe, Y., Baldini, I., Ghosh, J., Wallace, B., Varshey, K. “Biomedical Interpretable Entity Representations”. Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)

Motivation

Entities over text = typically embedded in dense vector spaces with pre-trained language models (BERT,etc).

[0.519, 0.917, -0.935, 0.891, 0.396, 0.711, 0.479, 0.417, 0.744, -0.254,
-0.174, 0.233, -0.315, 0.497, -0.516, 0.22, -0.679, 0.389, -0.683, 0.909,
23, 0.528, 0.116, 0.334, 0.717, 0.857, -0.262, 0.624, -0.178, -0.045, -0.
-0.952, 0.4, 0.356, 0.091, 0.976, -0.337, -0.002, 0.054, 0.512, -0.312,
.278, -0.409, -0.655, -0.294, -0.453, 0.735, 0.461, 0.282, -0.43, -0.838,
3, -0.736, -0.001, 0.889, -0.228, 0.645, 0.883, 0.805]

[0.656, 0.407, 0.568, -0.035, -0.842, -0.257, 0.202, -0.31, 0.886, 0.386,
34, -0.823, -0.929, -0.068, -0.238, 0.236, -0.463, 0.56, -0.687, -0.521,
88, 0.54, 0.047, -0.434, -0.009, 0.59, 0.971, 0.798, 0.202, 0.225, 0.131,
88, 0.44, -0.835, -0.032, -0.935, 0.318, 0.72, -0.23, -0.903, 0.912, -0.8
0.981, -0.23, 0.797, -0.785, -0.583, 0.055, -0.511, 0.413, -0.757, 0.914,
943, 0.62, -0.78, 0.888, 0.288, 0.807, -0.207, -0.284]

Motivation

Entities over text = typically embedded in dense vector spaces
with pre-trained language models (BERT,etc).

```
>>> word_embedding_for_happy
[0.519, 0.917, -0.935, 0.891, 0.396, 0.711, 0.479, 0.417, 0.744, -0.254,
-0.174, 0.233, -0.315, 0.497, -0.516, 0.22, -0.679, 0.389, -0.683, 0.909,
23, 0.528, 0.116, 0.334, 0.717, 0.857, -0.262, 0.624, -0.178, -0.045, -0.
 -0.952, 0.4, 0.356, 0.091, 0.976, -0.337, -0.002, 0.054, 0.512, -0.312,
.278, -0.409, -0.655, -0.294, -0.453, 0.735, 0.461, 0.282, -0.43, -0.838,
3, -0.736, -0.001, 0.889, -0.228, 0.645, 0.883, 0.805]
```

```
>>> word_embedding_for_sad
[0.656, 0.407, 0.568, -0.035, -0.842, -0.257, 0.202, -0.31, 0.886, 0.386,
34, -0.823, -0.929, -0.068, -0.238, 0.236, -0.463, 0.56, -0.687, -0.521,
88, 0.54, 0.047, -0.434, -0.009, 0.59, 0.971, 0.798, 0.202, 0.225, 0.131,
88, 0.44, -0.835, -0.032, -0.935, 0.318, 0.72, -0.23, -0.903, 0.912, -0.8
0.981, -0.23, 0.797, -0.785, -0.583, 0.055, -0.511, 0.413, -0.757, 0.914,
943, 0.62, -0.78, 0.888, 0.288, 0.807, -0.207, -0.284]
```

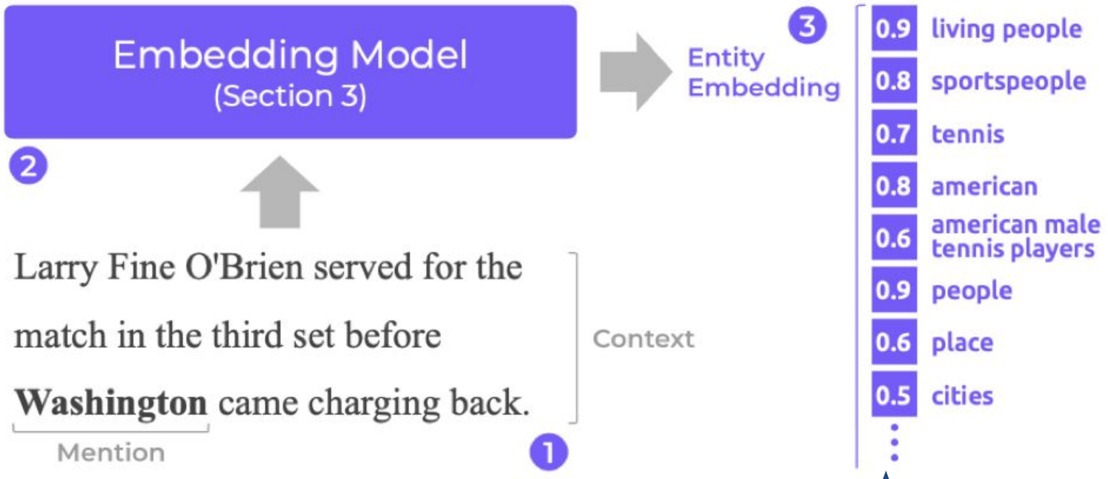
Not immediately interpretable.

Dense Entity Embeddings = Give good performance for entity-related tasks, but using them in those tasks requires additional processing in neural models.

IERs

Onoe et al* learn human readable interpretable entity representations that achieve high performance without additional learning ("out of the box")

fine grained entity types



represent probability of entity have corresponding properties

experiments using Ultra Fine Entity Type system **(10k)**
and Wikipedia Categories Type System **(60k)**

Problem setup: Interpretable Entity Representations

s = a sequence of **context words**,
m = an **entity mention span in s**.
t $\in [0, 1]^T$ binary vector of **entity types** over types in T

Goal: Learn parameters θ of a function f that
maps the mention m and its context s
⇒ to a vector t
that captures salient features of the entity mention in its context

High dimensional Multi-label classification task over entity types

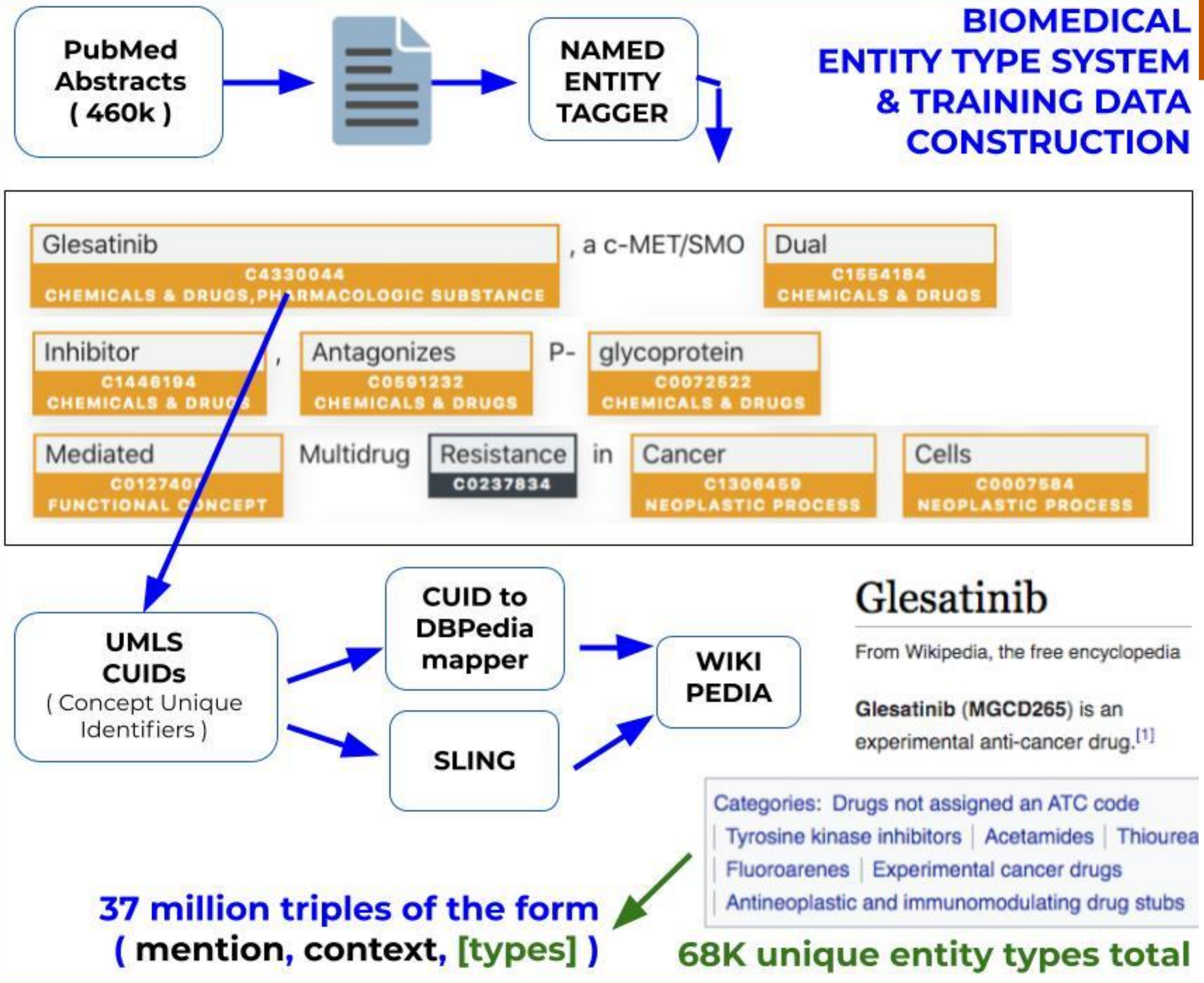
Can we adapt IERs for the **Biomedical Domain?**

[Glesatinib] is a dual inhibitor of c-Met and SMO
that is under phase II clinical trial for non-small cell lung cancer.

Most probable entity types for mention/context

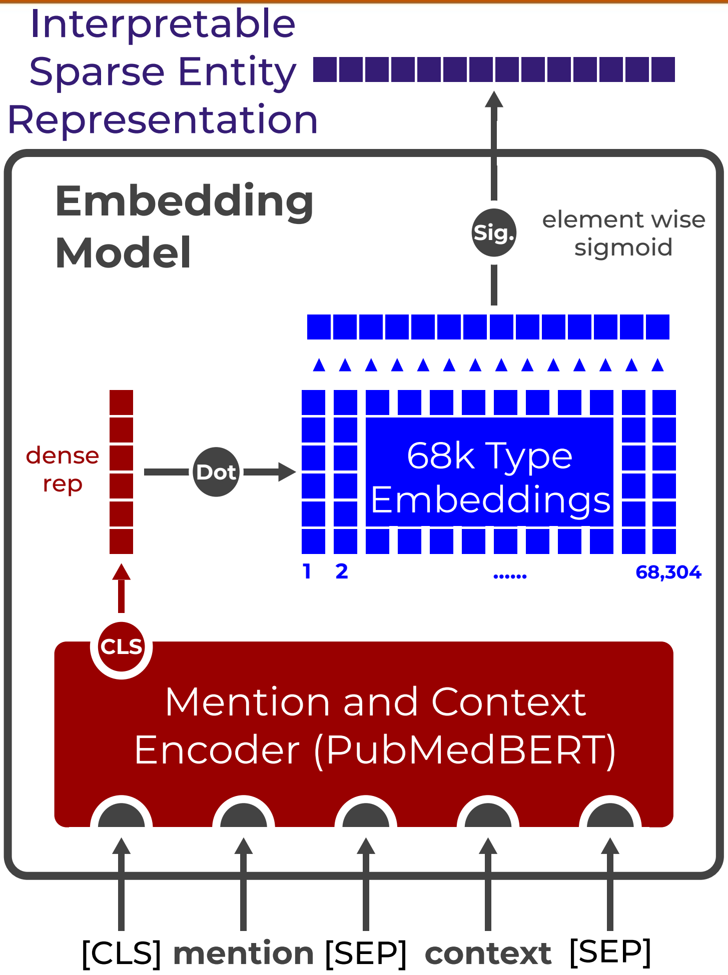
```
world health organization essential medicines : 0.4941
                                    pyridines : 0.4073
                                        diols : 0.3539
                            cancer treatments : 0.3260
                           carboxylate esters : 0.2376
                               chloroarenes : 0.1984
                                          rtt : 0.1879
                 hormonal antineoplastic drugs : 0.1768
                          antineoplastic drugs : 0.1037
                                     alcohols : 0.0771
                                     prodrugs : 0.0315
                                     peptides : 0.0300
                                methyl esters : 0.0223
                                        merck : 0.0191
                       transgender and medicine : 0.0135
                                  teratogens : 0.0130
 world anti-doping agency prohibited substances : 0.0124
                  peripherally selective drugs : 0.0103
                               human proteins : 0.0099
                                        ureas : 0.0090
                             withdrawn drugs : 0.0089
                       iarc group 2a carcinogens : 0.0073
                             prostate cancer : 0.0066
                                  mechanisms : 0.0066
                                chemotherapy : 0.0058
                          aromatase inhibitors : 0.0057
```

of 60k wiki entity types

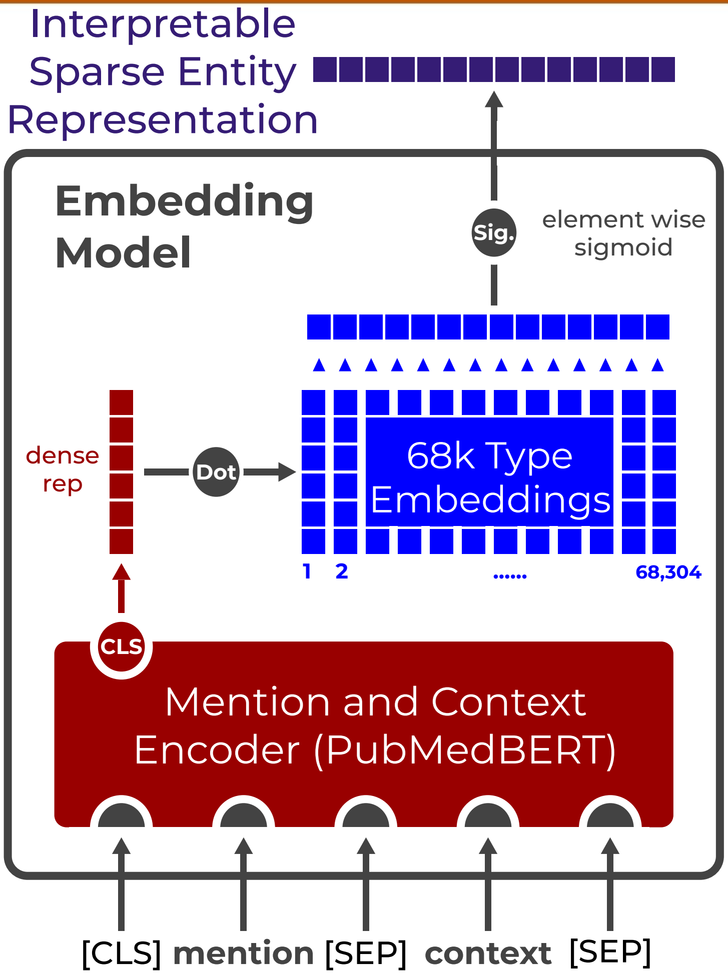


Distant Supervision to **construct** **Entity Type System** and **Training Data**.

Inference using Biomedical IERs



Inference using Biomedical IERs



Training loss:

Independent sum
of binary cross entropy losses
over all all entity types T
over all training examples D.

$$-\sum_{i}^{D}\sum_{j}^{T} t_{ij}^{*}\cdot\log(t_{ij}) + (1-t_{ij}^{*})\cdot\log(1-t_{ij}),$$

where t_{ij}^{*} *is the true label value (0 or 1)*
for data instance i's jth component

Inference via distance metric (cosine sim, dot prod)
between Biomedical IERs
without fine-tuning
(*leverages quantized based*
efficient similarity search)

Using BIERs

(1) **Named Entity Disambiguation** (NED) on Clinical Entities.

Model	Test Acc. Dot Prod	Test Acc. Cosine Sim
BIER-PubMedBERT (ours)	80.1	**84.0**
BIER-SciBERT (ours)	76.4	77.3
BIER-BioBERT (ours)	71.9	75.9
Onoe and Durrett (2020)	63.6	69.8
Popular Prior	73.9	-
PubMedBERT (Gu et al., 2020)	77.6	-
SciBERT (Beltagy et al., 2019)	77.4	-
BioBERT (Lee et al., 2019)	77.9	-

Table 2: BIER zero shot test results vs Logistic Regression Baselines trained on task data for NED task

(2) **Entity label Classification** for Cancer Genetics

Model	Test Acc. L2 Dist Dense	Test Acc. L2 Dist Sparse	Test Acc. Dot Prod Dense	Test Acc. Dot Prod Sparse
BIER-PubMedBERT	85.5	86.8	**88.2**	**87.5**
BIER-SciBERT	70.8	77.0	72.8	76.8
BIER-BioBERT	83.4	85.9	85.6	86.8
Onoe and Durrett (2020)	63.9	55.1	60.0	59.9
PubMedBERT	77.3	-	69.3	-
SciBERT	74.4	-	75.2	-
BioBERT	67.6	-	59.6	-

Table 3: Test accuracy on Cancer Genetics data using a nearest neighbor classifier (k=1) without fine-tuning based on sparse output or intermediate dense embeddings using L2 or Dot Product distance metrics.

Using BIERs

(2) Entity label Classification for Cancer Genetics

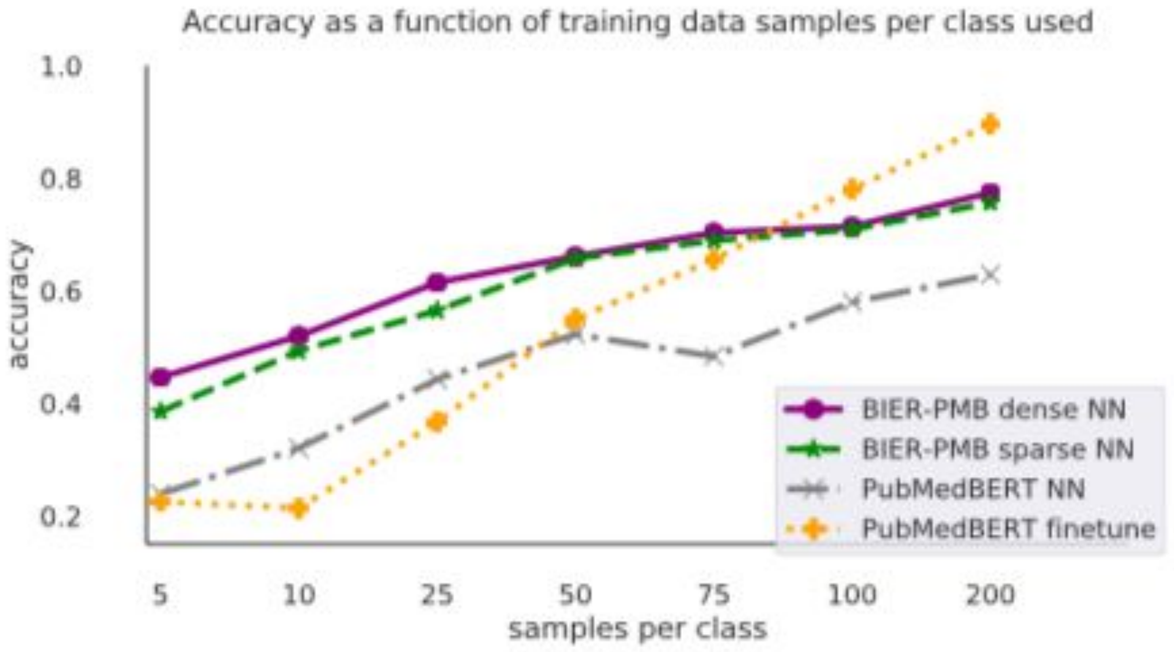


Figure 3: Results for the entity label classification task under varying amounts of supervision.

Summary of Contributions

Developed a **Biomedical Interpretable Entity Representations (BIERs) model**

Using training data (37 million)& a 68K biomed entity type system
obtained via a **novel distant supervision method linking PubMed to Wikipedia**

Empirically **BIERs outperforms the prior IERs work** on various biomedical tasks

Showed **BIERs outperforms Dense non-interpretable models when the supervision available is limited** (75 samples per class)

Propose an **oracle technique** using both the dense and sparse embeddings from a BIER model **to improve task performance** and **motivate the use of confidence measures for discovering when to inspect test cases**.

