

# The Project



CSE 598 Introduction to Deep Learning

# The Project

- You will work on a problem from beginning to end
- General guidelines for a standard project
  - The problem needs to be well defined
  - A dataset for the problem needs to be readily available (you do not have time to create one)
  - Your job will primarily be to:
    - understand the problem and get the data ready
      - this will take longer than you think
    - build models to solve that problem
      - replicate results in a paper: similar results is fine; alternative architecture is fine
      - it needs to be clear what is your idea and you implement
        - reusing code is fine, but you need to add more stuff
    - analyze the results and interpret them: qualitative analysis, requires some manual effort
      - code helps, but it won't give you the qualitative analysis

# The Project

- If you liked MiniTorch and want to understand more about Tensors and parallelization:
  - Complete Modules 3 and 4
    - You will get substantially less help from the course staff (compared to Modules 0-2)
    - but you do have test cases
      - If you figures out broadcasting and enjoy tinkering with a serious piece of software, this is a good choice
  - I expect most of you to not take this option
- You can work in groups (2 or 3 people max; ask and come up with a convincing argument if you want a larger group.

# The Project — Evaluation

- Project Proposal
  - Due on 10/27, soft deadline
    - You will get 10 points just for submitting (out of 100).
  - You will get feedback and a clear TODO list of what you need to do to get 100 in the project
    - You are the one who writes the TODO list. Follow the outline.
    - The course staff will not write a proposal for you. You are responsible for deciding the experiments you will run before you run them.
    - Submit by 10/29, get the proposal approved, and you will be fine.
      - Submit late or a bad proposal, and you will not do a good project.
- Project report: working code and report. A jupyter notebook is fine, but you need way more than working code.

# The Project — Proposal (1-2 pages)

- 1. General project idea (1-2 paragraphs)
- 2. The problem: Define the problem in terms of the input and output (1-2 paragraphs including examples)
  - What problem you are going to solve (not how)
  - Include examples
  - This has nothing to do with deep learning, machine learning, or models
  - Do not answer how you are going to solve the problem yet
- 3. Dataset(s): Describe the data you will work with
  - number of training, validation and test instances
  - number of labels / classes (or whatever you are going to predict)
  - Citation and source (link)
  - It needs to be readily available (downloadable), you do not have time to build a dataset

# The Project — Proposal (1-2 pages)

- 4. Baselines and Evaluation Metrics (1-2 paragraphs)
  - Define at least one baseline. A baseline is a model that is very simple
    - The simplest baseline is to choose a label randomly. If you only have 2 labels and the dataset is balanced, the random baseline gets 50% accuracy.
    - Look here for common baselines:  
<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.dummy>
      - “stratified”: generates predictions by respecting the training set’s class distribution.
      - “most\_frequent”: always predicts the most frequent label in the training set.
      - “prior”: always predicts the class that maximizes the class prior (like “most\_frequent”) and `predict_proba` returns the class prior.
      - “uniform”: generates predictions uniformly at random.
      - “constant”: always predicts a constant label that is provided by the user. This is useful for metrics that evaluate a non-majority class
  - Report results with the baseline in the proposal. Otherwise it will not be approved

# The Project — Proposal (1-2 pages)

- 4. Baselines and Evaluation Metrics (1-2 paragraphs)
  - Evaluation Metrics
    - Define how you are going to compare two models
      - If your network does not do better than the baseline, the network is useless
      - If you make the architecture more complicated and you do not observe improvements in results, your additions are useless.
    - Most likely (if classifying): Precision, Recall and F1-measure; Accuracy
      - The safest way is to use `classification_report(...)` from sklearn
      - [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\\_report.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html)

# The Project — Proposal (1-2 pages)

- 4. Baselines and Evaluation Metrics (1-2 paragraphs)
  - Evaluation Metrics
    - Define how you are going to compare two models
    - Most likely (if classifying): Precision, Recall and F1-measure; Accuracy
      - The safest way is to use `classification_report(...)` from sklearn
      - [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\\_report.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html)

```
>>> y_true = [0, 1, 2, 2, 2]
>>> y_pred = [0, 0, 2, 2, 1]
>>> target_names = ['class 0', 'class 1', 'class 2']
>>> print(classification_report(y_true, y_pred, target_names=target_names))
```

	precision	recall	f1-score	support
class 0	0.50	1.00	0.67	1
class 1	0.00	0.00	0.00	1
class 2	1.00	0.67	0.80	3
accuracy			0.60	5
macro avg	0.50	0.56	0.49	5
weighted avg	0.70	0.60	0.61	5