TEXAS The University of Texas at Austin

Completed Works - Post Proposal

Post Proposal Works

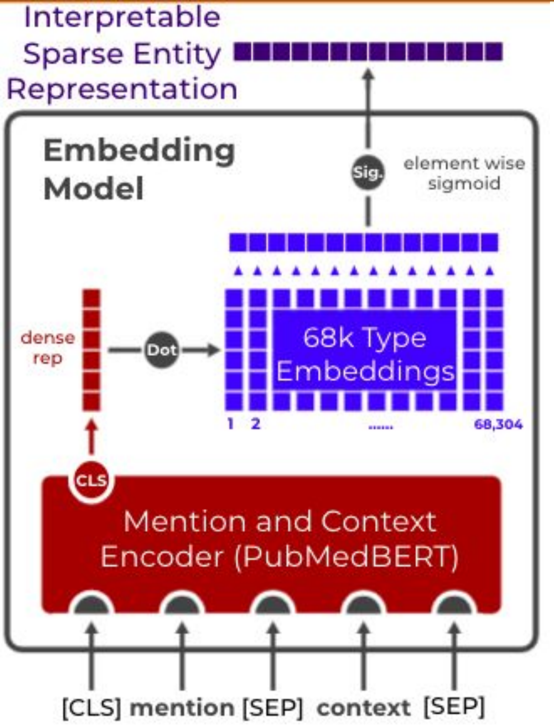
- Intermediate Entity-based Sparse Interpretable Representation Learning. *under submission*

- Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection *(WWW 22. Multimodal Understanding for the Web and Social Media workshop)*

Intermediate Entity-based Sparse Interpretable Representation Learning

Garcia-Olano, D., Onoe, Y., Wallace, B., Ghosh, J., “Intermediate Entity-based Sparse Interpretable Representation Learning”. Under Submission

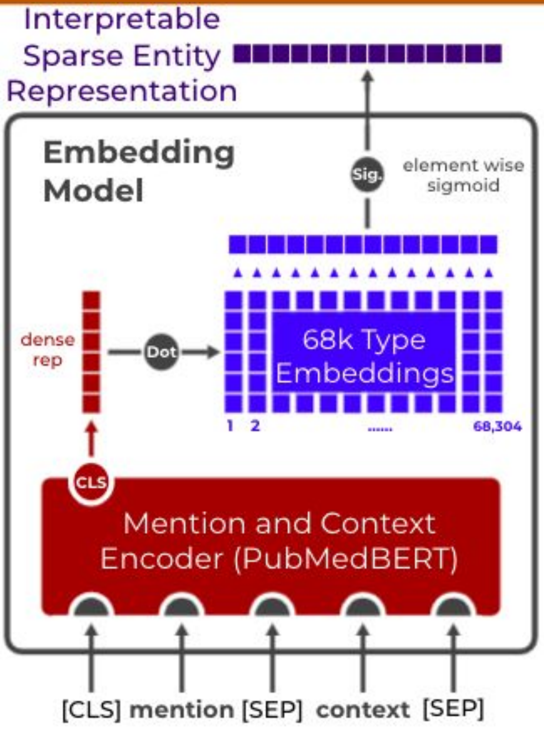
Recall IERs/BIERs



Pros

- Induce sparse embeddings that are human-readable, whose dimensions correspond to fine-grained entity types & values are predicted probabilities that a given entity type component aligns with an entity/context
- Perform well in zero-shot & low supervision settings.
- Compared with standard dense embeddings, these interpretable representations permit unique, fine-grained model analysis & debugging

Recall IERs/BIERs



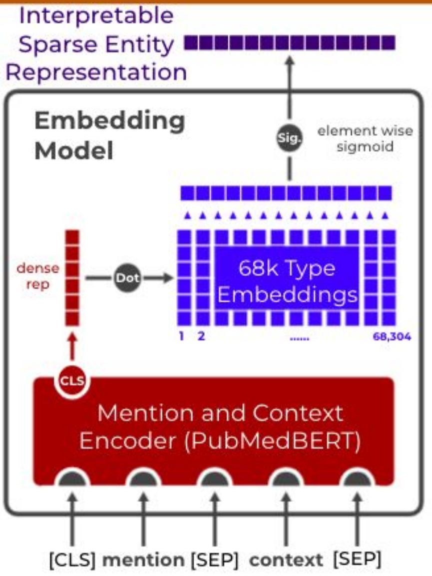
Pros (prior slide)

- Sparse human readable entity type embeddings
- Perform well in zero-shot & low supervision settings.
- Unique, fine-grained model analysis & debugging

Cons

- Lower accuracy on tasks with lots of training data
- **Fine-tuning** these representations improves accuracy on downstream tasks, but **destroys the semantics** of the dimensions as enforced in pre-training

Motivation



Pros (prior slide)

- Sparse human readable entity type embeddings
- Perform well in zero-shot & low supervision settings.
- Unique, fine-grained model analysis & debugging

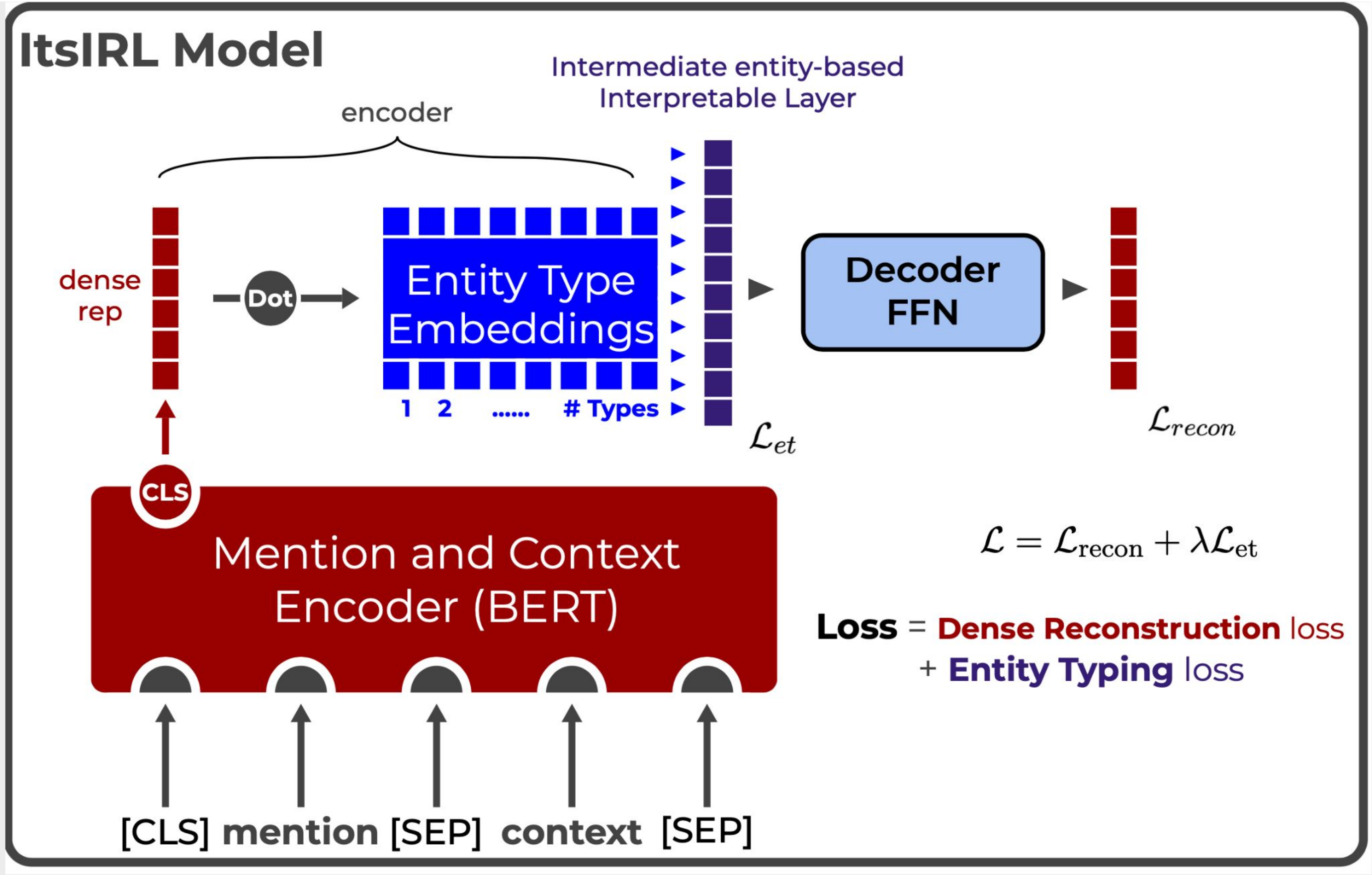
Cons

- Lower accuracy on tasks with lots of training data
- **Fine-tuning** improves accuracy on downstream tasks, but **destroys the semantics** of the dimensions

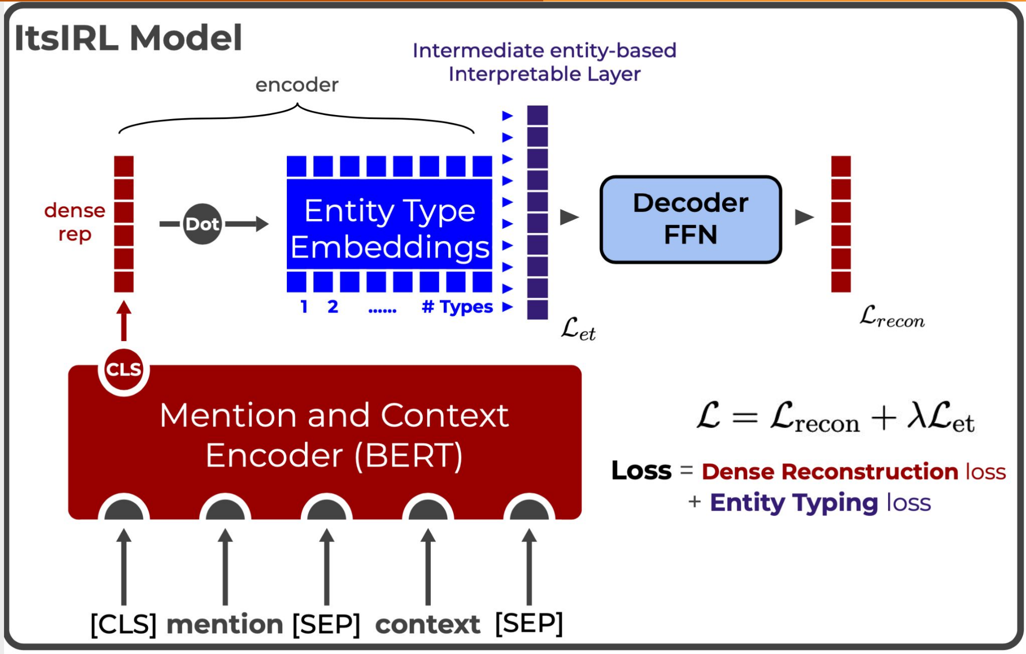
Motivating Question:

Can we maintain the interpretable semantics afforded by (B)IERs while improving predictive performance on downstream tasks?

Intermediate en**T**ity-based **s**parse **I**nterpretable **R**epresentation **L**earning



ItsIRL



We **pre-train** a encoder/decoder with a sparse and **interpretable**, **high dimensional latent space** and rich dense output representations.

The encoder induces a sparse embedding of entity types as in prior work on IERs,

but now **for downstream tasks** we can **freeze the encoder** (which yields interpretable entity representations) & **fine-tune the decoder.**

Empirically over two biomedical tasks
we show our model gives both

- **interpretable entity types** and
- **improved task performance.**

We propose two novel methods to study the model’s :

- interpretability via class-based **global prototypes** over entity types
- debugging ability via automated **entity type manipulation**

ItsIRL - Task 1

Cancer Genetics Classification [Pyysalo et al., 2013]
Data: ~11K training, 3.5k dev, & 7k test examples from PubMed articles
Task: Given a title/abstract & entity mention, **classify** the entity as one of 16 classes

Model	🔍	Test Acc
BIER-PMB*	✓	87.5
ItsIRL	✓	91.9
ItsIRL E2E*	-	95.7
PubMedBERT	-	96.1

Table 1: Cancer Genetics results
🔍 = interpretable types
PMB* = PubMedBERT
E2E* = End-To-End fine-tuned

ItsIRL - Task 1

Cancer Genetics Classification [Pyysalo et al., 2013]
Data: ~11K training, 3.5k dev, & 7k test examples from PubMed articles
Task: Given a title/abstract & entity mention, **classify** the entity as one of 16 classes

Model	🔍	Test Acc
BIER-PMB*	✓	87.5
ItsIRL	✓	91.9
ItsIRL E2E*	-	95.7
PubMedBERT	-	96.1

Ablations	Test Acc
ItsIRL - random init	88.9
ItsIRL - 1 layer decoder	68.1

<- importance of pre-training decoder
<- importance of size & pre-training of decoder

Table 1: Cancer Genetics results
🔍 = interpretable types

* varying layer depths for our decoder (3, 5, 8) gives similar performance across.

ItsIRL - Task 2

BIOSSES - Sentence Similarity Estimation System for the Biomedical Domain
Data: 64 train, 16 dev & 20 test cases (pairs of PubMed sentences)
Task: Predict similarity score (regression) between two sentences

Model	🔍	MSE
BIER-PMB*	✓	5.05
ItsIRL	✓	1.59
ItsIRL E2E*	-	1.15
PubMedBERT	-	1.14

Table 2: BIOSSES sentence similarity regression results.
🔍 = interpretable types
PMB* = PubMedBERT
E2E* = End-To-End fine-tuned

ItsIRL - Task 2

BIOSSES - Sentence Similarity Estimation System for the Biomedical Domain
Data: 64 train, 16 dev & 20 test cases (pairs of PubMed sentences)
Task: Predict similarity score (regression) between two sentences

			Type Sparsity		
Model	🔍	MSE	@.01	@.1	@.25
BIER-PMB*	✓	5.05	-	-	-
ItsIRL	✓	1.59	33.6	8.1	4.4
ItsIRL E2E*	-	1.15	5723	780	330
PubMedBERT	-	1.14	-	-	-

Table 2: BIOSSES sentence similarity regression results.
🔍 = interpretable types
PMB* = PubMedBERT
E2E* = End-To-End fine-tuned

Sparsity of Entity Type Layer at varying weight thresholds

← Sparsity of Interpretable layer
← Sparsity of Un-interpretable layer

ItsIRL - Global Prototypes

Positive class prototypes

1) Run the decoder fine-tuned model over the task training data.
2) Gather all correctly predicted instances for each class,
   sum their interpretable entity type layer representations & normalize them

$$\text{Positive class prototype} = \frac{\text{vec} - \min(\text{vec})}{\max(\text{vec}) - \min(\text{vec})}$$

vec is the sum of entity type layers for a given class.

ItsIRL - Global Prototypes

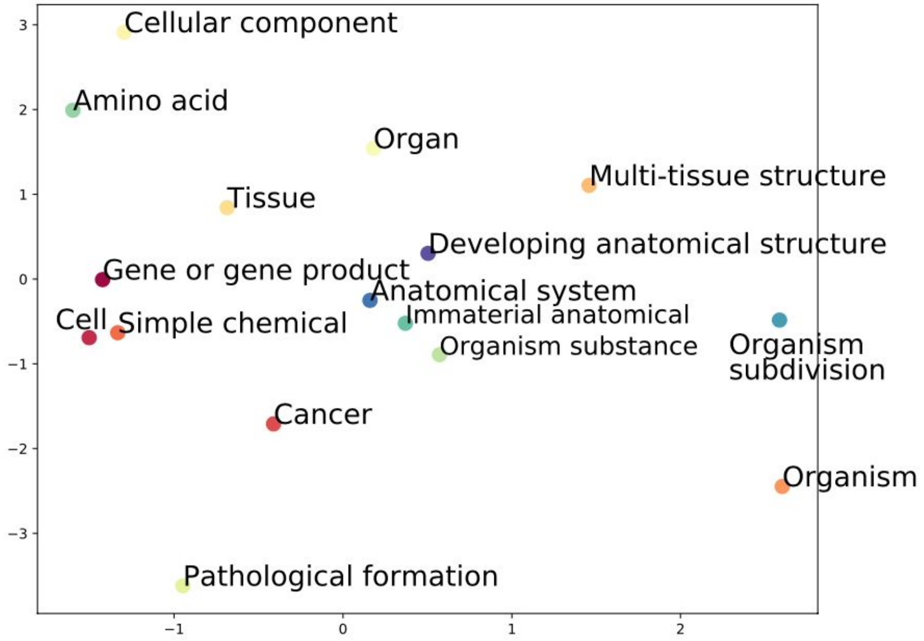
Positive class prototypes

1) Run the decoder fine-tuned model over the task training data.
2) Gather all correctly predicted instances for each class, sum their interpretable entity type layer representations & normalize them

$$\text{Positive class prototype} = \frac{\text{vec} - \min(\text{vec})}{\max(\text{vec}) - \min(\text{vec})}$$

vec is the sum of entity type layers for a given class.

Positive Class Prototypes in 2D via PacMap



ItsIRL - Global Prototypes

	Gene or gene product	Cell	Cancer	Simple chemical	Organism	Multi-tissue structure	Tissue
1	protein	cell	disease	ingredient	taxonomy	blood	tissue
2	ingredient	elementary particle	neoplasm	acid	mammals in 1758	angiology	cell
3	human	human cells	oncology	rtt	humans	soft tissue	human body
4	gene	battery	tissue	who essential medicines	tool-using mammals	nephron	connective tissue
5	coagulation	gene	abnormality	chemical compound	anatomically modern humans	blood vessel	endocrine system
6	cell	protein	cancer	measurement	postmodernism	human body	epithelium
7	cell growth	pancreas	syndrome	calcium	patient	lymphatic sys	angiology
8	endothelium	system	malignancy	hydroxyl	medical term.	lymphoid org.	blood vessel
9	homology	carboxylic acid	cell growth	glucose	prothrombin time	mononuclear phagocyte sys	histology
10	oncogene	ester	paraneoplastic syndromes	methyl group	bbc	gland	barcode

Table 3: Top Entity Types for 7 most frequent positive Prototype class embeddings

TEXAS
The University of Texas at Austin

class prototypes

Entity type weight

Entity type index

Gene or gene product	Cell	Cancer	Simple chemical	Organism	Multi-tissue structure	Tissue
protein (1.0, 5)	cell (biology) (1.0, 3)	disease (1.0, 2)	ingredient (1.0, 1)	taxonomy (biology) (1.0, 45)	blood (1.0, 47)	tissue (biology) (1.0, 34)
ingredient (0.742, 1)	elementary particle (0.346, 314)	neoplasm (0.897, 8)	acid (0.304, 18)	mammals described in 1758 (0.943,169)	angiology (0.843, 857)	cell (biology) (0.878, 3)
human (0.729, 7)	human cells (0.201, 145)	oncology (0.684, 28)	rtt (0.301, 4)	humans (0.943, 187)	soft tissue (0.792, 3067)	human body (0.814, 30)
gene (0.679, 6)	battery (electricity) (0.192, 485)	tissue (biology) (0.646, 34)	world health organization essential medicines (0.269, 25)	tool-using mammals (0.943, 186)	nephron (0.761, 1951)	connective tissue (0.385, 937)
coagulation (0.361, 37)	gene (0.184, 6)	abnormality (behavior) (0.604, 56)	chemical compound (0.206, 14)	anatomically modern humans (0.943,188)	blood vessel (0.682, 327)	endocrine system (0.345, 482)
cell (biology) (0.353, 3)	protein (0.177, 5)	cancer (0.582, 9)	measurement (0.19, 12)	post-modernism (0.943, 177)	human body (0.538, 30)	epithelium (0.325, 144)
F1score - 96.29	90.71	92.73	90.24	94.10	81.65	74.94
Support - 2520	1054	925	727	543	303	190

ItsIRL - Global Prototypes

Negative prototypes

Gather all incorrectly predicted instances, group by true vs predicted class, sum entity type layers & normalize

Truth / Pred	Cell / Cancer	Chemical / Gene	Cell / Gene	Organism / Gene	Tissue / Multi-tissue	Gene / Chemical	Cancer / Cell
1	cancer (1.0, 9)	ingredient (1.0, 1)	gene (1.0, 6)	gene (1.0, 6)	histology (1.0, 391)	ingredient (1.0, 1)	cell (biology) (1.0, 3)
2	disease (0.87, 2)	protein (0.61, 5)	protein (0.65, 5)	protein (0.93, 5)	blood (0.96, 47)	acid (0.58, 18)	neoplasm (0.41, 8)
3	neoplasm (0.73, 8)	receptor (biochemistry) (0.53, 52)	human (0.50, 7)	human (0.65, 7)	blood vessel (0.96, 327)	chemical compound (0.53, 14)	disease (0.38, 2)
4	malignancy (0.66, 20)	gene (0.49, 6)	allele (0.34, 71)	allele (0.43, 71)	angiology (0.92, 857)	derivative (chemistry) (0.42, 58)	t cell (0.36, 429)
5	rtt (0.55, 4)	human (0.41, 7)	ingredient (0.28, 1)	apoptosis (0.37, 87)	nephron (0.74, 1951)	protein (0.34, 5)	lymphocyte (0.35, 112)
6	oncology (0.46, 28)	enzyme (0.34, 29)	receptor (biochemistry) (0.25, 52)	wild type (0.35, 159)	circulatory system (0.64, 664)	purine (0.32, 781)	cancer (0.25, 9)
7	squamous-cell carcinoma	blood (0.29, 47)	transcription factors (0.25, 219)	ingredient (0.34, 1)	tongue (0.58, 158)	deciduous teeth (0.28, 3292)	lymphoblast (0.25, 1200)

Entity Types for 7 most frequent negative Prototypes

ItsIRL - Type Manipulation

Entity Type manipulation study

1. Generate coarse sets of entity types for each class based on string matching

Class	Term Rules Inclusion/Exclusion	Terms in Set
Cell	[cell]	357
Cancer	[cancer, neoplasm]	155
Gene or gene product	[' gene', 'gene ', ' genes', 'genes '] , ' not in ['generation', 'general']	434
Simple chemical	[chemical, chemical]	80
Organism	[' organ', 'organ ', 'organism'] not in ['organization']	172

Table 6: Terms used to create coarse Class specific Entity Type sets

ItsIRL - Type Manipulation

Entity Type manipulation study

1. Generate coarse sets of entity types for each class based on string matching

2. **3 strategies for manipulating entity types** at inference time

- **“Fixing”** incorrect entity types
  reduce weights of types from incorrectly predicted class’s coarse type set

- **“Promoting”** true entity types
  increase weights of entity types associated with the true label’s type set

- **Both “Fixing”** incorrect types **And “Promoting”** true types

ItsIRL - Type Manipulation

Entity Type manipulation study

1. Generate coarse sets of entity types for each class based on string matching

2. **3 strategies for manipulating entity types** at inference time

- **“Fixing”** incorrect entity types
  reduce weights of types from incorrectly predicted class’s coarse type set

- **“Promoting”** true entity types
  increase weights of entity types associated with the true label’s type set

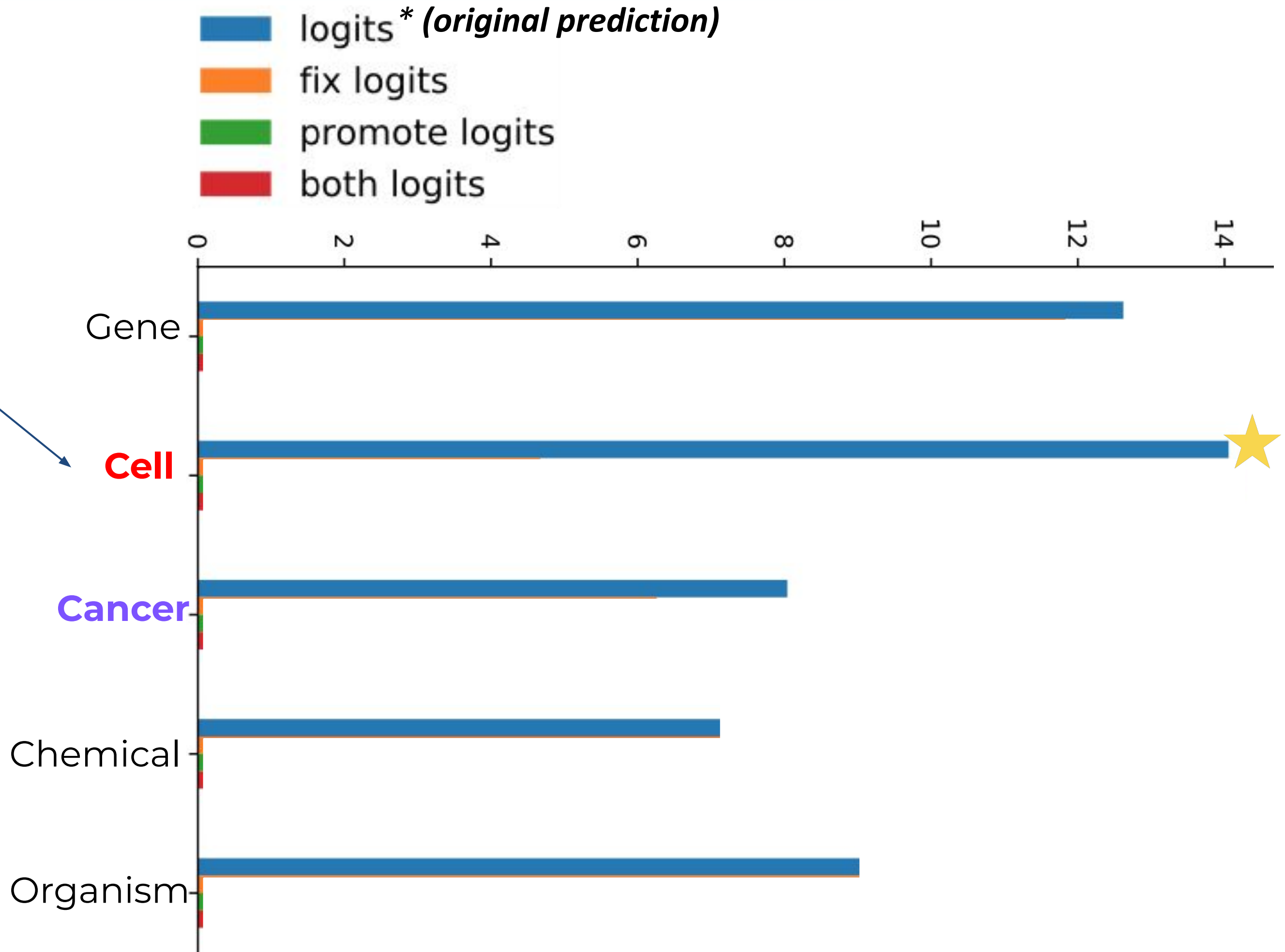
- **Both “Fixing”** incorrect types **And “Promoting”** true types

3. For each test error case, feed them through our model and run each of the 3 strategies on the corresponding entity type weights in the intermediate entity types layer & observe final class probabilities.

ItsIRL - Type Manipulation

Example on
Single Error Case

True Label: Cancer
Predicted: Cell



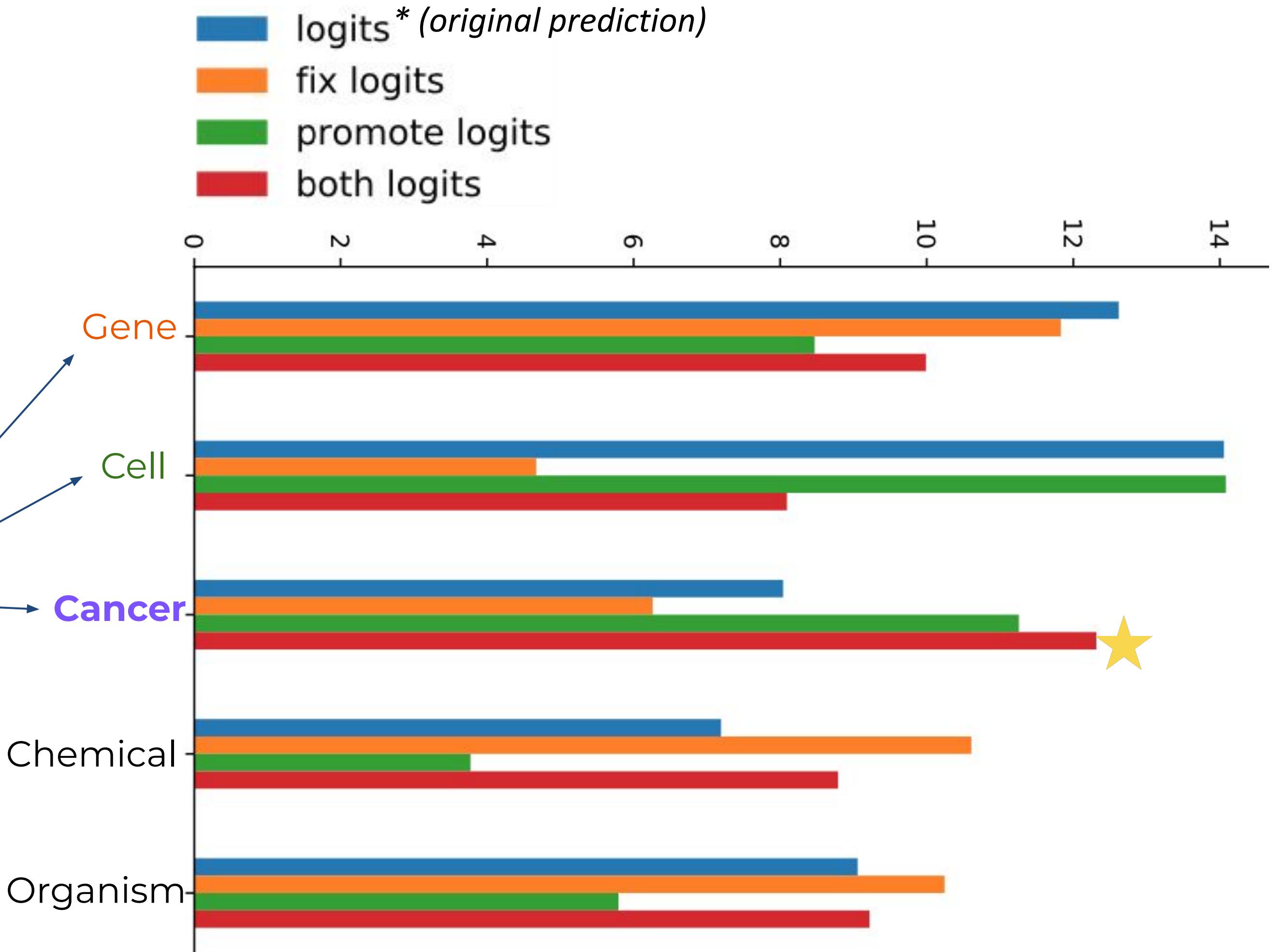
ItsIRL - Type Manipulation

Example on
Single Error Case

True Label: Cancer
Predicted: Cell

Results after Manipulation Techniques

1. **Fixing** -> Gene
2. **Promote** -> Cell
3. **Both** -> Cancer



ItsIRL - Type Manipulation

Results:

Model	Test Accuracy
ItsIRL	91.48
+ Fix types	93.91
+ Promote types	95.74
+ Both fix & promote	95.68
+ Best of 3 approach	**96.78**
PubMedBERT*	96.10

Table 4: Entity type manipulation results using Class Coarse sets to approximate non-expert

Summary of Contributions

- Intermediate enTity-based Sparse Interpretable Representation Learning (ItsIRL) an **extension to the IERs** architecture provides an intermediate interpretable layer and decoder that can be **fine-tuned for improved performance on downstream tasks.**

- **ItsIRL outperforms prior IER methods** and is competitive with uninterpretable dense language models on two biomedical tasks.

- Propose **entity type manipulation analysis** which facilitates **model understanding and debugging in an automated fashion** with even minimal, noisy supervision.

- Show how combining entity types over classes on the training set to create **positive and negative class prototypes** can be used to reveal task specific **global structure and semantics learned by our model**.

Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection

Garcia-Olano, D., Onoe, Y., Ghosh, J., "Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection". Proceedings of WWW 22 conference. Workshop on Multimodal Understanding for the Web and Social Media.

Knowledge Based VQA

Question: How many of them were born in the USA?

Image Caption: *Barack Obama* and his wife *Michelle* at the Civil Rights Summit at the LBJ Presidential Library, 2014.

Wikipedia Entities:
Barack_Obama Michelle_Obama

Question + Image Caption

Motivation

- VQA models are **expensive to pre-train** (many image, question pairs)
  Can **we improve upon their performance during fine-tuning?**

Motivation

- VQA models are **expensive to pre-train** (many image, question pairs)
  Can **we improve upon their performance during fine-tuning?**

- Quite a bit of work studying if **LMs can be used as knowledge bases**
  But less on **whether vision-language models can be?**

Motivation

- VQA models are **expensive to pre-train** (many image, question pairs)
  Can **we improve upon their performance during fine-tuning?**

- Quite a bit of work studying if **LMs can be used as knowledge bases**
  But less on **whether vision-language models can be?**

- Poerner et al 2020 show improved performance on entity-centric text tasks by using a simple, entity based, **knowledge injection technique into LMs**.
  Would this injection technique work as well for VQA models?

Motivation

- VQA models are **expensive to pre-train** (many image, question pairs)
  Can **we improve upon their performance during fine-tuning?**

- Quite a bit of work studying if **LMs can be used as knowledge bases**
  But less on **whether vision-language models can be?**

- Poerner et al 2020 show improved performance on entity-centric text tasks by using a simple, entity based, **knowledge injection technique into LMs**.
  Would this injection technique work as well for VQA models?

- Research on interpretability methods for single modalities is abundant,
  How would knowledge injection affect bi-modal explainability?

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT(Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow R^{d_{\text{Wikipedia}}}$

Background

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow R^{d_{\text{Wikipedia}}}$

E-BERT aligns **Wikipedia2Vec entity embeddings**
to BERT's wordpiece vector space
for entities found in task text inputs

Background

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow R^{d_{\text{Wikipedia}}}$

E-BERT aligns **Wikipedia2Vec entity embeddings**
to BERT's wordpiece vector space
for entities found in task text inputs

Learn map W during training

$$\sum_{x \in \mathbb{L}_{\text{WP}} \cap \mathbb{L}_{\text{Word}}} ||\mathbf{W}\mathcal{E}_{\text{Wikipedia}}(x) - \mathcal{E}_{\text{BERT}}(x)||_2^2$$

Background

Learn map **W** during training

$$\sum_{x \in \mathbb{L}_{\text{WP}} \cap \mathbb{L}_{\text{Word}}} ||\mathbf{W}\mathcal{E}_{\text{Wikipedia}}(x) - \mathcal{E}_{\text{BERT}}(x)||_2^2$$

At Inference map Wiki ents to BERT via **W**

$$\mathcal{E}_{\text{E-BERT}} : \mathbb{L}_{\text{Ent}} \rightarrow \mathbb{R}^{d_{\text{BERT}}}$$

$$\mathcal{E}_{\text{E-BERT}}(a) = \mathbf{W}\mathcal{E}_{\text{Wikipedia}}(a)$$

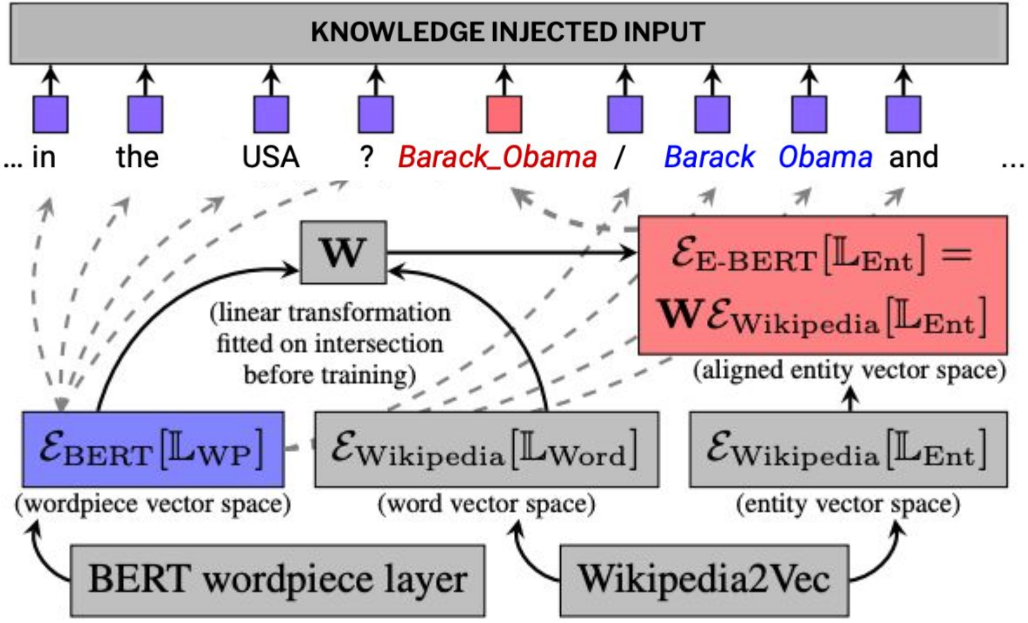


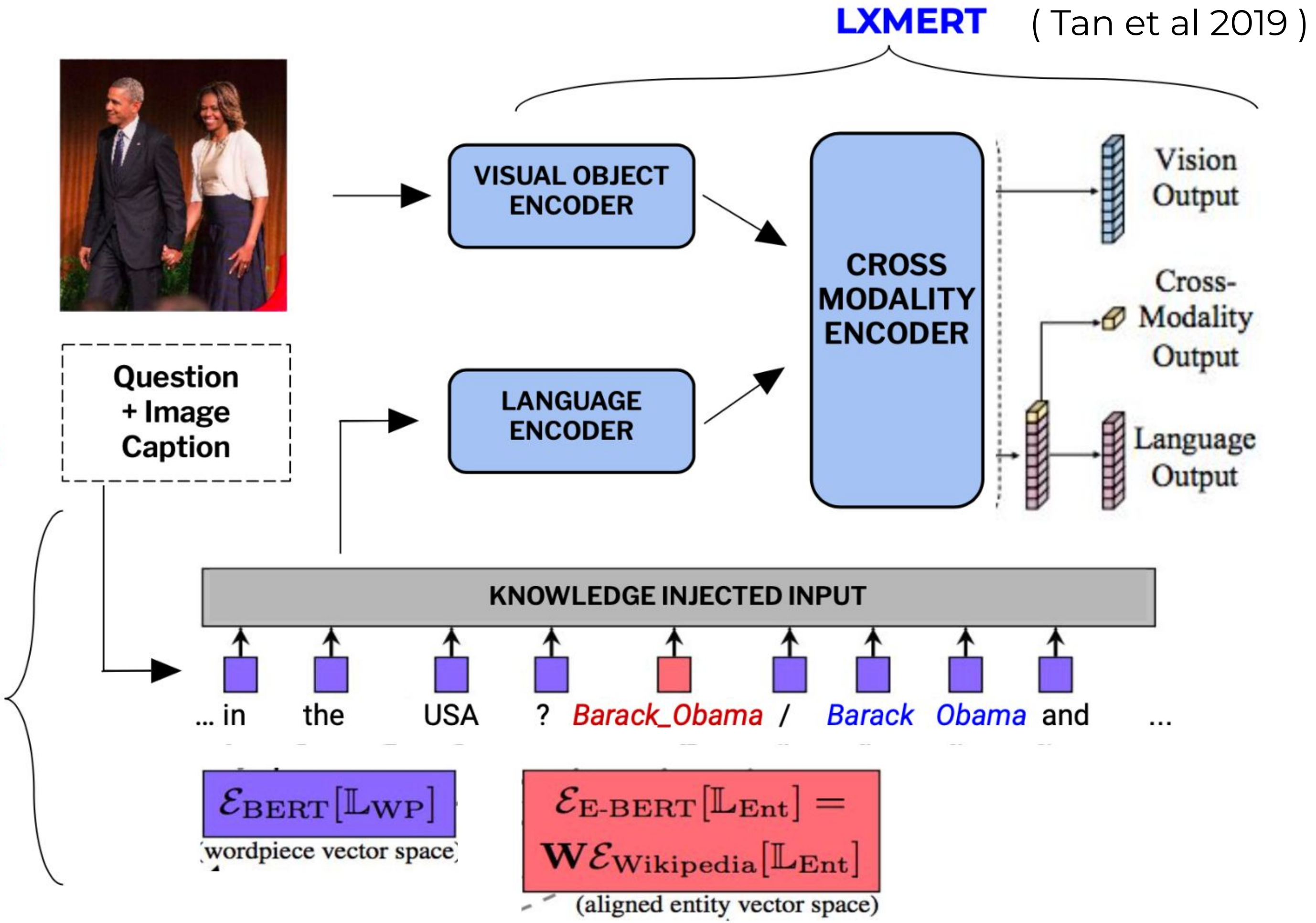
Figure 1: Schematic depiction of E-BERT-concat.

Our proposed architecture

Question: How many of them were born in the USA?

Image Caption: *Barack Obama* and his wife *Michelle* at the Civil Rights Summit at the LBJ Presidential Library, 2014.

Wikipedia Entities:
Barack_Obama Michelle_Obama



E-BERT concat

(Poerner et al 2020)

Datasets

KVQA (Sanket Shah, et al. AAAI 19)

- 24K images with text captions of politicians, actors, athletes, etc
- 183K image/question QA pairs (~ 7 questions per image)
- Metadata for the 18.8K unique Wikipedia entities
- Rare entities (only 65% exist in top million most occurring Wiki entities)

OKVQA (Marino, et al. CVPR 19)

- 14k image/question pairs for commonsense reasoning tasks (fewer entities)
- 10 human generated answers per questions while KVQA only has 1

Datasets

Entity span construction

KVQA	OKVQA
1) Question only (no spans)	1) Question only (no spans)
2) Question + Image Caption (no spans)	2) **13K** - no filtering to obtain entity spans for 13K QA pairs (92.8% of questions)
3) **NERper** - only entities of people	3) **4K** - semi-automated rules based technique to identify poor candidate spans which filters the set to 4K (28.6% of questions).
4) **NERagro** - all entities, no filtering	4) **2.5K** - manual filtering over unique entity spans to filter it down to 2.5K (17.8% of questions).
5) **KVQAmeta** - use metadata provided (less noise, more precise, only partial cover)	

KVQA Results

Table 1: KVQA overall accuracy results over 5 splits and entity spans per question (ents per Q), E-BERT representations injected per question (eberts per Q) and the percent of questions with E-BERT injections (Qs w/ eberts) for split 1

	Model	Type	Acc	ents per Q	eberts per Q	Qs w/ eberts
prior work	Shah 2019	-	49.50	-	-	-
	+ Caption	-	50.20	-	-	-
1.	Question	-	47.54	-	-	-
2.	+ Caption	-	50.25	-	-	-
3.	NERper	noisy	50.69	2.5	2.3	.94
4.	NERagro	noisy	50.77	3.3	**3.2**	.97
5.	KVQAmeta	noisy	**52.83**	1.4	1.4	**.99**

- Using E-BERT with entity spans from **KVQAMeta** gives 2.5 points higher accuracy. These spans are the closest to "gold spans" (quality over quantity) however there is still plenty of room for improvement.
- Multi-hop and multi-relationship questions improve by 6 & 5 points respectively (Table 3)
- The improvement for the lower quality derived entity spans (NERper and NERagro) still give .5 accuracy improvement.
- In all cases, more context can be gathered via retrieval mechanisms and E-BERT could be used on top of those results.

OKVQA Results

Table 2: OKVQA model results over 5 runs. * denotes models based on GPT-3 that are not directly comparable

Model	Mean	Std	Max	Median
OKVQA best	27.84	-	-	-
Shevchenko [29]	39.04	-	-	-
Wu et al [39]	40.50	-	-	-
PICA-Base (best) [41] *	43.3	-	-	-
PICA-Full (best) [41] *	48.0	-	-	-
LXMERT Plain	43.51	0.23	43.87	43.34
+ EBERT 13K	40.59	0.09	40.69	40.59
+ EBERT 4K	**43.67**	0.13	43.88	**43.66**
+ EBERT 2.5K	43.61	0.36	**44.10**	43.34

prior works

- Overall using E-BERT on LXMERT for OKVQA has much less effect since the data has very few, as a percentage, questions with entities and image captions (which are available externally from COCO) were not used
- Adding noisy entity spans (13K) hurts performance

Explainability Results

Table 4: KVQA Bi-modal (BM) and Transformer attention (TRF) explaination results for Questions where an E-BERT injected entity is in top 5 most important tokens.

Model	Type	BM ACC	BM Qs	TRF Acc	TRF Qs
Average		**59.74**	8.59	58.33	10.35

- For 7 out of 9 entity span set variations (NERper, NERagro, KVQAmeta), the questions which include E-BERT entities amongst their top 5 using BM-GAE provide better accuracy.
- This suggests that when using either method, an entity appearing in the top 5 most important tokens for a question/caption correlates with higher model accuracy (59.74 vs 51.04%) *

* Agrees with perturbation test results in Hila Chefer et al ICCV 2021.
"Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers."

Summary of Contributions

- We **analyzed how** efficient, entity based **knowledge injection via E-BERT** during fine tuning **affects** the performance of an existing model LXMERT on the task of **knowledge-based VQA** in terms of **accuracy & explainability**.

- We show substantial **improved accuracy on the entity rich KVQA dataset**, 2.5% top 1 acc, without the need to redo any costly pre-training.

- Model accuracy is **never harmed by knowledge injection on KVQA**, & only once for OKVQA, when the entity span set quality is very low.

- This work is **complementary to state of the art retrieval based methods** that gather additional context to improve VQA task performance since our method can be applied on top of those methods.

Future work

Future work

For the ItsIRL work,

1) **learning class entity type sets** in a data driven way and
(as opposed to the coarse string matching way we did in the paper)

2) **learning optimal error manipulation methods** for model debugging
(which technique: promote, fix or both works best for which error cases)

3) a nearest neighbor confidence measure approach
for **flagging test examples for inspection** that takes
a test case's entity type layer & matches it against
the entity type layers of positively predicted training examples

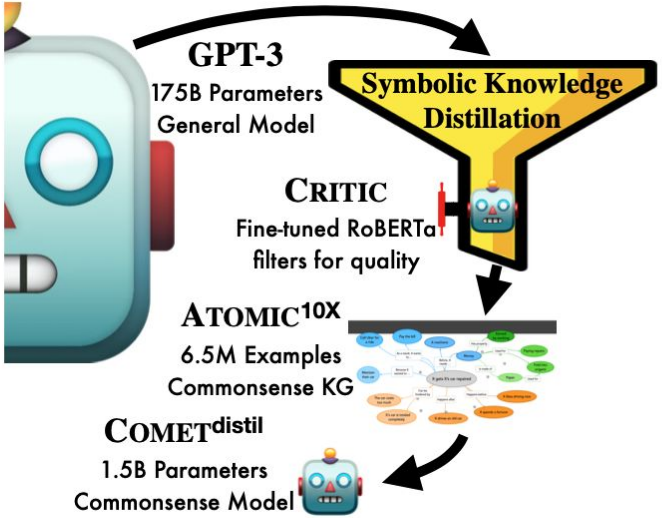
Future work

Application to **Large Language Models** (GPT3, Dall-E2, Imagen, etc)

Work around **prompting LLMs** and using smale-scale manual labeling to **learn in-process critic models** that filter & improve quality of generated texts.

- LLMs classifiers where high quality explanations are generated in-process (Wiegreffe., 2022)
- LLMs for automating knowledge base creation in commonsense reasoning (West, 2021).

- Extending to **different domains & use cases with in-process techniques**
- **Multi-modal setting** where a model could generate images that explain the behavior of the model as a whole



Summary

This dissertation argues in-process diagnostic techniques are useful for sequential data tasks both in accuracy & interpretability.

1. We constructed a **dual mention-entity encoder** that learns dense representations for efficient neural Entity Retrieval with an **in-process, iterative hard-negatives procedure** that can be inspected.

2. We adapted a **prototypical autoencoder** classifier to be compatible with **time series data**; allowing for **tunable prototype diversity** and improved **global and instance level explanations**. *(not shown)*

3. We learned a distantly supervised entity type system and data set for use in training a **Biomedical Interpretable Entity model** whose representations exist in a **semantically meaningful vector space** & whose **predictions may be diagnosed** with an oracle method.

Summary

4) Introduced the **ItsIRL** architecture that **extends BIERs** to allow for **task-centric fine tuning** on pre-trained representations without breaking the semantics of our learned entity type space.
We also proposed **two explainable diagnostic methods**, automated entity type manipulation & entity type based class prototypes, for **fine-grained model debugging** & **global model semantics interpretability**.

5) We analyzed how **efficient, entity based knowledge injection** via E-BERT during fine tuning affects an existing VQA model LXMERT on the task of **knowledge-based VQA** in terms of **accuracy & bi-modal explainability**.

List of Publications

- Garcia-Olano, D., Onoe, Y., Wallace, B., Ghosh, J. . "Intermediate Entity-based Sparse Interpretable Representation Learning" *under submission*

- Garcia-Olano, D., Onoe, Y., Ghosh, J., "Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection". Proceedings of WWW 22 conference. Workshop on Multimodal Understanding for the Web and Social Media.