# The Project — Proposal (1-2 pages)

- 4. Baselines and Evaluation Metrics (1-2 paragraphs)
  - Evaluation Metrics
    - Define how you are going to compare two models
    - Most likely (if classifying): Precision, Recall and F1-measure; Accuracy
      - The safest way is to use `classification_report(...)` from sklearn
      - [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\\_report.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html)

```
>>> y_true = [0, 1, 2, 2, 2]
>>> y_pred = [0, 0, 2, 2, 1]
>>> target_names = ['class 0', 'class 1', 'class 2']
>>> print(classification_report(y_true, y_pred, target_names=target_names))
```

	precision	recall	f1-score	support
class 0	0.50	1.00	0.67	1
class 1	0.00	0.00	0.00	1
class 2	1.00	0.67	0.80	3
accuracy			0.60	5
macro avg	0.50	0.56	0.49	5
weighted avg	0.70	0.60	0.61	5

# The Project — Proposal (1-2 pages)

- 5. Experiments and Results (2-3 paragraphs)
  - What networks you expect to try and what results you expect to get?
  - “I will replicate the architectures reported in paper X” is fine, but do list the networks and provide a full citation to the paper.
  - Running code available in GitHub is not enough
    - If you can download the implementation, you do not credit for running it
    - Reusing and adapting code is fine, reading APIs and using them for your needs is fine
- Following this outline guarantees that I will know what you plan to do
  - If you don't have an approved proposal, don't expect a good grade

# Ideas

- Generally speaking, anything you read in these venues is good
  - Natural Language Processing: ACL, NAACL, EMNLP, EACL, COLING, etc.
    - For datasets: LREC, SemEval
    - All available here: <https://aclanthology.org/>
  - Computer Vision: CVPR, ECCV, ECCV, etc.
  - AI, ML, etc.: AAAI, IJCAI, ICML, ICLR, NeurIPS, KDD, etc.
- A problem presented in any of those venues is most likely interesting for your project.
  - There are many other good venues to look at!

# Ideas - Natural Language Processing

- There are a ton of problems in which
  - the input text (one or more pieces of text)
  - the output is a label (the label indicates some degree of understanding about the input)
  - (We call this text classification: give me a bunch of text and I will put it in bins [one bin per label])
- One example: Natural Language Inference
  - Input:
  - Premise: A person on a horse jumps over an obstacle
  - Hypothesis A person is outdoors, on a horse
- Output:
  - **Entailment** / Neutral / Contradiction

# Ideas - Natural Language Processing

- The GLUE and SuperGlue Benchmarks provide a nicely formatted collection of corpora with interesting tasks (all of them are text classification):
  - <https://gluebenchmark.com/>, <https://super.gluebenchmark.com/>
  - Examples of tasks in the GLUE and SuperGLUE benchmarks

MultiRC

**Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*

**Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)*

# Ideas - Natural Language Processing

- The GLUE and SuperGlue Benchmarks provide a nicely formatted collection of corpora with interesting tasks (all of them are text classification):

BoolQ

**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*

**Question:** *is barq's root beer a pepsi product*    **Answer:** No

CB

**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?*

**Hypothesis:** *they are setting a trend*    **Entailment:** Unknown

COPA

**Premise:** *My body cast a shadow over the grass.*    **Question:** *What's the CAUSE for this?*

**Alternative 1:** *The sun was rising.*    **Alternative 2:** *The grass was cut.*

**Correct Alternative:** 1

# Ideas - Natural Language Processing

- The GLUE and SuperGlue Benchmarks provide a nicely formatted collection of corpora with interesting tasks (all of them are text classification):
  - <https://gluebenchmark.com/>, <https://super.gluebenchmark.com/>
  - Examples of tasks in the GLUE and SuperGLUE benchmarks

ReCoRD

**Paragraph:** (CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood

**Query** For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency    **Correct Entities:** US



# Ideas - Natural Language Processing

- The GLUE and SuperGlue Benchmarks provide a nicely formatted collection of corpora with interesting tasks (all of them are text classification):
  - <https://gluebenchmark.com/>, <https://super.gluebenchmark.com/>
  - Examples of tasks in the GLUE and SuperGLUE benchmarks