- Garcia-Olano, D., Onoe, Y., Baldini, I., Ghosh, J., Wallace, B., Varshey, K. "Biomedical Interpretable Entity Representations". Findings of the Association for Computational Linguistics (ACL-IJCNLP), Bangkok, Thailand, 2021

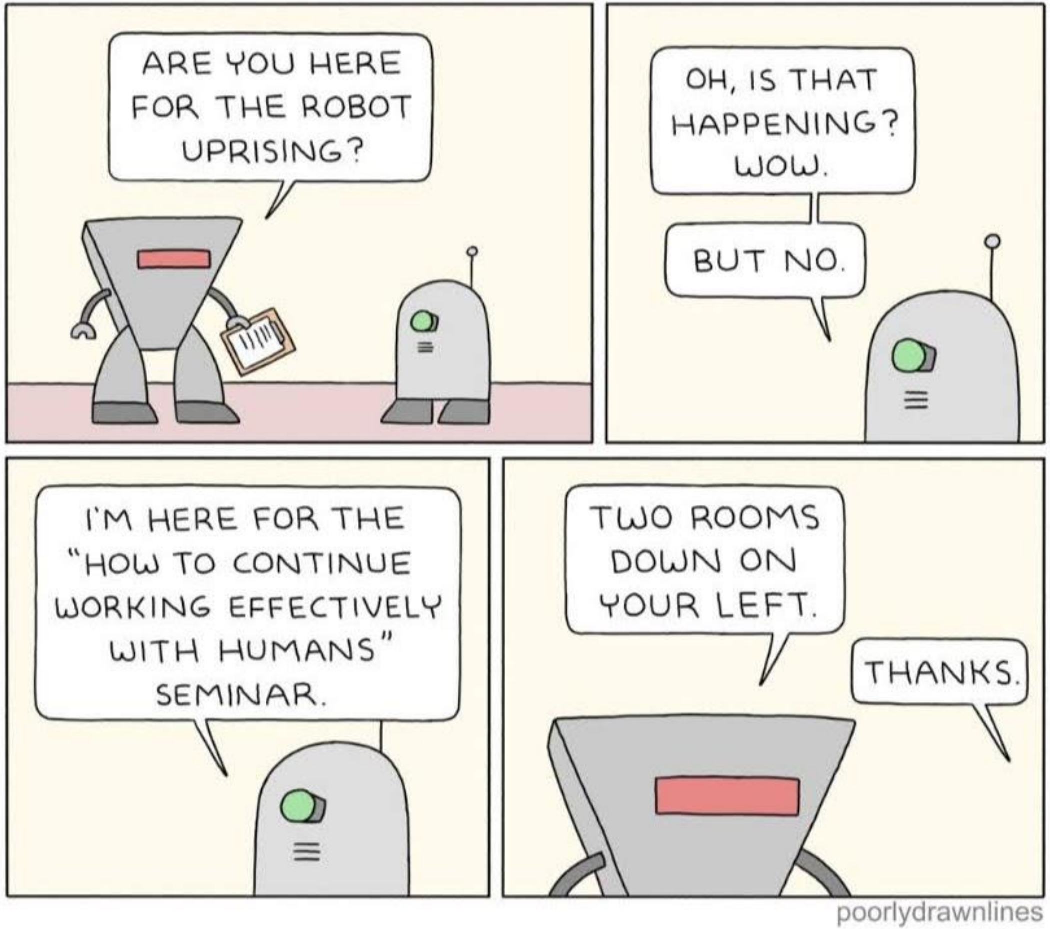
- Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, Eugene., Garcia-Olano, D. "Learning Dense Representations for Entity Retrieval". Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 2019.

- Garcia-Olano, D., Gee, A., Ghosh, J., Paydarfar, D. "Deep Classification of Time-Series Data with Learned Prototype Explanations". Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, California, PMLR 97, 2019

- Sankaran, K., Garcia-Olano, D., Javed, M., Alcala-Durand, M., De Unánue, A., van der Boor, P., Potash, E., Avalos, R., Encinas, L., Ghani, R., "Applying Machine Learning Methods to Enhance the Distribution of Social Services in Mexico". Presented at UChicago Data Science for Social Good. arXiv:1709.05551. 2017.

- Garcia-Olano, D. Arias, M, Larriba Pey, J. "Automated construction and analysis of political networks via open government and media sources". European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML). Riva del Garda, Italy, 2016

Thank you!



www.diegoolano.com
Twitter: @dgolano

BACKUP SLIDES

Prior State of the Art for Entity Resolution:

- Train on (Mention, Context, Entity) Triples.

2 Stages

(1) Retrieve Candidates

- Construct a Mention to Entities Lookup **"Alias" Table**.

  9.8 Million unique mention strings

  5.7 Million unique entities

(2) Re-Rank them

- **Limitations**

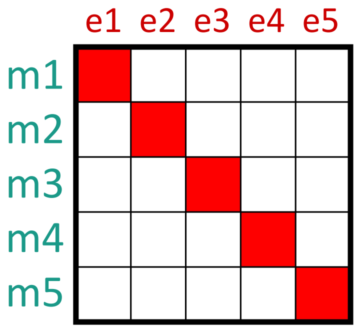
1) Low Recall
2) Context not considered. Can't predict unseen entities

Training with Sampled Softmax

The dual encoder learns a mention encoder φ and an entity encoder ψ,

where the **score** of a mention-entity pair (m, e) is:

$$\mathbf{s}(m, e) = \cos \mathrm{sim}(\ \varphi(m),\ \ \psi(e)\ )$$



These pairs constitute only positive examples,
so we use **in-batch random negatives** (Henderson et al., 2017;):

We build the all-pairs similarity matrix for all mentions & entities in a batch.
& **optimize a softmax loss** on each row of the matrix.

We do this **sampled softmax** (Jozefowicz et al, 2016)
in place of a full softmax
because the normalization term
is *intractable* to compute over all 5.7M entities.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Training with Sampled Softmax

For each training pair (m_i, e_i) in a batch of ***B*** pairs, the loss is computed as:

$$L(m_i, e_i) = -f(m_i, e_i) + \log \sum_{j=1}^{B} \exp(f(m_i, e_j))$$

$$\text{where} \quad f(m_i, e_j) = a \cdot s(m_i, e_j)$$

Iterative Hard Negative Sampling

Random negatives are not enough to train an accurate entity resolution model

So after learning an initial model using random negatives,
we propose to identify more challenging **"hard negatives"** via the following:

1. Encode all mentions and entities found in training pairs using current model.
2. For each mention, retrieve the most similar 10 entities (i.e., its nearest neighbors).
3. Select all entities ranked above the correct one as negative examples.

Iterative Hard Negative Sampling

Random negatives are not enough to train an accurate entity resolution model

So after learning an initial model using random negatives,
we propose to identify more challenging **"hard negatives"** via the following:

1. Encode all mentions and entities found in training pairs using current model.
2. For each mention, retrieve the most similar 10 entities (i.e., its nearest neighbors).
3. Select all entities ranked above the correct one as negative examples.

We merge these new hard negative mention/entity pairs
with the original positive pairs to construct an additional task
& resume training the dual encoder using logistic loss on them.

For a pair (m, e) with label y ∈ {0, 1}, the **hard negative loss** is defined as:

$$L_h(m, e; y) = -y \cdot \log f(m, e) - (1 - y) \cdot \log(1 - f(m, e))$$

$$\text{where} \quad f(m, e) = g(a_h \cdot s(m, e) + b_h)$$

Multi-task loss & Task Results

The hard negative task is mixed with the original random negatives task

$$L_{multi} = L_{orig} + L_{hard}$$

System	R@1	Entities
AT-Prior	71.9	5.7M
AT-Ext	73.3	5.7M
Chisholm and Hachey (2015)	80.7	800K
He et al. (2013)	81.0	1.5M
Sun et al. (2015)	83.9	818K
Yamada et al. (2016)	85.2	5.0M
Nie et al. (2018)	86.4	5.0M
Barrena et al. (2018)	87.3	523K
DEER (this work)	87.0	5.7M

Table 1: Comparison of relevant TACKBP-2010 results using Recall@1 (accuracy). While we cannot control the candidate entity set sizes, we attempt to approximate them here.

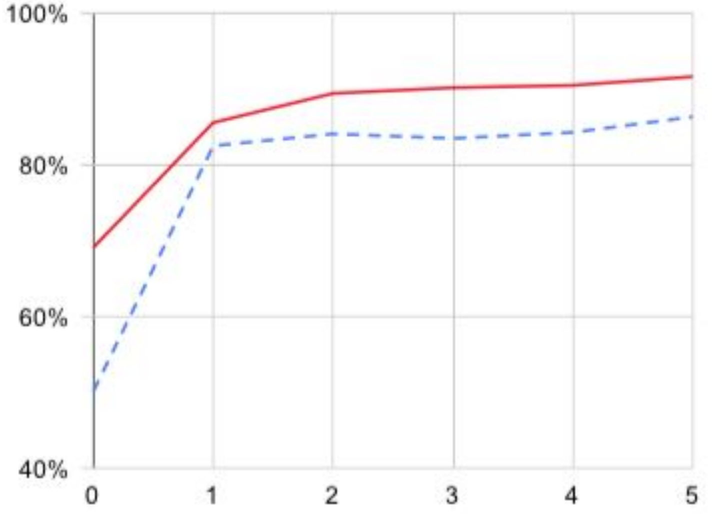


Figure 2: Recall@1 improvement for successive iterations of hard negative mining for Wikinews (solid) and TACKBP-2010 (dashed).

Retrieval at Scale

Inference is done by computing cosine similarity between
the test mention/context encoding and each of the cached entity encodings.

Approximate Search using quantization-based approaches (Guo et al. (2016))
can be used to speed up retrieval greatly!

Method	Mean search time (ms)	Wikinews R@100
Brute force	291.9	97.88
AH	22.6	97.22
AH+Tree	3.3	94.73

Table 3: Comparison of nearest-neighbor search methods using the DEER model. The benchmark was conducted on a single machine. AH indicates quantization-based asymmetric hashing; AH+Tree adds an initial tree search to further reduce the search space.

T-SNE visualization

Inspecting Entity Encodings for Semantic Meaning

Figure 3.5: t-SNE visualization of our learned embeddings for select country Wikipedia page entities. More at diegoolano.com/deer/

Hard Negative In-Process Explanations

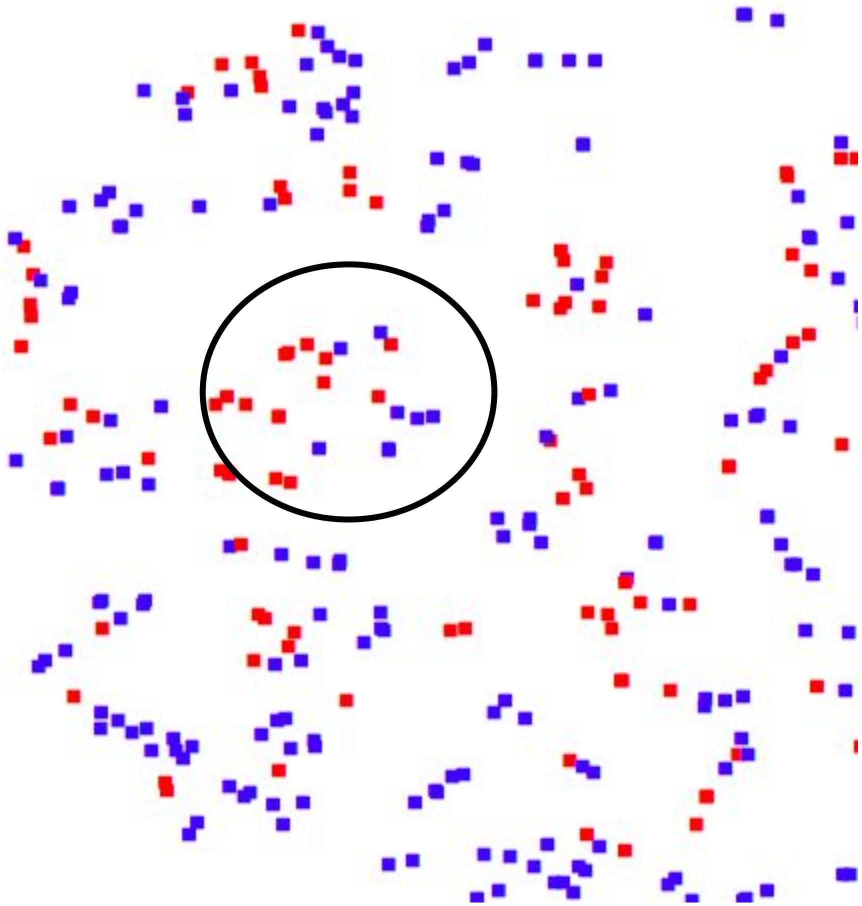
At inference time,
given a test mention/context,

1) Get K nearest mention/contexts from training set

2) Collectively assess how each of them performed over iterations (gather the hard negatives along with the true entities)

3) Get top entity prediction(s) for the test mention/context via cosine similarity

4) Utilize 2 and 3 results to calculate confidence measures for the final entity prediction

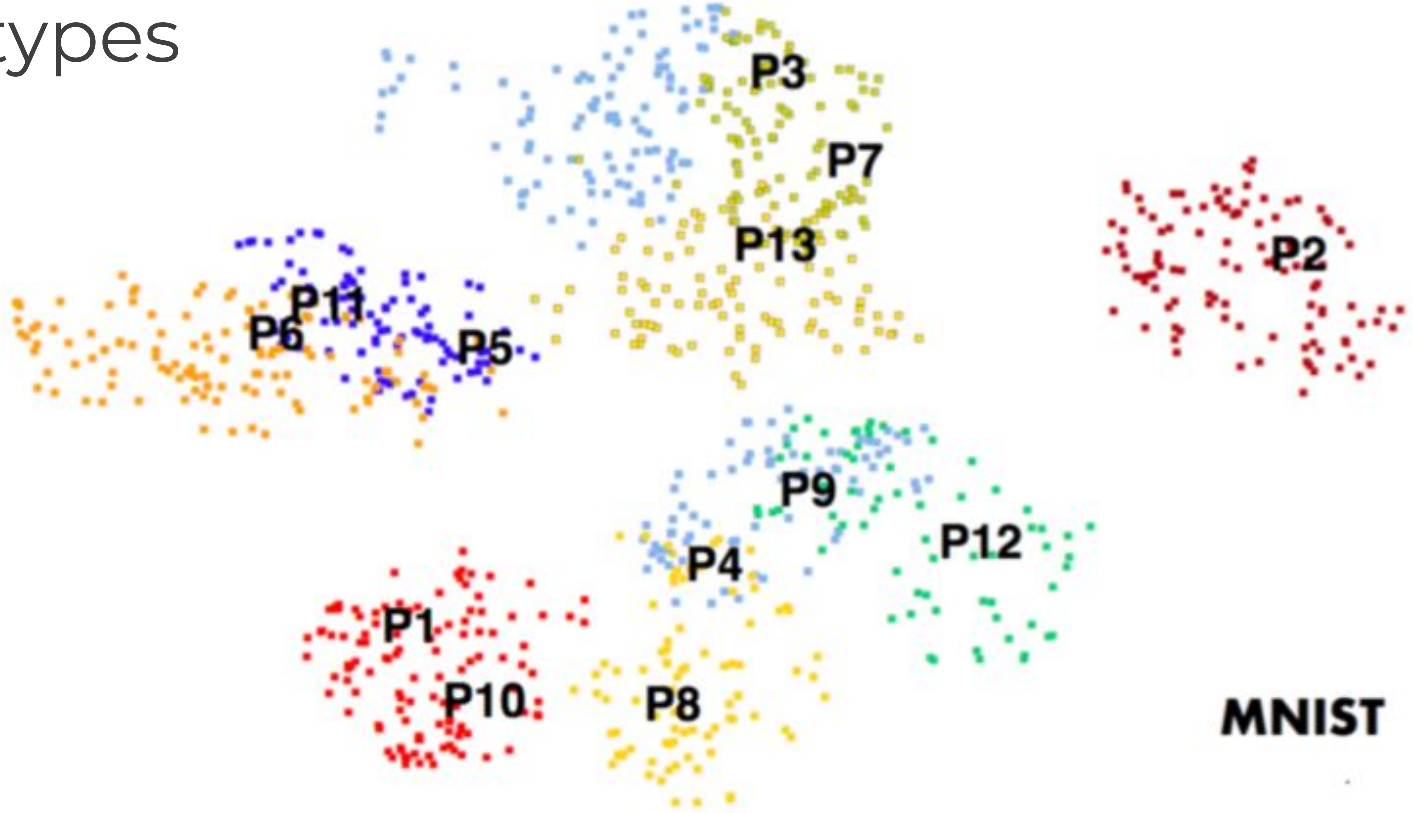


Explaining Deep Classification of Time-Series Data with Learned Prototypes

Garcia-Olano, D.*, Gee, A.*, Ghosh, J., Paydarfar, D. "Deep Classification of Time-Series Data with Learned Prototype Explanations". International Conference on Machine Learning (ICML 2019 time series workshop)

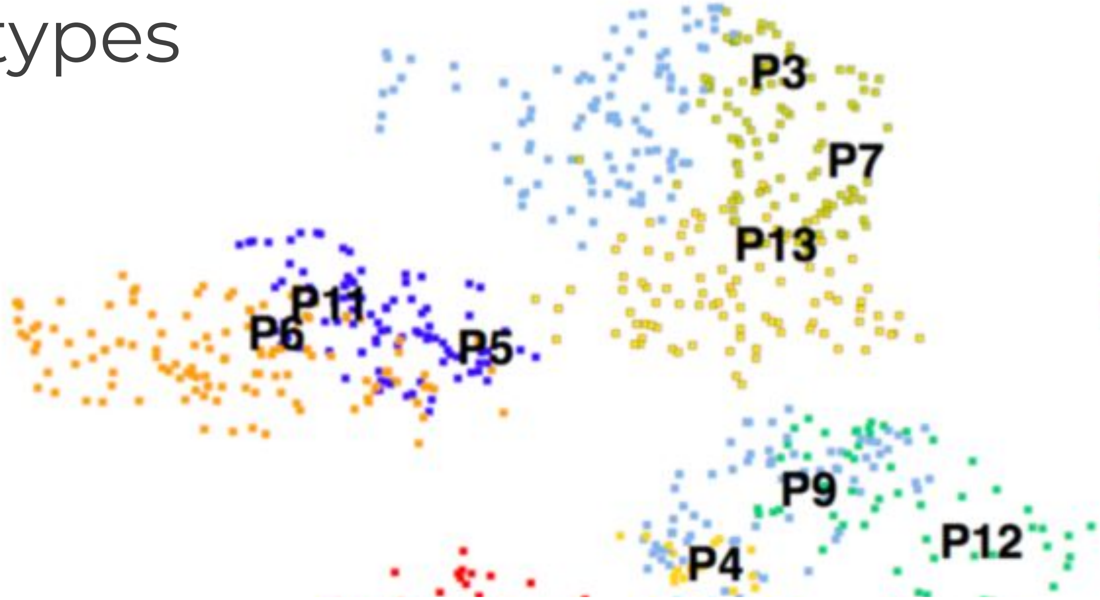
* *equal contribution*

Prototypes



*Li et al. Deep learning for case-based reasoning through prototypes. (2017)

Prototypes

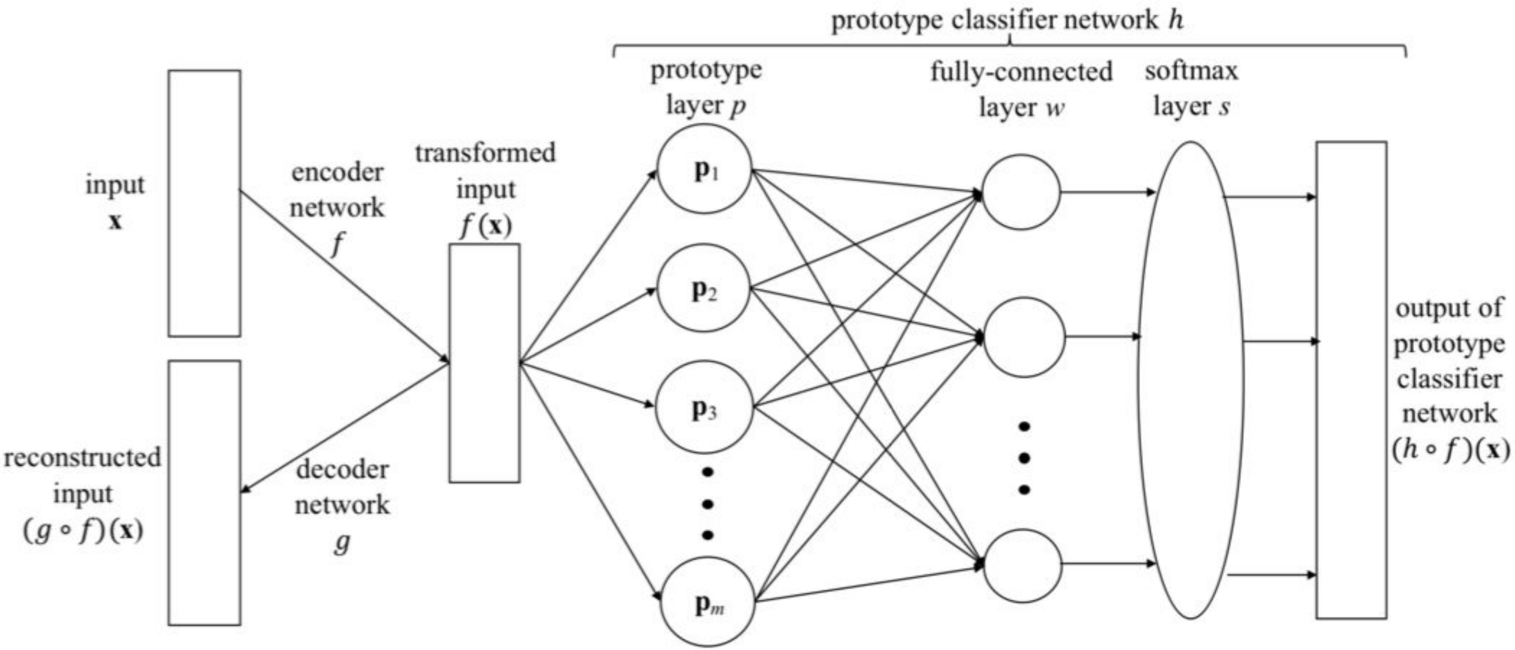


totypes. (2017)

Motivation

Prototype Classifier Network

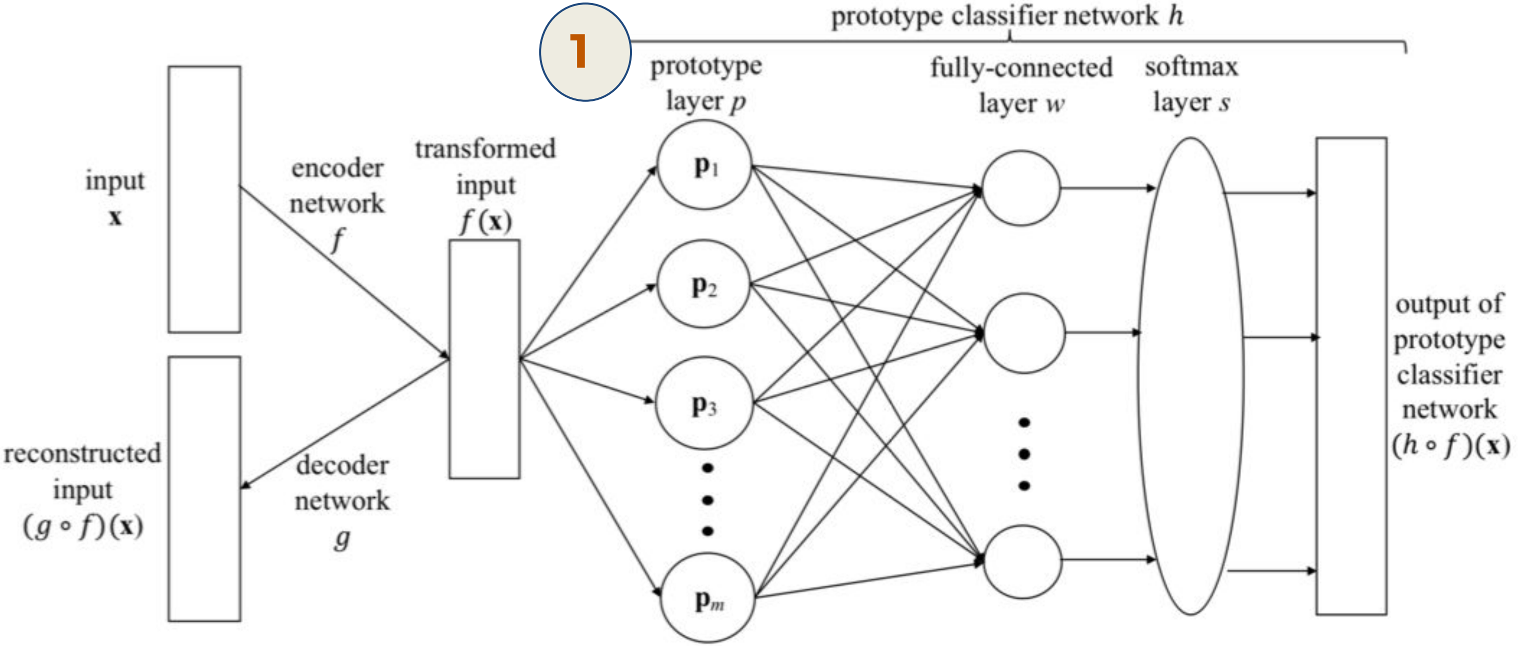
n data points
m prototypes



Motivation

Prototype Classifier Network

n data points
m prototypes

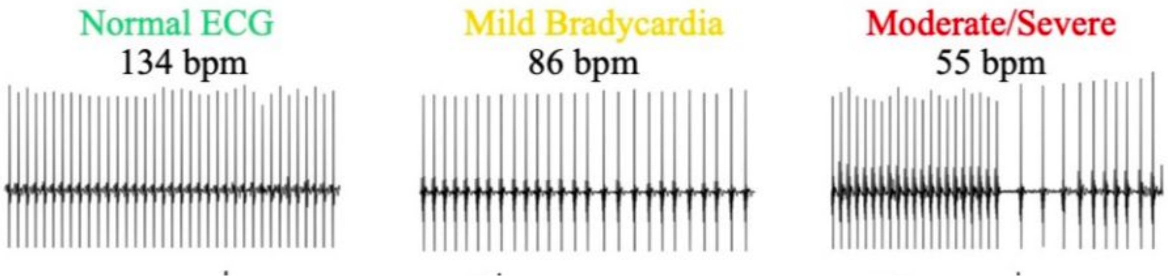


1

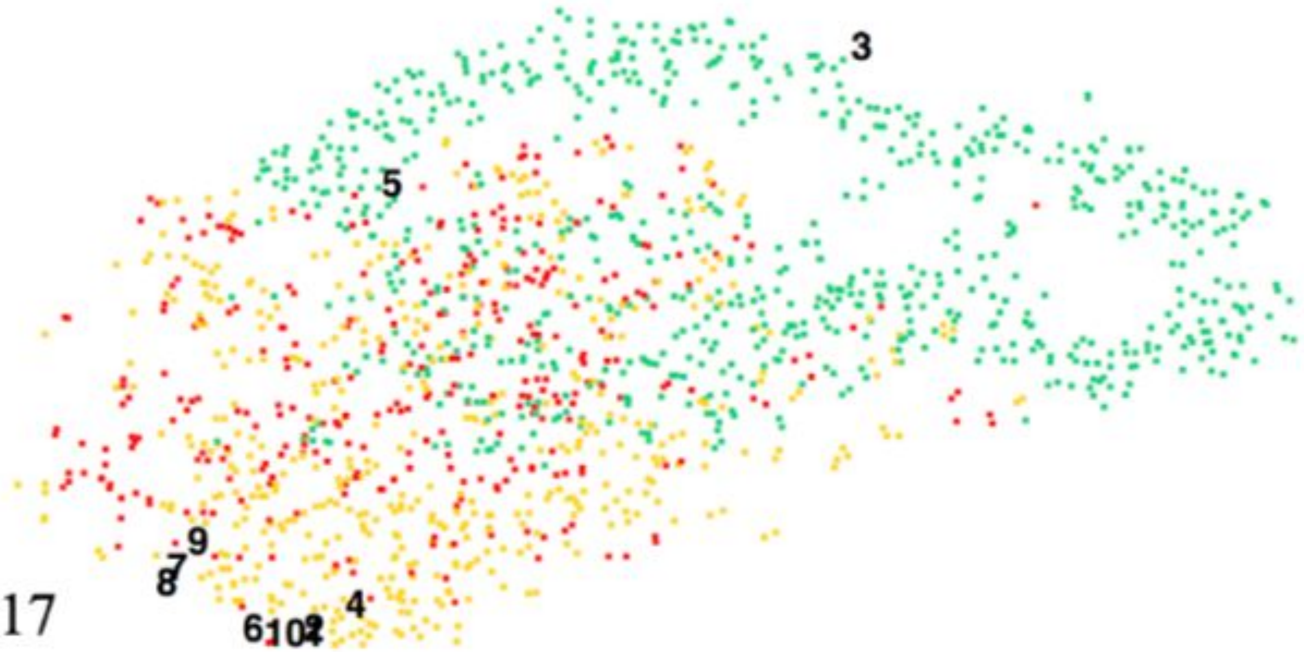
$$\underbrace{\text{Feature Vector } f(x) \in \mathbb{R}^q \qquad \text{Prototypes } p_i \in \mathbb{R}^q}_{\|f(x) - p_i\|_2^2 = \rho_i}$$

$$\mathcal{L}((f, g, h), X) = E(h \circ f, X) + \lambda_R R(g \circ f, X) + \lambda_1 R_1(p_1, \ldots, p_m, X) + \lambda_2 R_2(p_1, \ldots, p_m, X)$$

Predicting Bradycardia from ECG signals



Prior work
Latent Space Representation for Bradycardia task



Loss from
Li *et al.* 2017

Prototype Classifier Network Updated

$$\mathcal{L}((f,g,h),X) = E(h \circ f, X) + \lambda_R R(g \circ f, X) + \lambda_1 R_1(p_1, ..., p_m, X) + \lambda_2 R_2(p_1, ..., p_m, X) + \lambda_{pd} PDL(p_1, ..., p_m) \quad (2)$$

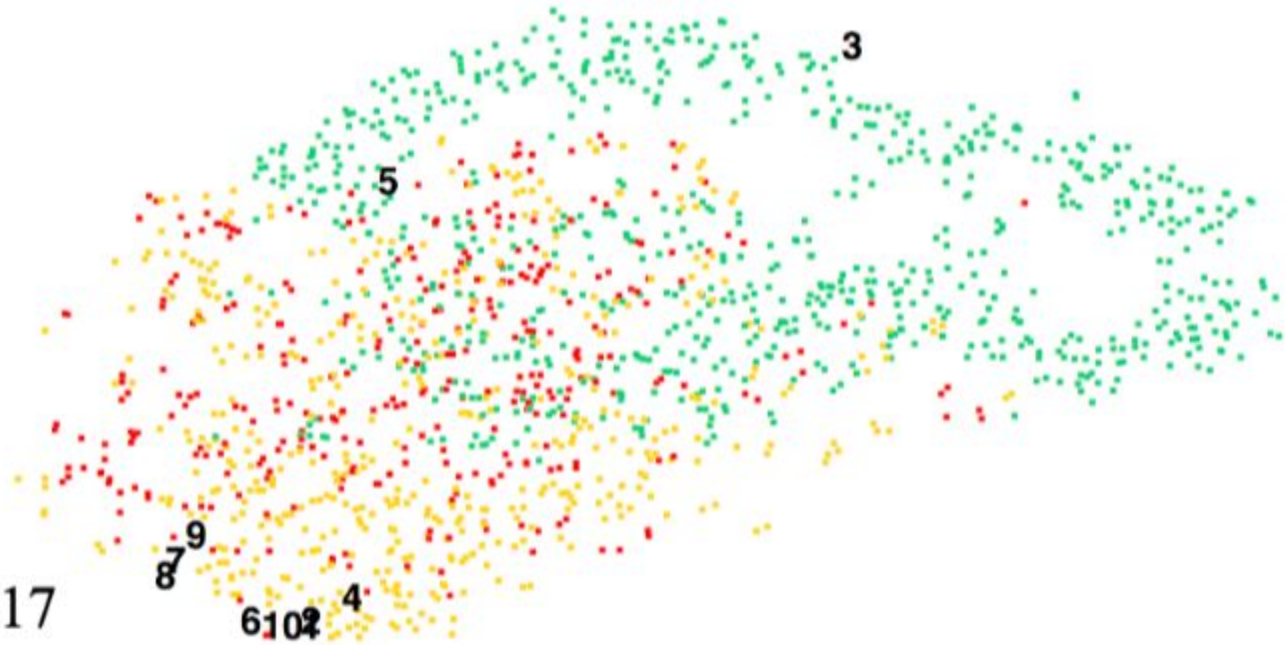
Prototype Diversity Loss

$$\lambda_{pd} PDL(p_1, ..., p_m) = \frac{1}{log(\frac{1}{m}\sum_{j=1}^{m} min_{i>j \in [1,m]} \|p_i - p_j\|_2^2) + \epsilon} \quad (1)$$

$$R_1(p_1, ..., p_m, X) = \frac{1}{m}\sum_{j=1}^{m} min_{i \in [1,n]} \|p_j - f(x_i)\|_2^2, \quad (3)$$

$$R_2(p_1, ..., p_m, X) = \frac{1}{n}\sum_{i=1}^{n} min_{j \in [1,m]} \|f(x_i) - p_j\|_2^2 \quad (4)$$

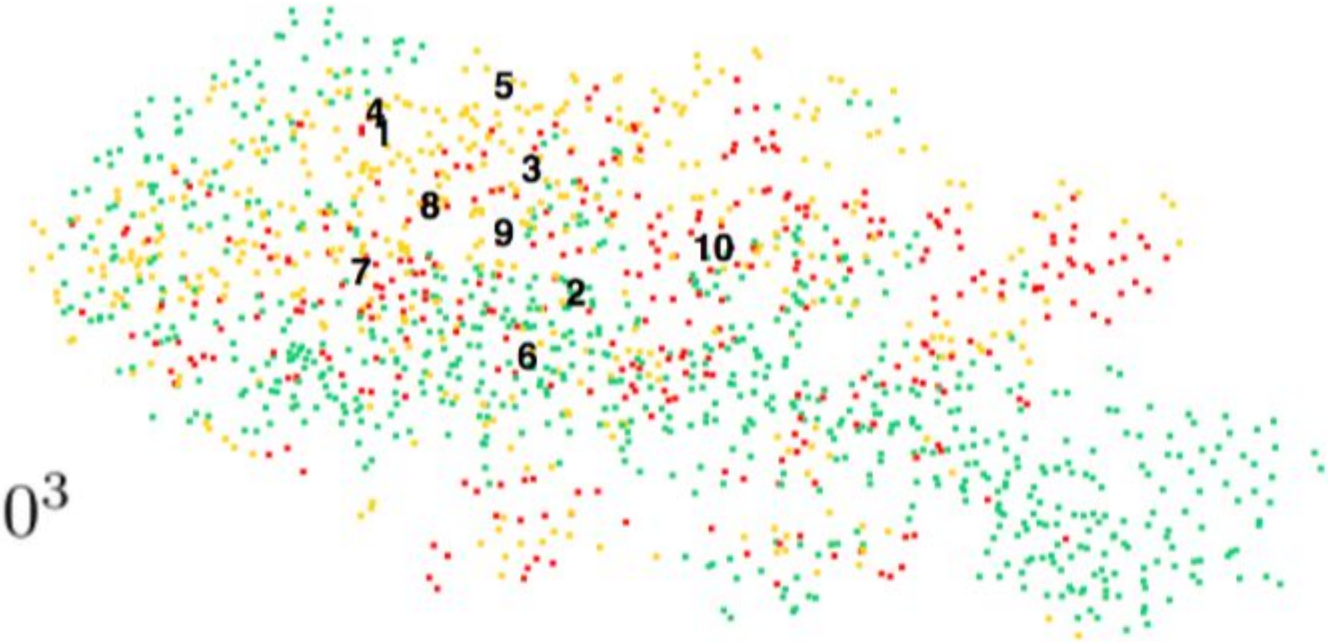
Prior work:
Latent Space Representation for Bradycardia task



Loss from
Li *et al.* 2017

$(\lambda_{pd} = 0)$

Our work:
Latent Space Representation for Bradycardia task



$$\lambda_{pd} = 10^3$$

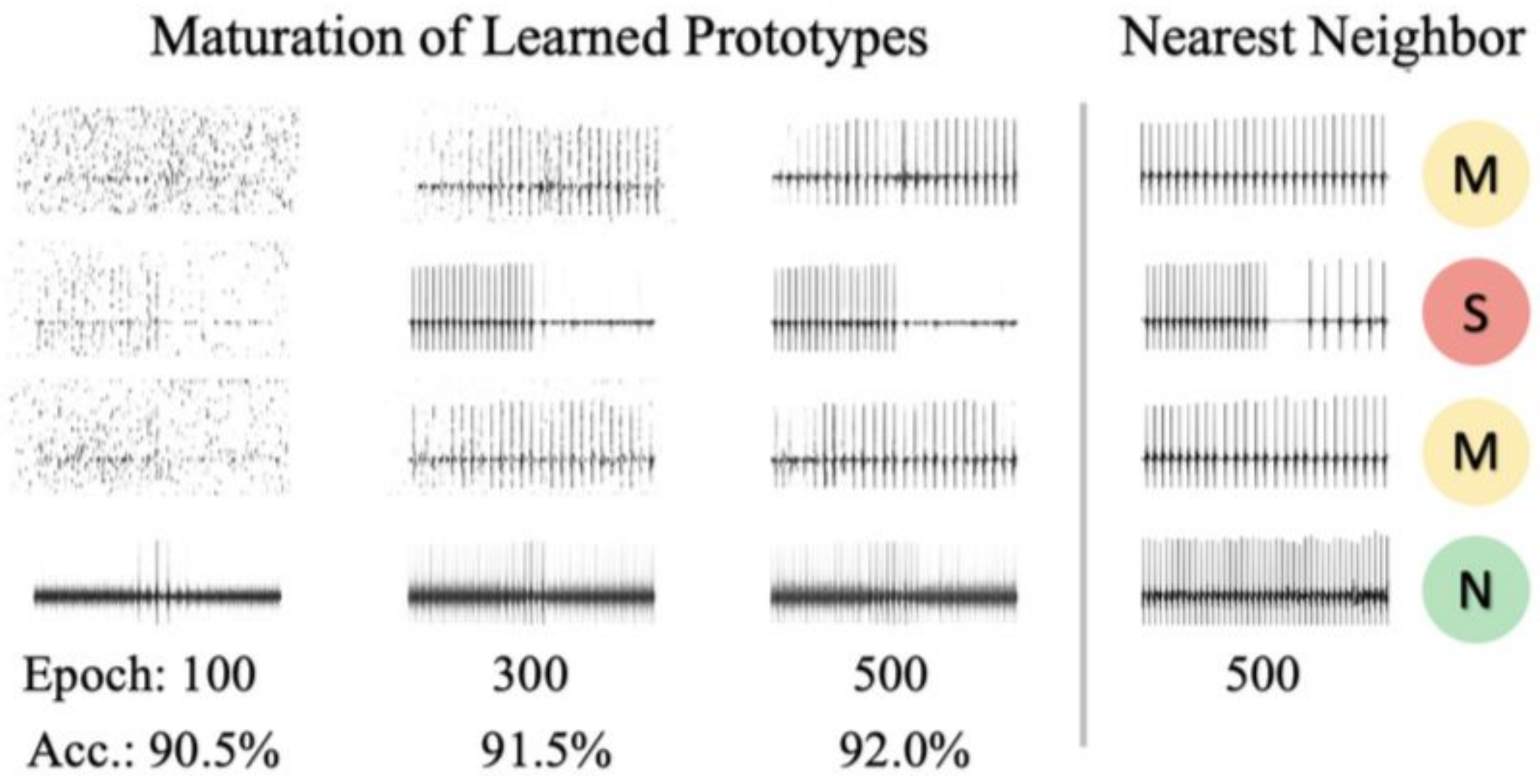
Results and Prototype Diversity

ECG: Bradycardia			
λ_{pd}	Accu.	Ψ_N	Ψ_C
0	$92.1 \pm 0.1\%$	0.83 ± 0.04	0.78 ± 0.19
500	$92.7 \pm 1.0\ \%$	0.86 ± 0.07	0.89 ± 0.19
$1e3$	$92.4 \pm 1.3\%$	0.87 ± 0.11	0.89 ± 0.19
$2e3$	**$93.1 \pm 0.4\%$**	**0.90 ± 0.04**	**1.00 ± 0.00**

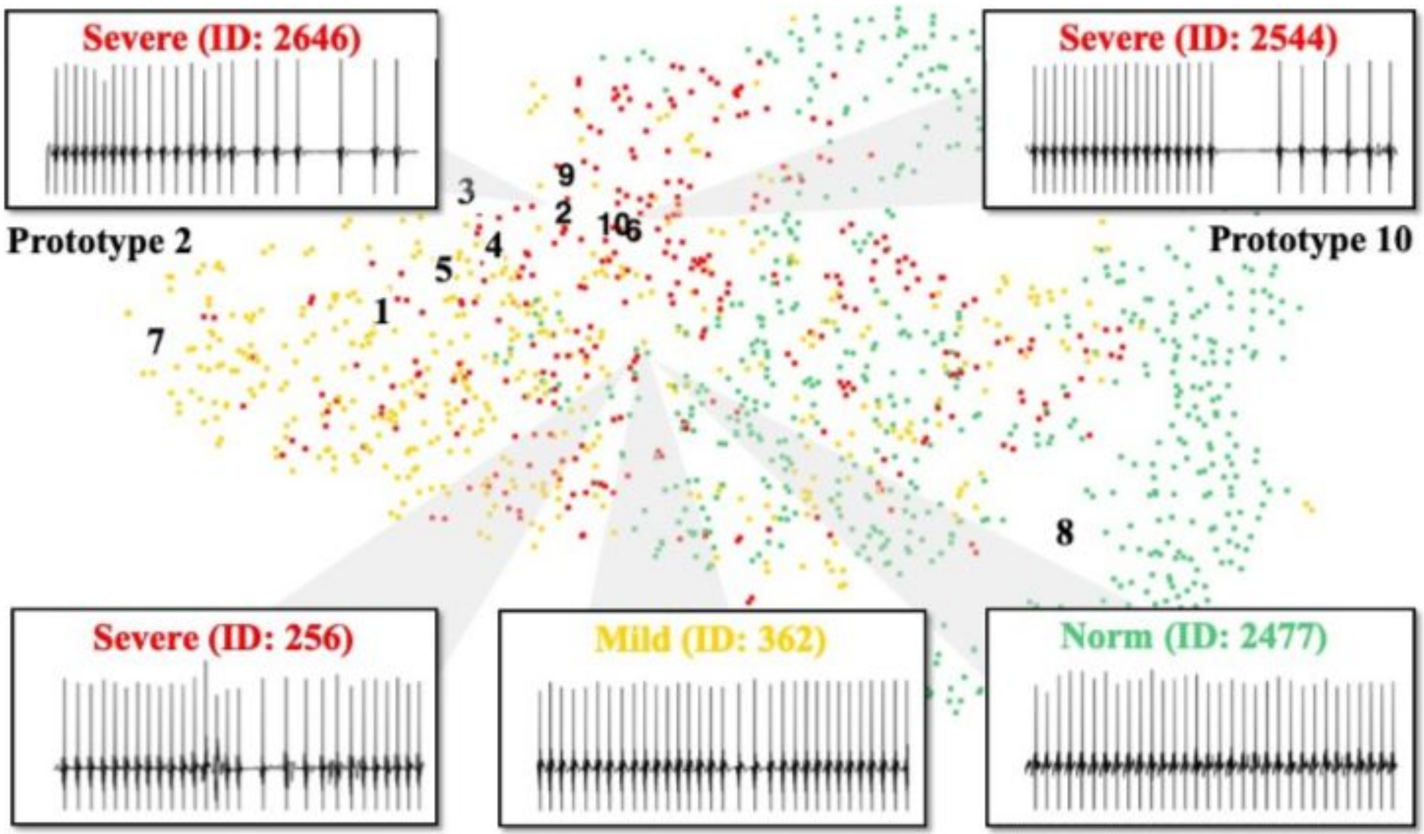
Prototype neighbor diversity Ψ_N

Prototype class diversity Ψ_C

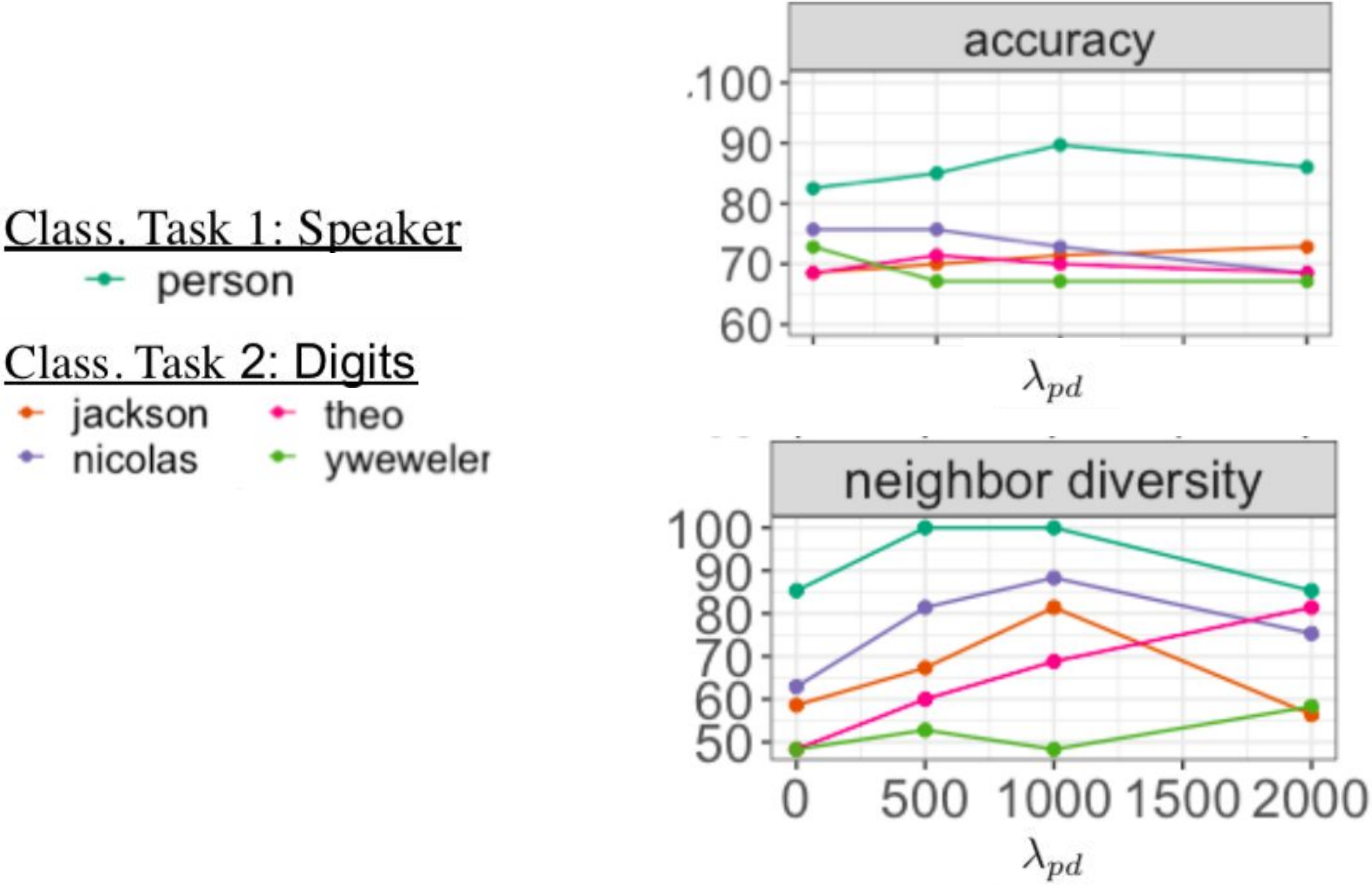
Explainability via Prototypes



Decoded Representations of Prototypes



Spoken MNIST Performance



Explainability via Prototypes

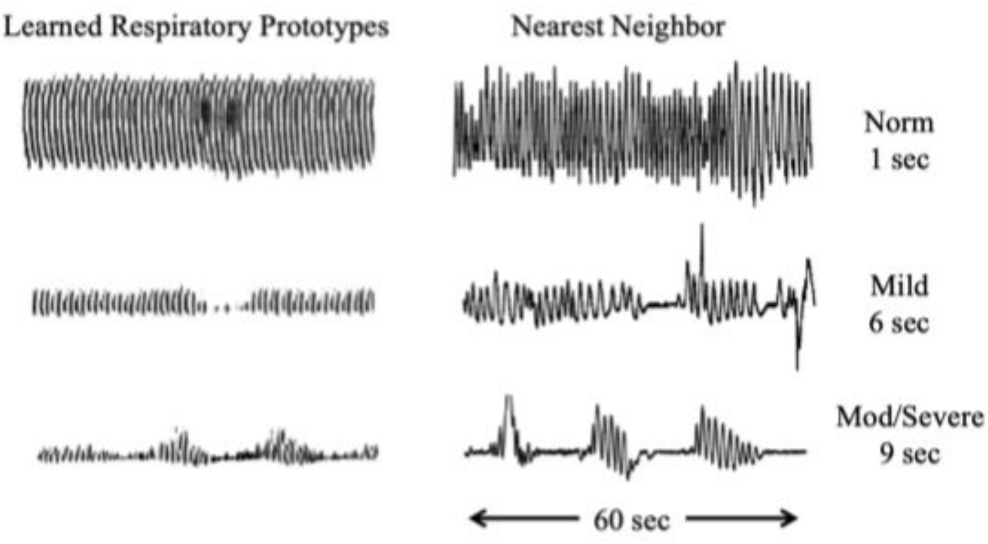


Figure 7: Learned prototypes showcase the diversity of features across classes that are important for understanding respiration morphology while classifying apnea events. For this classification task, we observe a variety of prototypes (at epoch 500) that learn various cases with cessation of breathing (6 and 9 second gaps) and the global features within the segment that are important for the model's classification. *(8-prototypes,* $\lambda_{pd} = 500$).