RTE

**Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*

**Hypothesis:** *Christopher Reeve had an accident.*     **Entailment:** False

WiC

**Context 1:** *Room and board.*     **Context 2:** *He nailed boards across the windows.*

**Sense match:** False

WSC

**Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.*     **Coreference:** False



# Ideas - Natural Language Processing

- Many more tasks and domains:
  - If you want to experiment with social media instead of cleaner text, check out TweetEval
    - <https://github.com/cardiffnlp/tweeteval>

- **Emotion Recognition**, [SemEval 2018 \(Emotion Recognition\)](#) - 4 labels: anger , joy , sadness , optimism
- **Emoji Prediction**, [SemEval 2018 \(Emoji Prediction\)](#) - 20 labels: ❤️ , 😍 , 😂 , ... 🌲 , 📷 , 🤪
- **Irony Detection**, [SemEval 2018 \(Irony Detection\)](#) - 2 labels: irony , not irony
- **Hate Speech Detection**, [SemEval 2019 \(Hateval\)](#) - 2 labels: hateful , not hateful
- **Offensive Language Identification**, [SemEval 2019 \(OffensEval\)](#) - 2 labels: offensive , not offensive
- **Sentiment Analysis\***, [SemEval 2017 \(Sentiment Analysis in Twitter\)](#) - 3 labels: positive , neutral , negative
- **Stance Detection\***, [SemEval 2016 \(Detecting Stance in Tweets\)](#) - 3 labels: favour , neutral , against

# Ideas - Natural Language Processing

- More text classification tasks:
  - Many times answers to binary questions are not “yes” or “no”:
    - <https://www.aclweb.org/anthology/2020.emnlp-main.601.pdf>
    - <https://github.com/google-research-datasets/circa>

**“Want to get some dinner together?”**

“I know a restaurant we could get a reservation at.”

“I have already eaten recently.”

“I hope to make it home by supper but I’m not sure I can.”

“Dinner would be lovely.”

“I’d rather just go to bed.”

“There’s a few new restaurants we could go to.”

“I would like that.”

“We could do dinner this weekend.”

“I would like to go somewhere casual.”

“I’d like to try the new Italian place.”

Table 1: A polar question with 10 indirect responses, taken from our corpus.

# Ideas - Natural Language Processing

- More text classification tasks:
  - Many times answers to binary questions are not “yes” or “no”:
    - <https://www.aclweb.org/anthology/2020.emnlp-main.601.pdf>
    - <https://github.com/google-research-datasets/circa>

<b>Yes</b> Q: Do you have any pets? A: My cat just turned one year old.	<b>Probably yes / sometimes yes</b> Q: Do you like mysteries? A: I have a few that I like.	<b>Yes, subject to some conditions</b> Q: Do you enjoy drum solos? A: When someone's a master.
<b>No</b> Q: Do you have a house? A: We are in a 9th floor apartment.	<b>Probably no</b> Q: Are you interested in fishing this weekend? A: It's supposed to rain.	<b>In the middle</b> Q: Did you find this week good? A: It was the same as always.

Table 6: Example question and answer pairs where all 5 annotators agreed on the label.

# Ideas - Natural Language Processing

- Inference / entailment from tables
  - <https://www.aclweb.org/anthology/2020.findings-emnlp.27.pdf>
  - <https://github.com/google-research/tapas>

Rank	Player	Country	Earnings	Events	Wins
1	Greg Norman	Australia	1,654,959	16	3
2	Billy Mayfair	United States	1,543,192	28	2
3	Lee Janzen	United States	1,378,966	28	3
4	Corey Pavin	United States	1,340,079	22	2
5	Steve Elkington	Australia	1,254,352	21	2

- Entailed:* Greg Norman and Steve Elkington are from the same country.  
Greg Norman and Lee Janzen both have 3 wins.
- Refuted:* Greg Norman is from the US and Steve Elkington is from Australia.  
Greg Norman and Billy Mayfair tie in rank.
- Counterfactual:* **Greg Norman** has the highest earnings.  
~~Steve Elkington~~ has the highest earnings.
- Synthetic:* 2 is less than wins when Player is Lee Janzen.  
The sum of Earnings when Country is Australia is 2,909,311.

# Ideas - Natural Language Processing

- All text classification tasks can be solved with a network that:
  - takes as input text (you will transform words into ids, and then embeddings)
  - uses some neural architecture to go from the embeddings to an output layer with softmax activation
    - the output layer will have dimensionality == number of labels
- Here is the recipe in Pytorch:
  - [https://pytorch.org/tutorials/beginner/text\\_sentiment\\_ngrams\\_tutorial.html](https://pytorch.org/tutorials/beginner/text_sentiment_ngrams_tutorial.html)
  - Get a simple network working first
    - You must get better results than the baseline, otherwise something is wrong
  - Then, start adding more stuff to the network to (hopefully) do better

# Ideas - Natural Language Processing

- State-of-the-art models use transformers
  - widely use library: <https://huggingface.co/transformers/>
- More sources of datasets:
  - Huge hub of many datasets: <https://github.com/huggingface/datasets>
  - Conversations: <https://convokit.cornell.edu/>
  - Read the original papers so you understand what the problem is

# Ideas - Natural Language Processing

- Not all natural language processing tasks are classification: sequence2sequence
  - machine translation: <https://www.statmt.org/wmt21/>
  - summarization
  - dialogue systems

**If you cannot define a metric to measure success,  
you need to find a different problem**

# Ideas - Computer Vision

- Recognize dog breeds:
  - [https://d2l.ai/chapter\\_computer-vision/kaggle-dog.html](https://d2l.ai/chapter_computer-vision/kaggle-dog.html)
  - Experiment with (just some examples):
    - different data augmentation
    - simpler and more complicated architectures
    - different filters and strides
    - color vs. grayscale
  - Figure out (just some examples):
    - which breeds are harder and easier to identify (and confuse)
  - And many other options



# Ideas - Computer Vision

- Recognize dog breeds:
  - [https://d2l.ai/chapter\\_computer-vision/kaggle-dog.html](https://d2l.ai/chapter_computer-vision/kaggle-dog.html)
  - German Shepherd vs. Chihuahua



# Ideas - Computer Vision

- Recognize dog breeds:
  - [https://d2l.ai/chapter\\_computer-vision/kaggle-dog.html](https://d2l.ai/chapter_computer-vision/kaggle-dog.html)

**Curly-Coated Retriever**



**American Water Spaniel**



# Ideas - Computer Vision

- Recognize dog breeds:
  - [https://d2l.ai/chapter\\_computer-vision/kaggle-dog.html](https://d2l.ai/chapter_computer-vision/kaggle-dog.html)

**Brittany**



**Welsh Springer Spaniel**



# Ideas - Computer Vision

- Colorized pictures
  - Is it difficult to get data?
  - <https://arxiv.org/abs/1603.08511>

ID

**Input**

**LEARCH**  
(Deshpande et al. 2015)

**Ours**

**Ground truth color**



1

# Ideas - Language and Vision

- Many problems are multimodal: you need language and vision
  - Pronoun Coreference in dialogues about pictures
    - <https://www.aclweb.org/anthology/D19-1516.pdf>
    - [https://github.com/HKUST-KnowComp/Visual\\_PCR](https://github.com/HKUST-KnowComp/Visual_PCR)

●

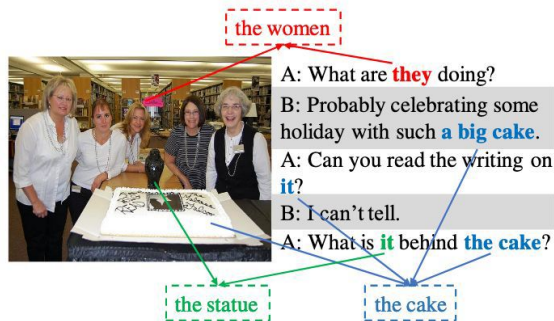


Figure 1: An example of a visual-related dialogue. Two people are discussing the view they can both see. Pronouns and noun phrases referring to the same entity are marked in same color. The first “it” in the dialogue labeled with blue color refers to the object “the big cake” and the second “it” labeled with green color refers to the statue in the image.