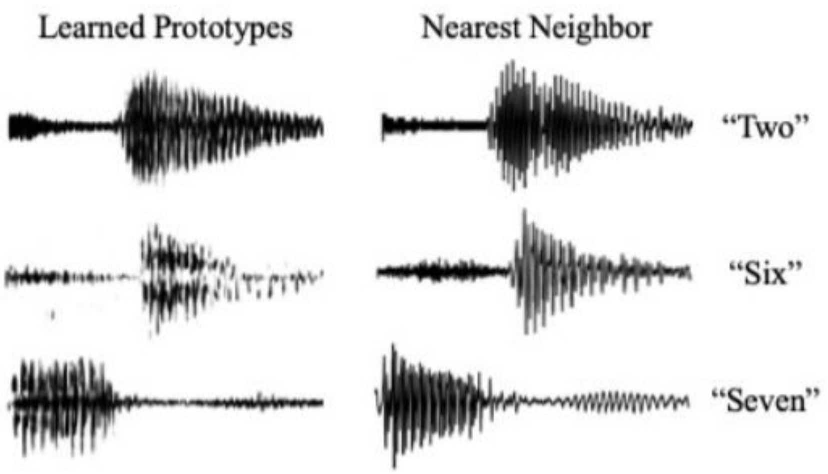


Figure 8: Learned prototypes from audio waveforms of spoken digits by Nicolas from the FSDD *(*$\lambda_{pd} = 500$*)*.

Explainability via Prototypes

Spoken Digit Global Explainability

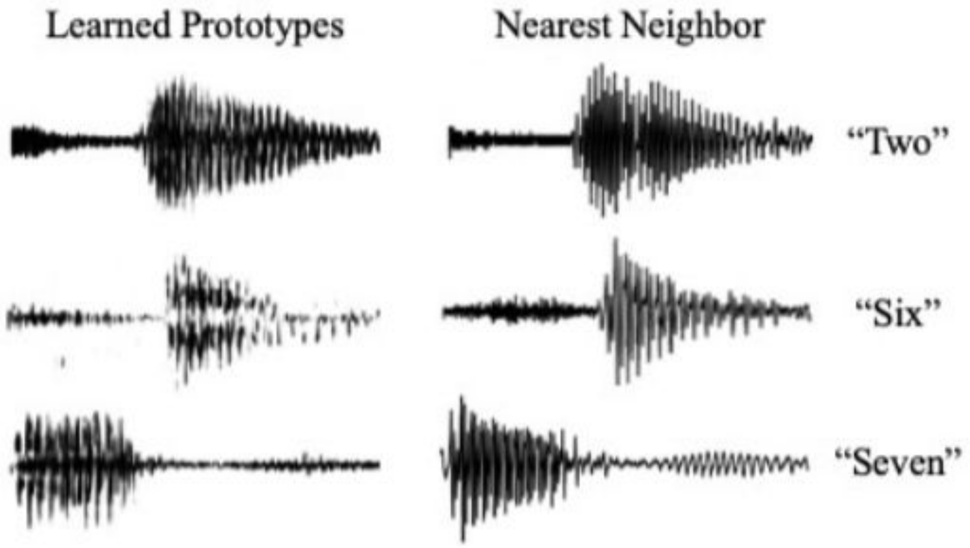
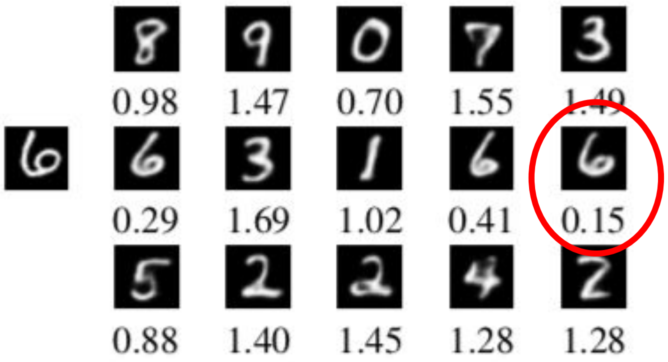
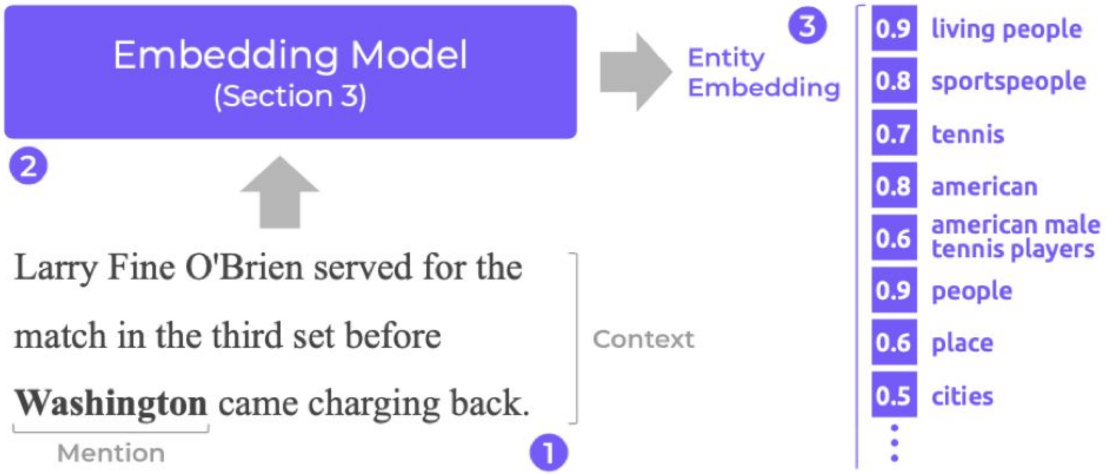


Figure 8: Learned prototypes from audio waveforms of spoken digits by Nicolas from the FSDD $(\lambda_{pd} = 500)$.

Instance Explainability



Onoe et al* learn human readable interpretable entity representations that achieve high performance without additional learning ("out of the box")



"Interpretable Entity Representations Through Large Scale Typing"
Yasumasa Onoe & Greg Durrett . Findings of EMNLP 2020

Can we adapt IERs for the **Biomedical Domain?**

[Glesatinib] is a dual inhibitor of c-Met and SMO
that is under phase II clinical trial for non-small cell lung cancer.

Using BIERs

(2) Entity label Classification for Cancer Genetics

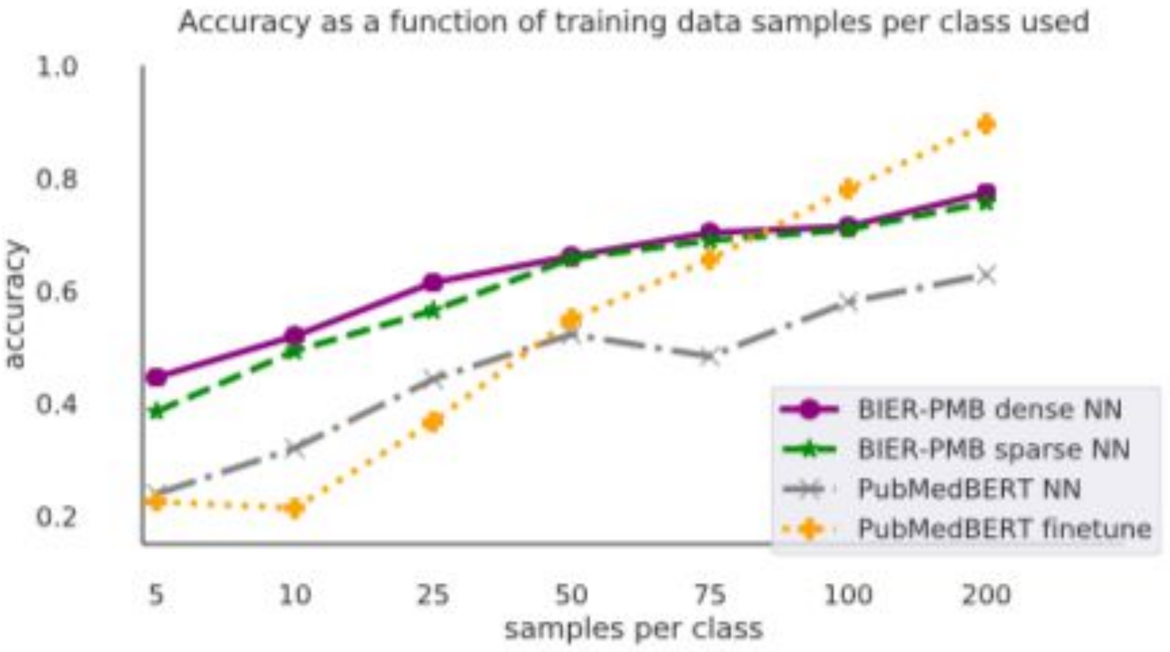


Figure 3: Results for the entity label classification task under varying amounts of supervision.

Using BIERs

(1) **Named Entity Disambiguation** (NED) on Clinical Entities.

Given a entity mention, context & set of candidate entities identify which of the candidates is the true one linked to the mention.

Model	Test Acc. Dot Prod	Test Acc. Cosine Sim
BIER-PubMedBERT (ours)	80.1	**84.0**
BIER-SciBERT (ours)	76.4	77.3
BIER-BioBERT (ours)	71.9	75.9
Onoe and Durrett (2020)	63.6	69.8
Popular Prior	73.9	-
PubMedBERT (Gu et al., 2020)	77.6	-
SciBERT (Beltagy et al., 2019)	77.4	-
BioBERT (Lee et al., 2019)	77.9	-

Table 2: BIER zero shot test results vs Logistic Regression Baselines trained on task data for NED task

Using BIERs

(2) Entity label Classification for Cancer Genetics

Model	Test Acc. L2 Dist Dense	Test Acc. L2 Dist Sparse	Test Acc. Dot Prod Dense	Test Acc. Dot Prod Sparse
BIER-PubMedBERT	85.5	86.8	**88.2**	**87.5**
BIER-SciBERT	70.8	77.0	72.8	76.8
BIER-BioBERT	83.4	85.9	85.6	86.8
Onoe and Durrett (2020)	63.9	55.1	60.0	59.9
PubMedBERT	77.3	-	69.3	-
SciBERT	74.4	-	75.2	-
BioBERT	67.6	-	59.6	-

Table 3: Test accuracy on Cancer Genetics data using a nearest neighbor classifier (k=1) without fine-tuning based on sparse output or intermediate dense embeddings using L2 or Dot Product distance metrics.

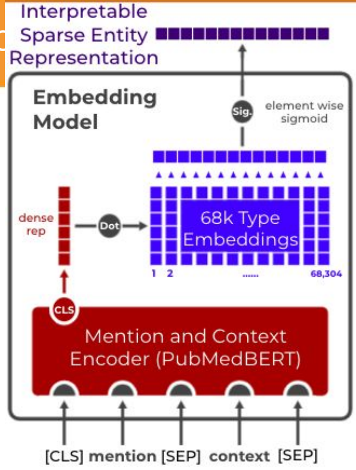
Using BIERs

(2) Entity label Classification for Cancer Genetics

Model	Test Acc.			
	L2 Dist		Dot Prod	
	Dense	Sparse	Dense	Sparse
BIER-PubMedBERT	85.5	86.8	**88.2**	**87.5**
BIER-SciBERT	70.8	77.0	72.8	76.8
BIER-BioBERT	83.4	85.9	85.6	86.8
Onoe and Durrett (2020)	63.9	55.1	60.0	59.9
PubMedBERT	77.3	-	69.3	-
SciBERT	74.4	-	75.2	-
BioBERT	67.6	-	59.6	-

Table 3: Test accuracy on Cancer Genetics data using a nearest neighbor classifier (k=1) without fine-tuning based on sparse output or intermediate dense embeddings using L2 or Dot Product distance metrics.

Debugging with BIERs



Allows for error analysis at the component level to identify areas lacking in supervision and/or possible changes to the type system.

How well the model could have done had it known to fallback to using the intermediate dense embedding in cases where the sparse representation led to an **incorrect prediction**

	Test Acc.			
Task	Dense	Sparse	Combined	Δ
NED	84.0	81.0	**91.7**	+7.7
ELC	87.5	88.2	**91.9**	+3.7

Table 5: Results for both tasks showing improvements that could have been achieved by combining intermediate dense and interpretable sparse output embeddings generated by the same BIER-PubMedBERT model.

Using BIERs

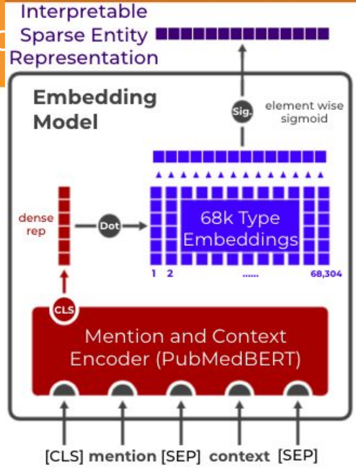
(1) **Named Entity Disambiguation** (NED) on Clinical Entities.

Given a entity mention, context & set of candidate entities,
identify which of the candidates is the true one linked to the mention.

Debugging with BIERs

Allows for error analysis at the component level
to identify areas lacking in supervision
and/or possible changes to the type system.

Debugging with BIERs



Allows for error analysis at the component level to identify areas lacking in supervision and/or possible changes to the type system.

How well the model could have done had it known to fallback to using the intermediate dense embedding in cases where the sparse representation led to an **incorrect prediction**

	Test Acc.			
Task	Dense	Sparse	Combined	Δ
NED	84.0	81.0	**91.7**	+7.7
ELC	87.5	88.2	**91.9**	+3.7

Table 5: Results for both tasks showing improvements that could have been achieved by combining intermediate dense and interpretable sparse output embeddings generated by the same BIER-PubMedBERT model.

Motivation for **future work** on developing a dynamic approach to making predictions that is a function of model confidence.

Error analysis using BIERS

context: The presence of activating TSH-R mutations has also been demonstrated in differentiated **thyroid carcinomas.** At present, the percentage of such a modification is low, unless referred to selected series of tumors.

mention: **thyroid carcinomas**

label: **Cancer**

Sparse NN model pred	Dense NN model pred
thyroid (label: Organ)	**esophageal carcinomas (label: Cancer)**
Types	**Types**
('gland', 0.99965), ('thyroid', 0.99932), ('rtt', 0.999), ('head_and_neck_cancer', 0.99093), ('neck', 0.97243), ('head_and_neck_anatomy', 0.93763), ('head', 0.86131), ('squamous-cell_carcinoma', 0.0024), ('ingredient', 0.00078), ('thyroid disease', 0.00047), ('nitrous_oxide', 0.00034), ('thyroid cancer', 0.0003), ('endocrine diseases', 0.00019),	('thyroid cancer', 0.99994), ('squamous-cell_carcinoma', 0.9998), ('thyroid', 0.99925), ('cancer', 0.99133), ('gland', 0.99039), ('nitrous_oxide', 0.01965), ('pancreatic_cancer', 0.00152), ('neck', 0.00023), ('thyroid_neoplasm', 0.00019), ('rtt', 0.00014), ('endocrine diseases', 2e-05), ('head', 1e-05), ('malignancy', 1e-05),

Error analysis using BIERS

context: The presence of activating TSH-R mutations has also been demonstrated in differentiated **thyroid carcinomas.** At present, the percentage of such a modification is low, unless referred to selected series of tumors.

mention: **thyroid carcinomas**

label: **Cancer**

Sparse NN model pred	Dense NN model pred	Counterfactual Sparse NN model pred
thyroid (label: Organ)	**esophageal carcinomas (label: Cancer)**	**medullary thyroid carcinoma (label: Cancer)**
Types	**Types**	**Types**
('gland', 0.99965), ('thyroid', 0.99932), ('rtt', 0.999), ('head_and_neck_cancer', 0.99093), ('neck', 0.97243), ('head_and_neck_anatomy', 0.93763), ('head', 0.86131), ('squamous-cell_carcinoma', 0.0024), ('ingredient', 0.00078), ('thyroid disease', 0.00047), ('nitrous_oxide', 0.00034), ('thyroid cancer', 0.0003), ('endocrine diseases', 0.00019),	('thyroid cancer', 0.99994), ('squamous-cell_carcinoma', 0.9998), ('thyroid', 0.99925), ('cancer', 0.99133), ('gland', 0.99039), ('nitrous_oxide', 0.01965), ('pancreatic_cancer', 0.00152), ('neck', 0.00023), ('thyroid_neoplasm', 0.00019), ('rtt', 0.00014), ('endocrine diseases', 2e-05), ('head', 1e-05), ('malignancy', 1e-05),	('cancer', 0.99994), ('rtt', 0.99964), ('nitrous_oxide', 0.99907), ('esophagus', 0.00159), ('endocrine diseases', 0.00013), ('pancreatic_cancer', 1e-04), ('gland', 4e-05), ('squamous-cell_carcinoma', 2e-05), ('neck', 2e-05), ('thyroid cancer', 1e-05), ('head_and_neck_anatomy', 1e-05), ('gastrointestinal cancer', 1e-05), ('head_and_neck_cancer', 0.0),

TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

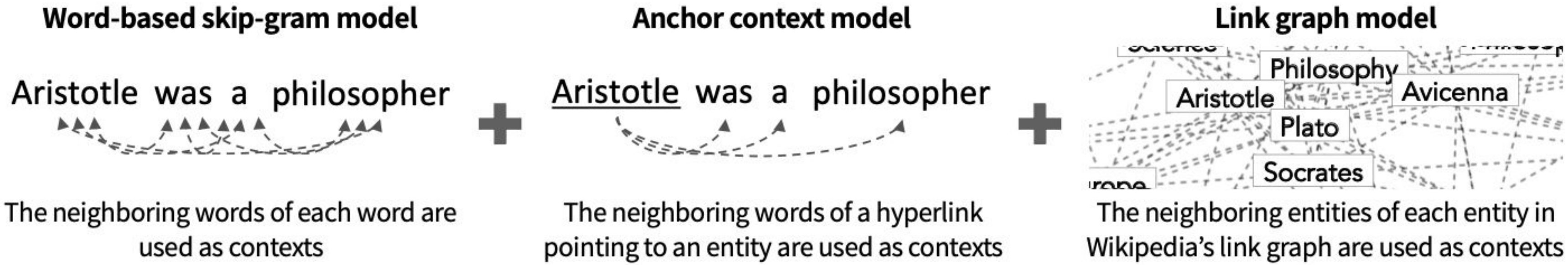


Figure 3: Wikipedia2Vec learns embeddings by jointly optimizing word-based skip-gram, anchor context, and link graph models.