# Ideas - Language and Vision

- Many problems are multimodal: y
  - Pronoun Coreference in dialogues at
    - <https://www.aclweb.org/anthology>
    - <https://github.com/HKUST-Kr>

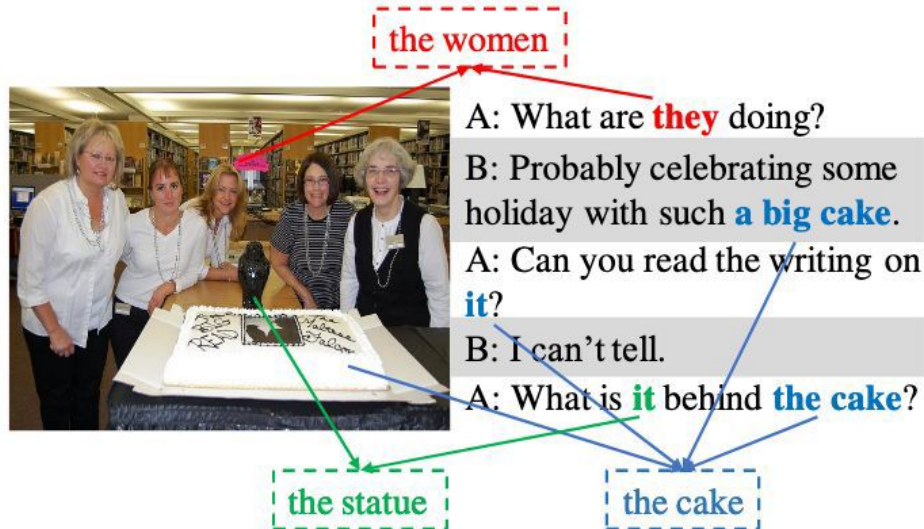
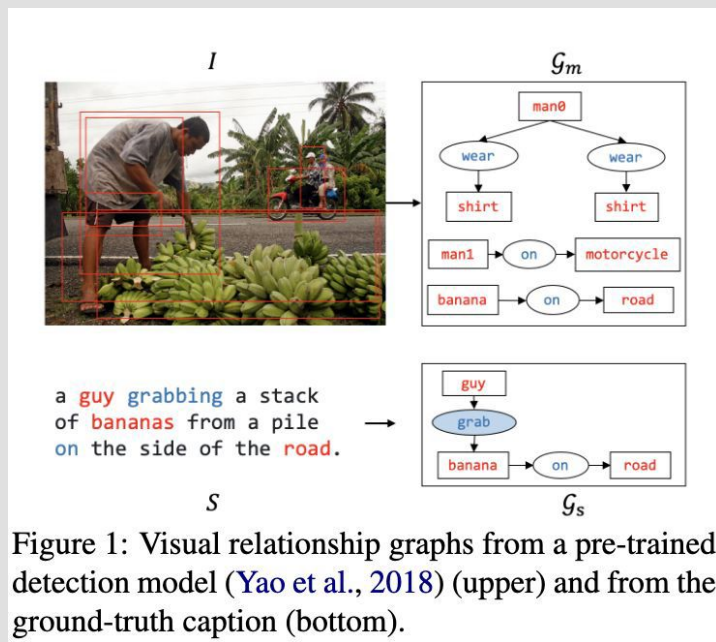


Figure 1: An example of a visual-related dialogue. Two people are discussing the view they can both see. Pronouns and noun phrases referring to the same entity are marked in same color. The first “it” in the dialogue labeled with blue color refers to the object “the big cake” and the second “it” labeled with green color refers to the statue in the image.

# Ideas - Language and Vision

- Many problems are multimodal: you need language and vision
  - Image captioning
    - <https://www.aclweb.org/anthology/2020.acl-main.664.pdf>



# Ideas - Language and

- Many problems are multi-modal
  - Image captioning
    - <https://www.aclweb.org/>

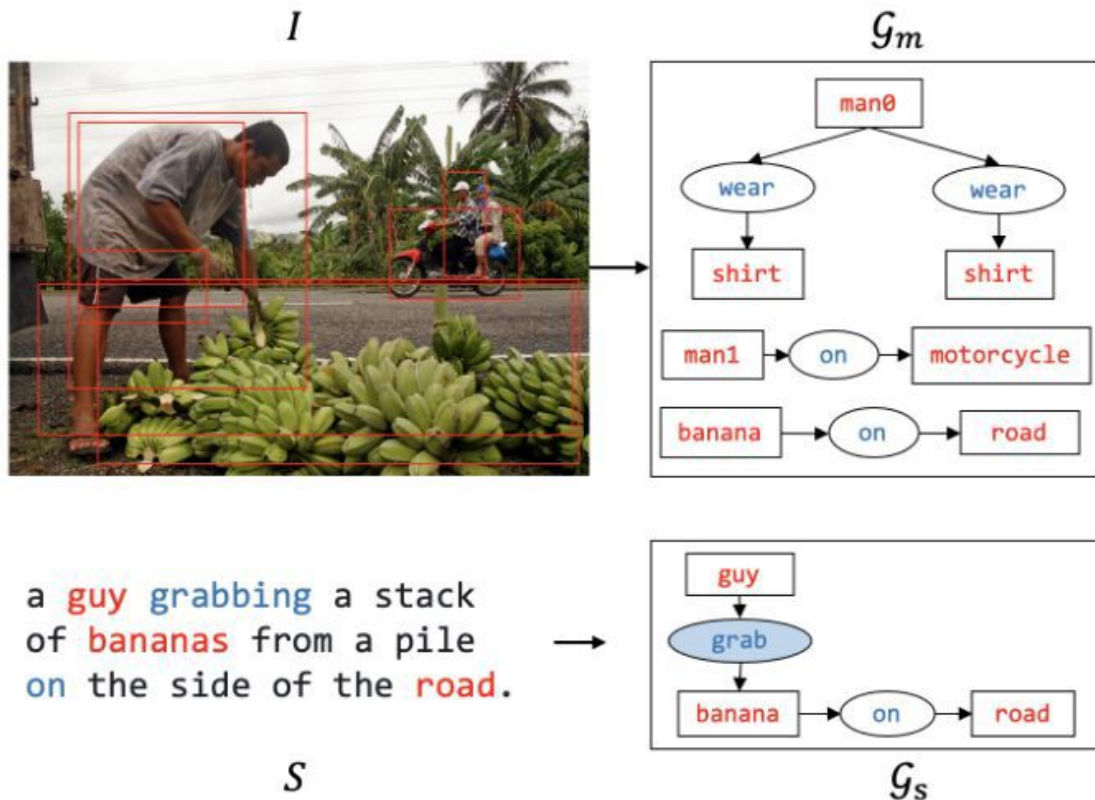


Figure 1: Visual relationship graphs from a pre-trained detection model (Yao et al., 2018) (upper) and from the ground-truth caption (bottom).



# Ideas - Language and Vision

- Visual Question Answering
  - Input: an image and a questions about the image
  - Output: the answer
  - <https://visualqa.org/>

Who is wearing glasses?

man



woman

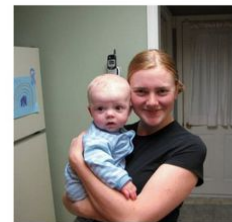


Where is the child sitting?

fridge



arms



Is the umbrella upside down?

yes



no



How many children are in the bed?