KVQA Task and Motivation

Existing solutions to KVQA (largely)**:**

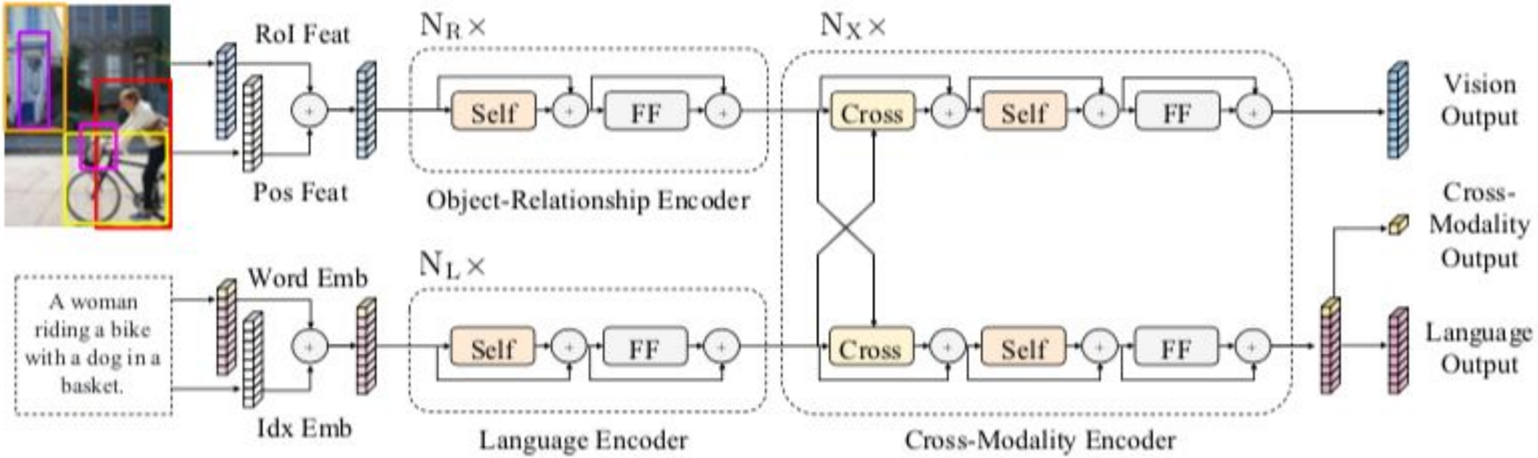
- Lack in-process explainable techniques
- Are entity centric and could benefit from grounding
- Treat the image modality as a sequence of region features

Existing VQA Solution

LXMERT

Tan et al

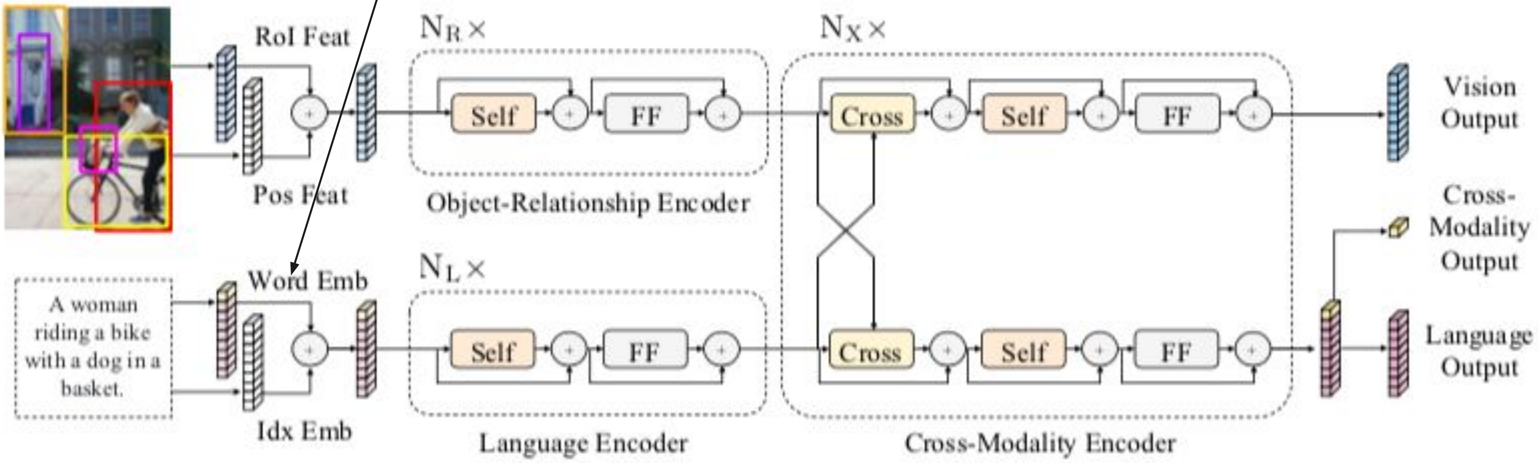
EMNLP 2019



EBERT "concat" method.

E-BERT(ent) -> mapper(WikiVec(ent)) + "/" + BERT(ent)

LXMERT



\+ **Entity Enhanced "EBERT"** (Poerner, et al EMNLP 2020) **over Language maps Wiki knowledge graph embeddings to BERT space (knowledge injection)**

Learn map W during training

$$\sum_{x \in \mathbb{L}_{\mathrm{WP}} \cap \mathbb{L}_{\mathrm{Word}}} ||\mathbf{W}\mathcal{E}_{\mathrm{Wikipedia}}(x) - \mathcal{E}_{\mathrm{BERT}}(x)||_2^2$$

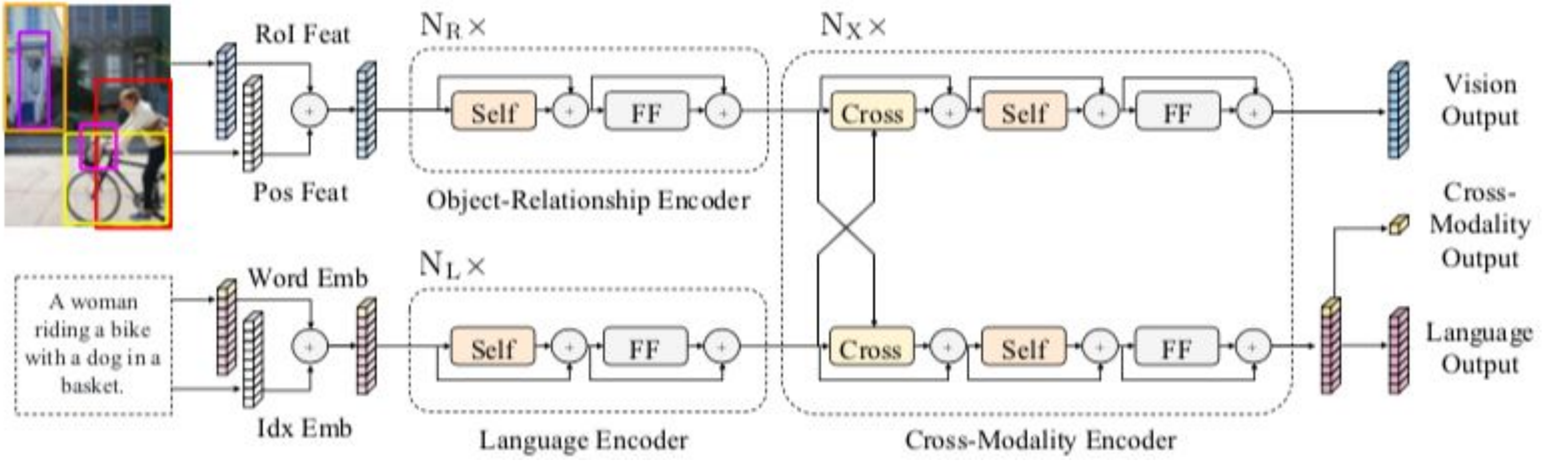
At Inference map Wiki ents to Bert via W

$$\mathcal{E}_{\text{E-BERT}} : \mathbb{L}_{\mathrm{Ent}} \rightarrow \mathbb{R}^{d_{\mathrm{BERT}}}$$

$$\mathcal{E}_{\text{E-BERT}}(a) = \mathbf{W}\mathcal{E}_{\mathrm{Wikipedia}}(a)$$

KVQA Existing Solution

LXMERT



Pre-training on MS COCO and Visual Genome

9.18 M image, question pairs

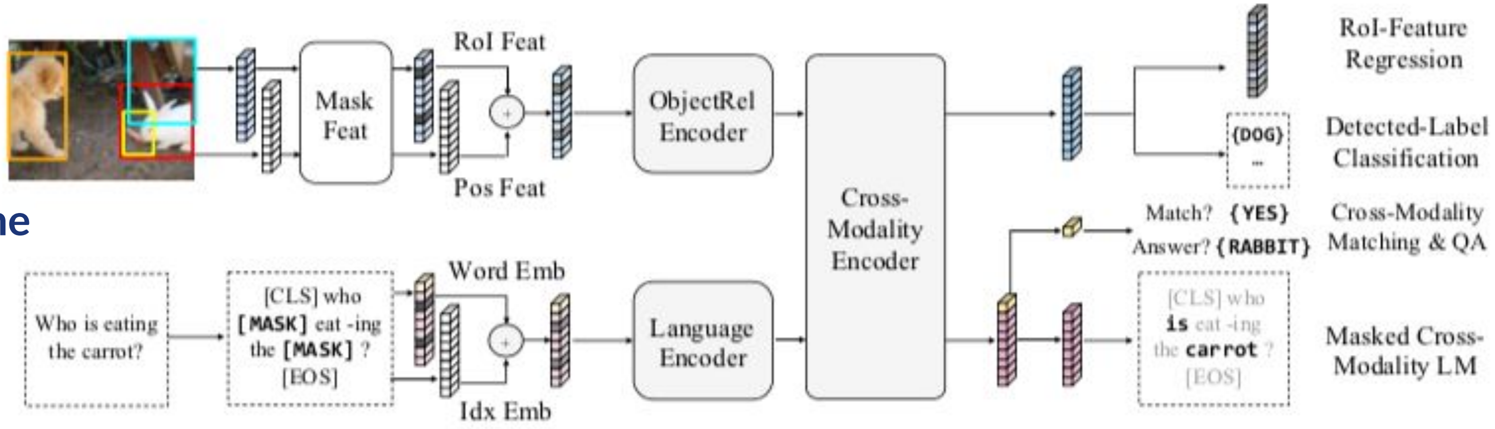


Table 3: KVQA results by question type accuracy (top half) and confidence (bottom 4 rows of unconstrained logits). Not shown NERper has highest accuracy for spatial question types (31.42). Average E-BERT refers to averages over NERper, NERagro and KVQAmeta for each link type (as is, links, noisy)

Model	Type	1-hop	multi hop	multi rel	bool	multi entity	cmp	spatial	subtr	count	inter	Acc / Conf
Percent	with	81.80	18.20	53.58	24.63	24.96	16.81	15.22	12.07	7.89	1.82	-
Question	-	44.89	57.98	47.40	86.37	72.14	81.67	28.12	19.68	84.62	65.00	47.27
+ Caption	-	46.36	65.47	51.57	**87.21**	72.46	80.91	29.17	19.33	85.03	70.29	49.84
KVQAmeta	links	48.87	70.61	55.43	86.69	**73.68**	**82.50**	31.14	**22.21**	84.82	**71.47**	52.83
KVQAmeta	noisy	**48.88**	**71.55**	**56.14**	86.63	73.57	82.15	31.14	21.23	**85.70**	70.00	**53.01**
Average	E-BERT	47.38	67.48	53.04	86.24	72.98	81.85	30.48	20.58	85.15	68.46	51.04
Best E-BERT	- Caption	2.52	6.08	4.57	-0.13	1.22	1.59	2.25	2.88	0.67	1.18	3.17
Question	-	-0.01	1.32	0.05	3.20	2.21	2.89	-1.69	-1.79	5.57	1.76	0.23
+ Caption	-	0.50	2.70	1.00	4.26	3.15	3.85	-1.18	-1.83	5.97	3.52	0.90
KVQAmeta	links	1.08	4.26	1.99	4.65	3.54	4.16	-0.71	-1.52	6.86	3.54	1.66
KVQAmeta	noisy	**1.52**	**4.84**	**2.48**	**5.87**	**4.34**	**5.02**	**-0.44**	**-1.51**	**7.31**	**5.24**	**2.12**

Table 6: KVQA entity knowledge injection explainability on split 1 for various entity span sets. For instance, 11.48 % of inference questions have E-BERT entities in their top 5 tokens for the NERper plain entity set model and overall 78% of questions in that entity set have E-BERT injected entities.

Model	Type	bimodal generic top1	bimodal generic top5	bimodal generic top10	transformer attention top1	transformer attention top5	transformer attention top10	Qs w/ EBERT
NERper	as is	0.66	11.48	31.23	0.29	6.13	22.64	.78
NERper	links	0.32	8.67	33.32	0.39	6.90	25.24	.79
NERper	noisy	0.13	4.75	21.62	0.73	7.11	23.38	.94
NERagro	as is	0.31	4.93	19.60	0.38	7.41	28.32	.91
NERagro	links	0.56	14.75	44.46	1.10	18.52	50.02	.97
NERagro	noisy	1.30	20.53	44.94	1.43	18.23	40.95	.97
KVQAmeta	as is	0.12	2.77	8.52	0.18	6.30	15.56	.87
KVQAmeta	links	0.39	4.26	12.96	4.06	12.57	23.80	.95
KVQAmeta	noisy	0.15	5.15	23.75	0.42	10.02	36.19	.99

TEXAS
The University of Texas at Austin

Thanks for listening!

Code/data: https://github.com/diegoolano/kbvqa

Pre-print: https://arxiv.org/abs/2112.06888

www.diegoolano.com
Twitter: @dgolano

ItsIRL - Global Prototypes

Positive class prototypes

1) Run the decoder fine-tuned model over the task training data.
2) Gather all correctly predicted instances for each class, sum their interpretable entity type layer representations & normalize them

$$\text{Positive class prototype} = \frac{\text{vec} - \min(\text{vec})}{\max(\text{vec}) - \min(\text{vec})}$$

where vec is the sum of entity type layers for a given class.

	Gene or gene product	Cell	Cancer	Simple chemical	Organism	Multi-tissue structure	Tissue
1	protein	cell	disease	ingredient	taxonomy	blood	tissue
2	ingredient	elementary particle	neoplasm	acid	mammals in 1758	angiology	cell
3	human	human cells	oncology	rtt	humans	soft tissue	human body
4	gene	battery	tissue	who essential medicines	tool-using mammals	nephron	connective tissue
5	coagulation	gene	abnormality	chemical compound	anatomically modern humans	blood vessel	endocrine system

Table 3: Top Entity Types for 7 most frequent positive Prototype class embeddings

ItsIRL - Type Manipulation

Class	Term Rules Inclusion/Exclusion	Terms in Set
Cell	[cell]	357
Cellular component	[cell]	357
Cancer	[cancer, neoplasm]	155
Gene or gene product	[' gene', 'gene ', ' genes', 'genes '] , ' not in ['generation', 'general']	434
Simple chemical	[chemical, chemical]	80
Organism	[' organ', 'organ ', 'organism'] not in ['organization']	172
Organism substance	[' organ', 'organ ', 'organism'] not in ['organization']	172
Organism subdivision	[' organ', 'organ ', 'organism'] not in ['organization']	172
Organ	[' organ', 'organ ', 'organism'] not in ['organization']	172
Tissue	[tissue, tissue]	15
Multi-tissue structure	[tissue, tissue]	15
Amino acid	[amino, amino , amino acid]	22
Pathological formation	[pathological]	3
Immaterial anatomical entity	[anatomical , anatomical, anatomical]	11
Developing anatomical structure	[anatomical , anatomical, anatomical]	11
Anatomical system	[anatomical , anatomical, anatomical]	11

Table 6: Terms used to create coarse Class specific Entity Type sets

ItsIRL - Type Manipulation

Digging Into the Results:

Technique	% of Errors Corrected	Best Method
Promoting	50.0%	11 out of 15
Both P & F	49.3%	10 out of 15
Fixing	28.5%	6 out of 15
Best of 3	61.0%	15 out of 15

Obtained using noisy, non-expert term sets.

True	Predicted	Errs	BOTH	PRMT	FIX	Best%
Chemical	Gene	65	64	48	59	98.4
Cell	Cancer	41	31	41	0	100
Cell	Gene	34	34	34	0	100
Multi-Tis	Tissue*	22	0	0	7	31.8
Gene	Chemical	17	3	3	10	58.8
Organ	Tissue	16	12	10	12	75
Cancer	Cell	16	0	14	0	87
Gene	Organism	15	6	0	15	100
Cell	Chemical	14	14	14	4	100
Amino	Gene	14	14	14	14	100
Pathol	Cancer	14	0	0	0	0
Organism	Cell	14	0	0	0	0
Organism	Gene	12	0	2	0	16.7
Organ	Multi-Tissue	10	0	1	0	10
Multi-Tis	Cancer	10	0	0	0	0
Chemical	Amino	10	10	10	10	100
Cancer	Org. Sub.	10	10	10	0	100
Cell	Tissue	10	10	10	5	100
Cell	Celu Comp*	10	10	10	0	100
	Raw Total	592	292	296	169	361
	Percent	100	49.3	50	28.5	61

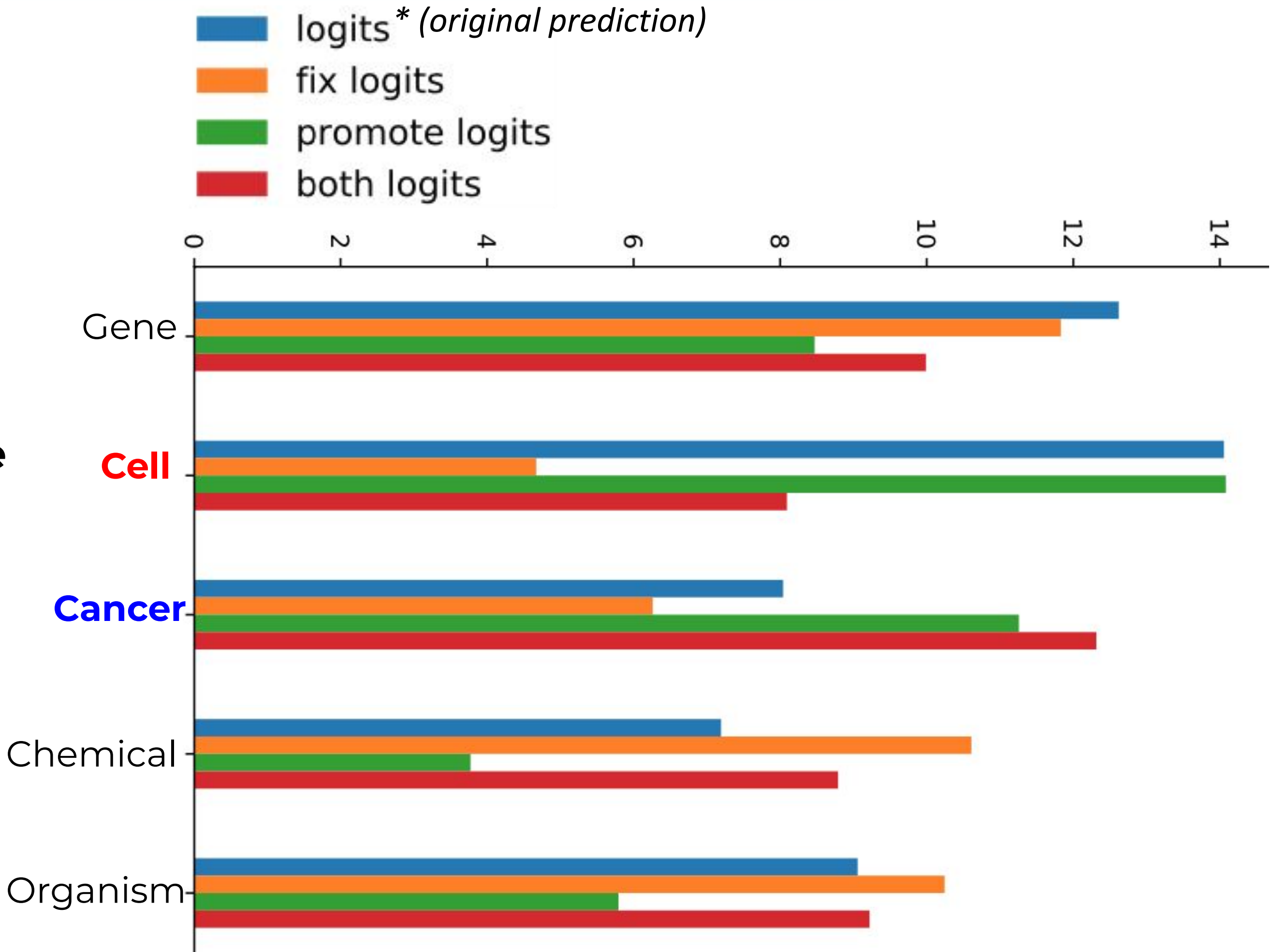
ItsIRL - Type Manipulation

Example on
Single Error Case

True Label: Cancer
Predicted: Cell

Results after Manipulation Technique

1. Fixing -> Gene
2. Promote -> Cell
3. Both -> Cancer



ItsIRL - Recap

- We propose Intermediate Entity-based Sparse Interpretable Representation Learning **(ItsIRL),** a pre-trained which provides an intermediate interpretable layer whose decoded dense representation output can be fine-tuned and used for performance on downstream tasks.

- Empirically we show the model **outperforms prior IERs work** and is competitive with dense language models on two biomedical tasks.

- To demonstrate the utility of the kind of interpretability afforded by ItsIRL, we propose a **counterfactual entity type manipulation analysis** which allows for modeling debugging. Using coarse class type sets, we show this technique can allow ItsIRL to surpass performance against dense non-interpretable models.

- We finally show how combining entity types over classes on the training set to create **positive and negative class prototypes** can be used to explain task specific global structure and semantics learned by our model.

UNICEN
Universidad Nacional del Centro de la Provincia de Buenos Aires

Predicting Gender Bias in Judicial Proceedings (Azul)

Dr. Maria DeArteaga at UT Austin

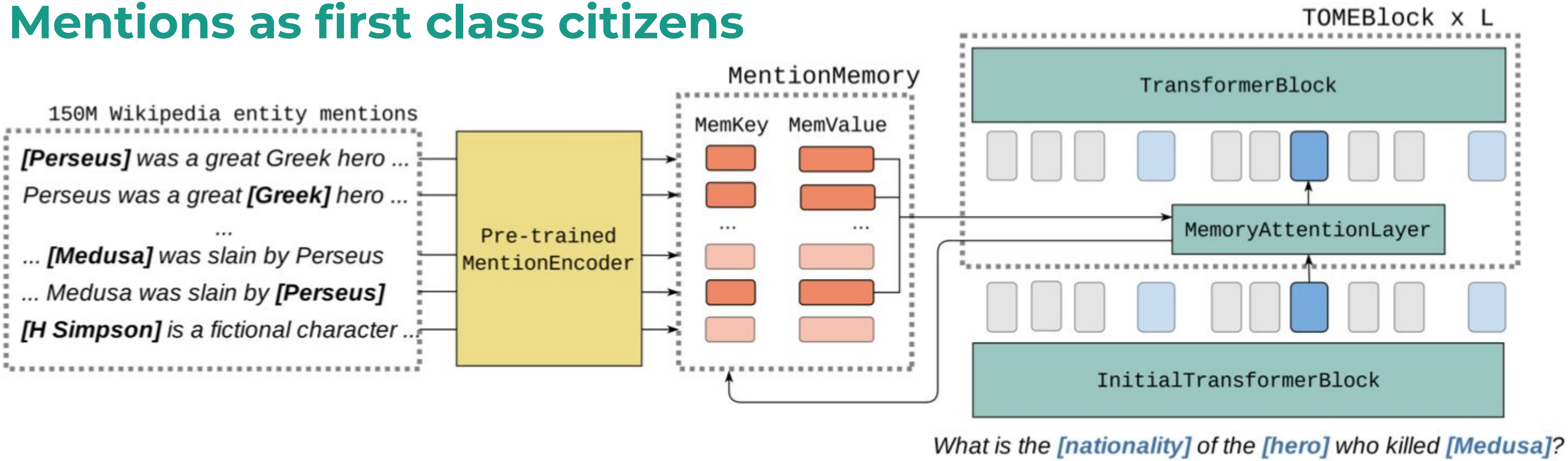
- Imbalanced task at the document and sentence level studying human bias and not learned biased representations.

- Setting up infrastructure for the data, sequence tagging, labeling and human in the loop feedback for iterative learning of spanish language model

- Giving workshops about diverse topics in AI/NLP

Mentions & Model Memories for improved Entity Learning, Retrieval & Reasoning over different domains

Mentions as first class citizens



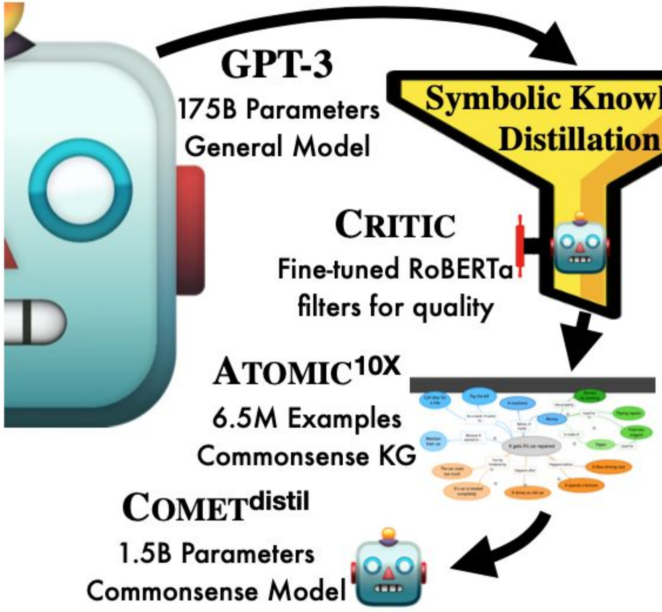
- Mention Memory: incorporating textual knowledge into transformers through entity mention attention. **ICLR 2022**
- MOLEMAN: Mention-Only Linking of Entities with a Mention Annotation Network - **ACL 2021**

Mentions & Model Memories for improved Entity Learning, Retrieval & Reasoning over different domains

1. Application of Mention and Memory techniques to **different tasks**, **specific product domains**, and possibly expanding to **multimodal entity centric settings**

2. How can **explainability methods** can be leveraged to guide them (via memory banks) and explain their internal reasoning.

Symbolic Knowledge Distillation and Human Critics for KB creation, model learning and explanations.



Yejin Choi's group at UW/Ai2

- Symbolic Knowledge Distillation from General Language Models to Commonsense Models

Symbolic Knowledge Distillation and Human Critics for KB creation, model learning and explanations.

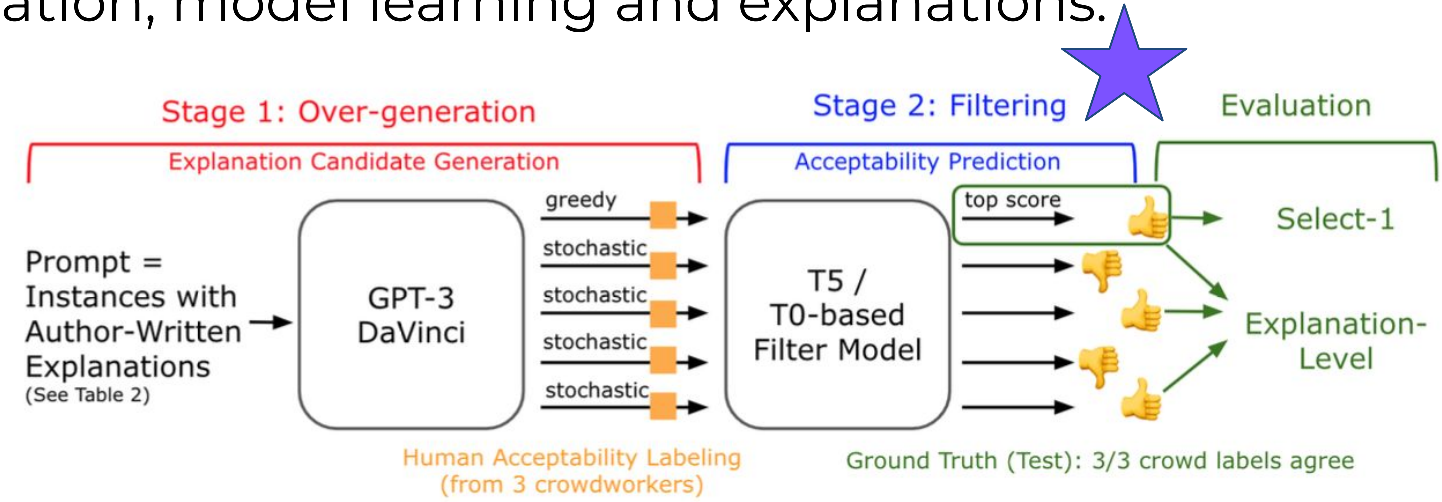


Figure 1: Illustration of our overgeneration + filtration pipeline approach to produce human-acceptable generated explanation for CommonsenseQA and SNLI instances (see examples in Table 1). Authors of this work write explanations to prompt GPT-3, generating five explanations per instance during Stage 1. An acceptability filter, trained with human binary acceptability judgments, determines which of these generated explanations as plausible. Our metrics evaluate the predicted ratings both at the explanation and at the instance level.

- Reframing Human-AI Collaboration for Generating Free-Text Explanations (2021)

Symbolic Knowledge Distillation and Human Critics for KB creation, model learning and explanations.

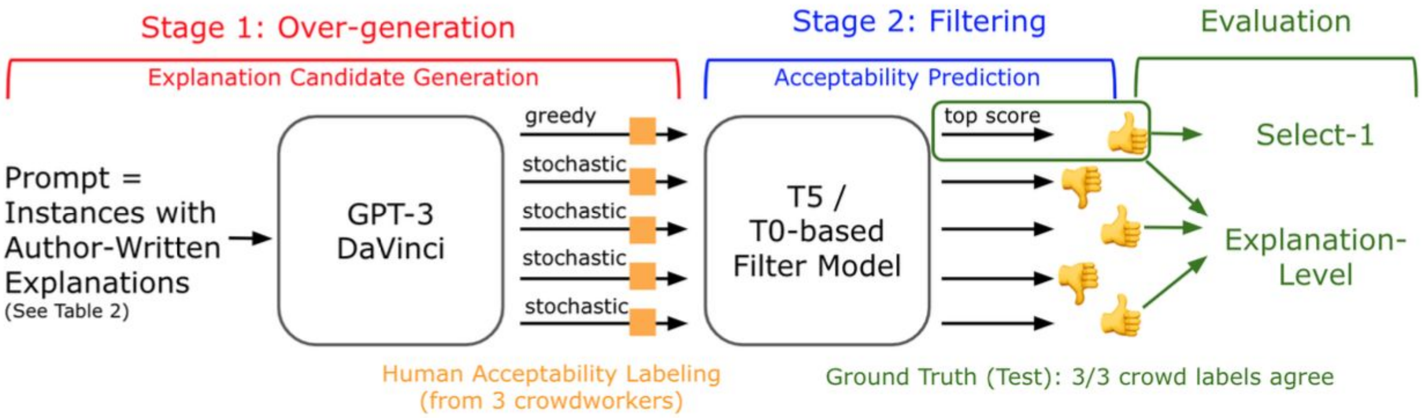
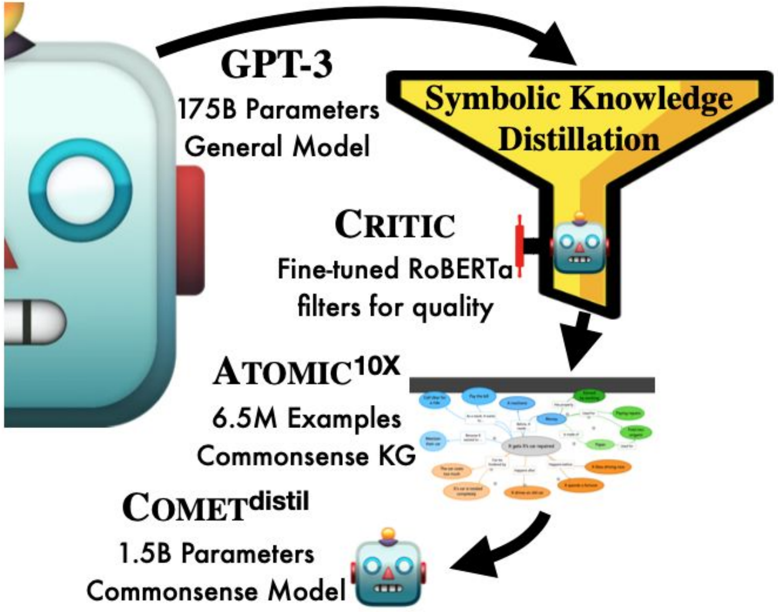


Figure 1: Illustration of our overgeneration + filtration pipeline approach to produce human-acceptable generated explanation for CommonsenseQA and SNLI instances (see examples in Table 1). Authors of this work write explanations to prompt GPT-3, generating five explanations per instance during Stage 1. An acceptability filter, trained with human binary acceptability judgments, determines which of these generated explanations as plausible. Our metrics evaluate the predicted ratings both at the explanation and at the instance level.

- Reframing Human-AI Collaboration for Generating Free-Text Explanations (2021)

Symbolic Knowledge Distillation and Human Critics for KB creation, model learning and explanations.



Yejin Choi's group at UW/Ai2

- Symbolic Knowledge Distillation from General Language Models to Commonsense Models

Symbolic Knowledge Distillation and Human Critics for KB creation, model learning and explanations.

Combining this single aspect symbolic knowledge distillation and human filtering method with other in-process techniques to

1) generate **knowledge bases for specific domains/tasks** and learning models on top of them or
2) train **explainer models for specific tasks/domains**

data augmentation for learning more robust models

Multi-modal setting where a model generates “constrained” images explaining the behavior of different layers and/or the model as a whole

Improving accuracy, robustness and transparency for **multi-modal models**

Study how the self supervised, teacher-student **Data2Vec** framework could be expanded and made more transparent,

Both in the modalities they focused on (text, speech and vision), but in particular for **text/graph** and **text/tabular** modalities,
(common settings for business with less general research focus)

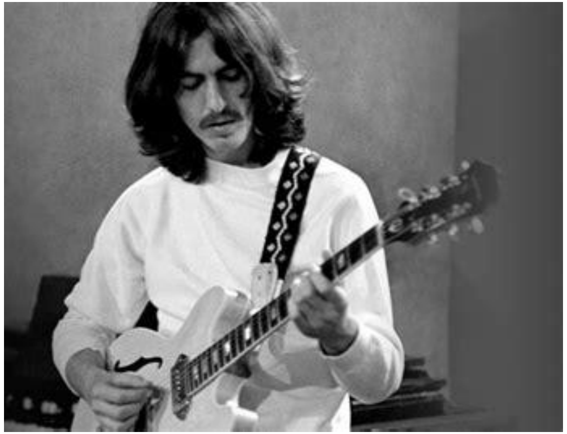
Adding **in-process explainability** to better understand the model, possible use of approximate influence functions from pre-training during inference time?

"Data2vec: A General Framework for Self-supervised Learning in Speech, Vision & Language" - 2022
"Benchmarking Multimodal AutoML for Tabular Data with Text Fields" NeurIPS 2021

Motivation

Example Query: What is George Harrison's favorite Nintendo game?

George Harrison

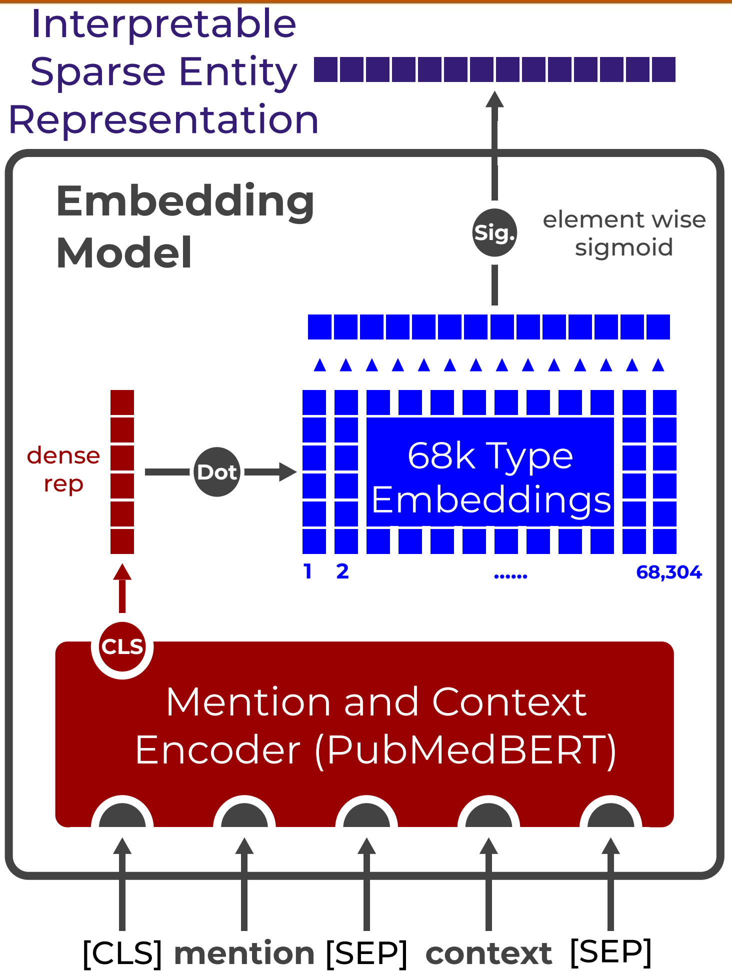


Q2643

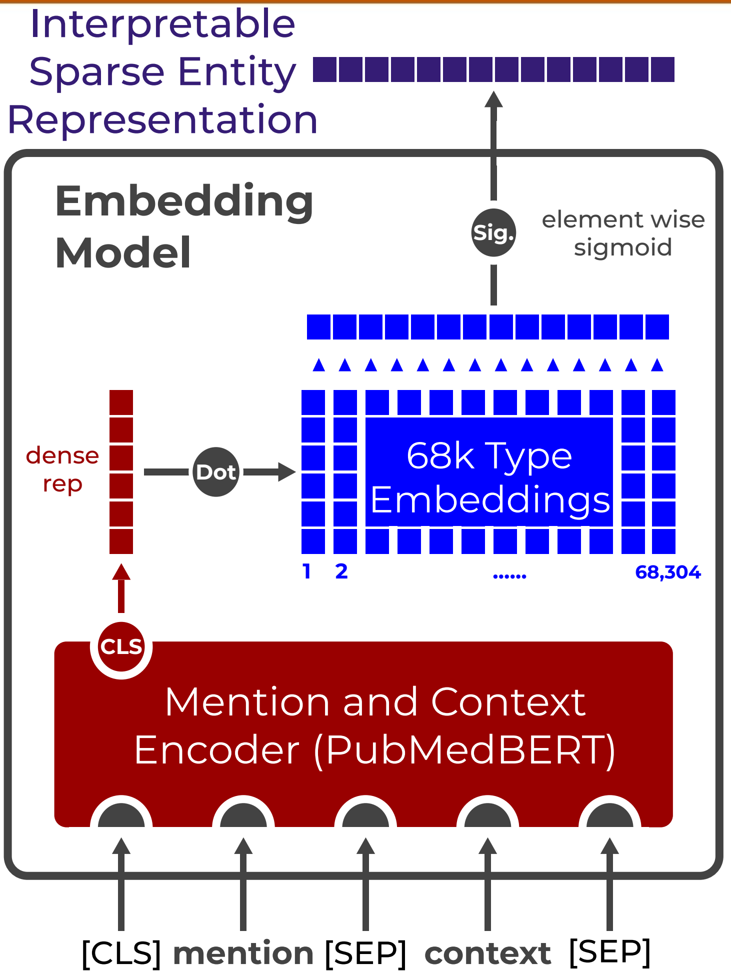
George Harrison

Q5540278

Training Biomedical IERs



Training Biomedical IERs



Training loss:

Independent sum
of binary cross entropy losses
over all all entity types T
over all training examples D.

$$-\sum_{i}^{D} \sum_{j}^{T} t^*_{ij} \cdot \log(t_{ij}) + (1 - t^*_{ij}) \cdot \log(1 - t_{ij}),$$

Future work

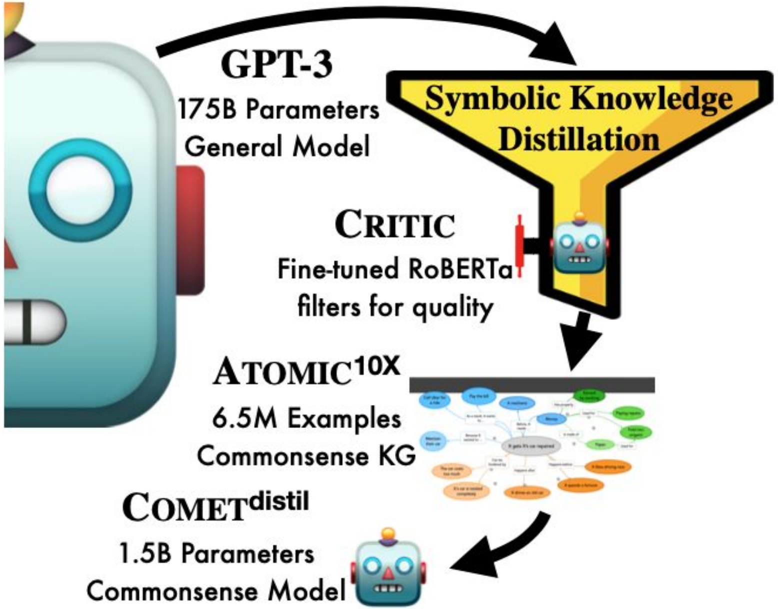
Application to **Large Language Models** (GPT3, Dall-E2, Imagen, etc)

Work around **prompting LLMs** and using smale-scale manual labeling
to **learn in-process critic models** that filter & improve quality of generated texts.

Future work

Application to **Large Language Models** (GPT3, Dall-E2, Imagen, etc)

Work around **prompting LLMs** and using smale-scale manual labeling
to **learn in-process critic models** that filter & improve quality of generated texts.



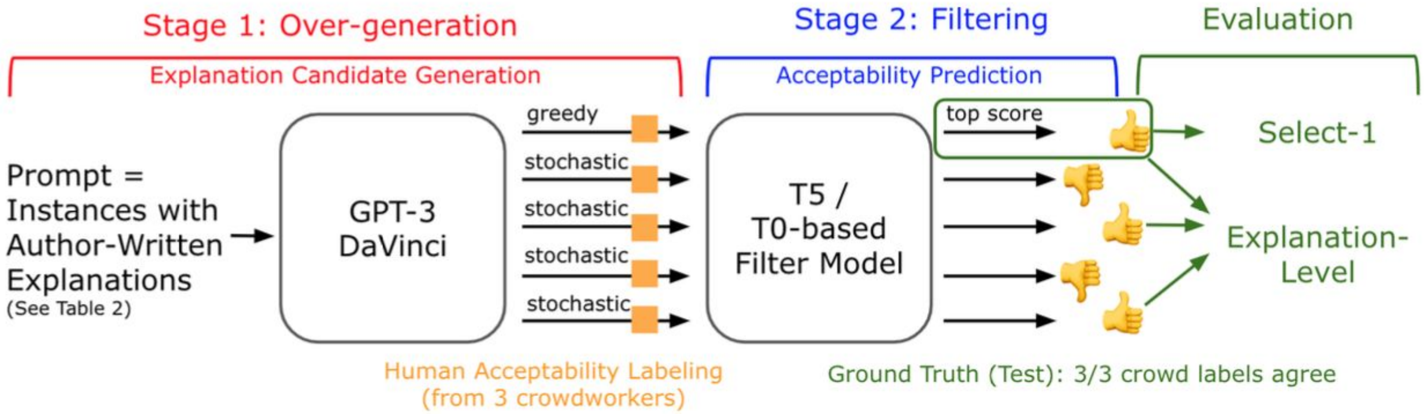
LLMs for automating knowledge base
creation in commonsense reasoning

(West et al 2021)
Symbolic Knowledge Distillation from General Language Models to Commonsense Models

Future work

Application to **Large Language Models** (GPT3, Dall-E2, Imagen, etc)

Work around **prompting LLMs** and using smale-scale manual labeling to **learn in-process critic models** that filter & improve quality of generated texts.



LLMs classifiers where high quality explanations are generated in-process (Wiegreffe., 2022)

Reframing Human-AI Collaboration for Generating Free-Text Explanations

Types of Explainable AI

Post Hoc explanations

Feature Attribution: which features contributed most to a model's output

- Path Integrated Gradients (IG)
- Shapley Additive Explanations (SHAP)
- Interpretability with Differential Masking

Influential examples: which training data most influenced a model's output

- Influence Functions
- Representer Point Selection for Explaining Deep Neural Networks

Counterfactuals: minimal change that would have led to a different output

BERT probing: assess how well a LM encodes semantic/syntatic properties of language by evaluating ("probing") on downstream tasks

Post Hoc Open Issues

Issues with Post Hoc secondary model explainers

Feature importance/saliency methods
- Need Baselines (Shap / IG)
- Are local/linear approximations of the actual model faithful explanations?
- Can we interpret Attention weights as explanations?

Influence functions:
- Expensive to compute
- Correlation to true influence for deep architectures is questionable

Counterfactuals:
- Semantic distance and meaning with text?

BERT probing:
- Don't generalize past probing tasks and don't "explain" model decisions

Explaining a network's behavior in a way that it wasn't expressly trained for can be problematic & makes assumptions that often do not hold (Chen, Rudin '20)

Post Hoc Open Issues

Issues with Post Hoc secondary model explainers

Feature importance/saliency methods
- Need Baselines (Shap / IG)
- Are local/linear approximations of the actual model faithful explanations?
- Can we interpret Attention weights as explanations?

Influence functions:
- Expensive to compute
- Correlation to true influence for deep architectures is questionable

Counterfactuals:
- Semantic distance and meaning with text?

BERT probing:
- Don't generalize past probing tasks and don't "explain" model decisions

Explaining a network's behavior in a way that it wasn't expressly trained for can be problematic & makes assumptions that often do not hold (Chen, Rudin '20)

In-Process

Prototypes: learn "prototypical" representations
- Deep Learning for Case-Based Reasoning through Prototypes

Deep k-NN models: utilize layer representations as additional "clustering" features
- Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust DL

Concept based Models: layer specific additional task loss with supervision
- Concept bottleneck models
- On completeness-aware concept-based explanations in deep neural networks

Retrieval as Explanation: for tasks involving entity retrieval as an intermediate step
- REALM: retrieval-augmented language model pre-training
- Entities as experts: Sparse memory access with entity supervision

Feature Importance as an auxiliary loss during training:
- Incorporating Priors with Feature Attribution on Text Classification

Require access and modifications to the underlying model ….

which is fine for critical applications!

Completed Work (Pre-Proposal)

Learning Dense Representations for Entity Retrieval. (CoNLL 2019)	Constructed a **dual mention-entity encoder** that learns dense representations for efficient neural **Entity Retrieval** with an **in-process, iterative hard negatives procedure** for **model learning and inference time inspection**.
Deep Classification of Time-Series Data with Learned Prototype Explanations. (ICML 19)	Adapted a **prototypical autoencoder** classifier to be compatible with **time series data** and allow for **tunable prototype diversity** leading to improved accuracy and **global and instance level explanations**.
Biomedical Interpretable Entity Representations. (ACL-IJCNLP 2021)	Learned a distantly supervised entity type system and data set for use in training a **Biomedical Interpretable Entity model** whose representations exist in a **semantically meaningful vector space** & whose **predictions may be interpreted and diagnosed** with an oracle method.

Completed Works - Post Proposal

Intermediate Entity-based Sparse Interpretable Representation Learning (*under submission*)	Extended **BIERs** to allow for task-centric **fine tuning** on pre-trained representations without breaking the semantics of our learned entity type space and introduced two **explainable diagnostic methods,** automated entity type manipulation & entity type based class prototypes, for **fine-grained model debugging** & **global model semantics interpretability**.
Improving and Diagnosing Knowledge Based Visual Question Answering via Entity Enhanced Knowledge Injection (WWW 22)	Analyzed how **efficient, entity based knowledge injection** via E-BERT during fine tuning affects an existing VQA model LXMERT on the task of **knowledge-based VQA** in terms of **accuracy & bi-modal explainability**.

Explainability Results

Figure 2: Two examples of KVQA questions where E-BERT is beneficial for KVQAmeta noisy entity set model. The rows show visual and token explanations for BM-GAE over the question/text (left column) and the 5 variants "Question", "+Caption", NERagro, NERper and KVQAmeta we explore . Next to each models name is their prediction and whether this top1 prediction is correct (1) or not, and then whether the correct answer exists in the top 5 predictions of the model which are additionally shown along with their logit values. Below that we see the top 5 most important tokens found by the explanation method followed by the set of Entities used for possible knowledge injection