2

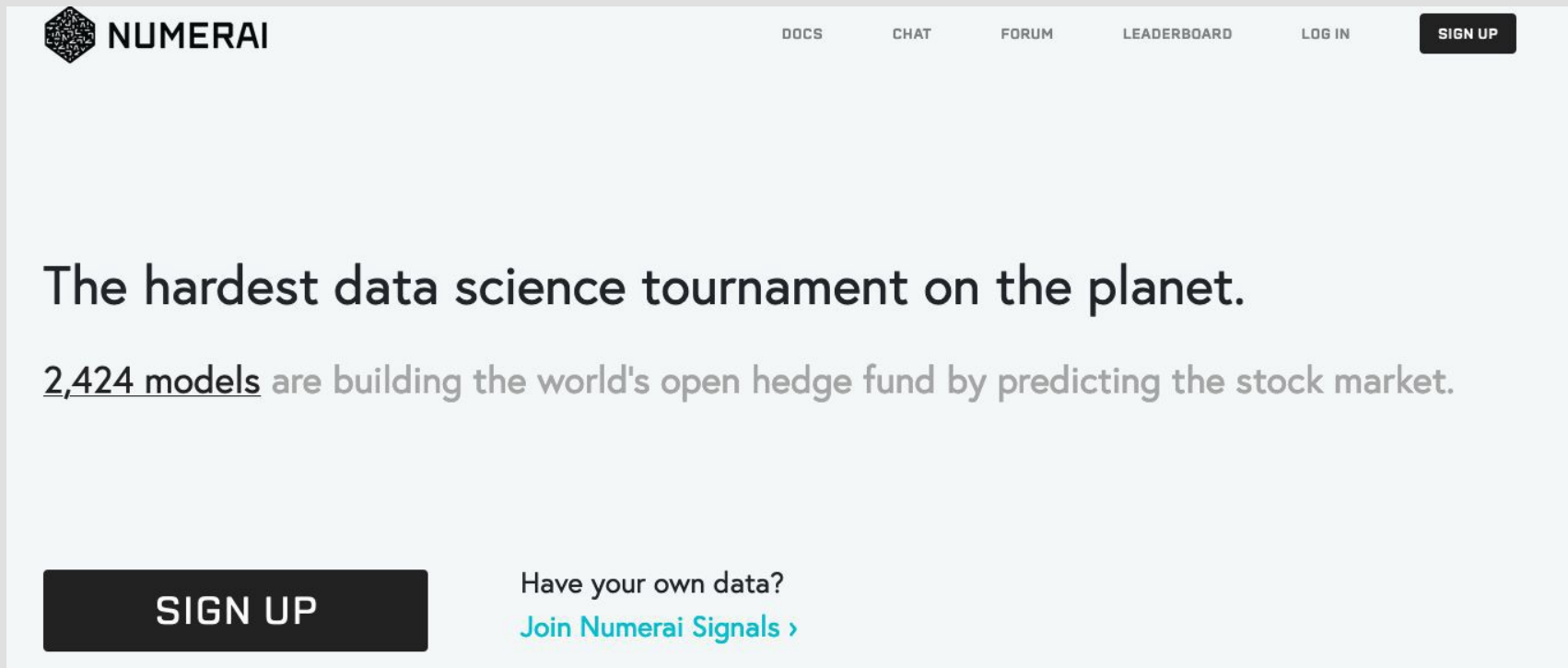


1



# Ideas - Other Problems

- Numerai



The screenshot shows the Numerai website. At the top is a navigation bar with the Numerai logo (a hexagon with a complex internal pattern) and the word "NUMERAI" on the left. To the right of the logo are links for "DOCS", "CHAT", "FORUM", "LEADERBOARD", and "LOG IN". A dark "SIGN UP" button is positioned on the far right of the navigation bar. Below the navigation bar, the main content area features the headline "The hardest data science tournament on the planet." followed by the text "2,424 models are building the world's open hedge fund by predicting the stock market." At the bottom left, there is a large dark "SIGN UP" button. To its right, the text "Have your own data?" is displayed above a link "Join Numerai Signals >" in a teal color.

NUMERAI

DOCS CHAT FORUM LEADERBOARD LOG IN SIGN UP

The hardest data science tournament on the planet.

2,424 models are building the world's open hedge fund by predicting the stock market.

SIGN UP

Have your own data?  
[Join Numerai Signals >](#)

# Ideas - Other Problems

- Numerai
  - Do not buy cryptocurrency and do not spend any money, PLEASE
  - Evaluation may be complicated: weekly submissions, feedback
    - Maybe you can do your own split
      - better: get together with classmates and compete against each other using the same splits
    - Can you evaluate in the comfort of your home?

## Start with hedge fund quality data.

It is clean and regularized, designed to be usable right away. Obfuscated, so it can be given out for free.

Got your own data? Try [Numerai Signals](#)

### Download Dataset

id	era	feature1	...	feature310	target
n2b2e3dd163cb422	era1	0.75	...	0.00	0.25
n177021a571c94c8	era1	1.00	...	0.25	0.75
n7830fa4c0cd8466	era1	0.25	...	1.00	0.00
nc584a184cee941b	era1	0.25	...	0.00	1.00
nc5ab8667901946a	era1	0.75	...	0.25	0.25
n84e624e4714a7ca	era1	0.00	...	0.75	1.00

# More Ideas

## QuAC : Question Answering in Context

**Eunsol Choi**<sup>★♥</sup>    **He He**<sup>★◇</sup>    **Mohit Iyyer**<sup>★♣†</sup>    **Mark Yatskar**<sup>★†</sup>  
**Wen-tau Yih**<sup>†</sup>    **Yejin Choi**<sup>♥†</sup>    **Percy Liang**<sup>◇</sup>    **Luke Zettlemoyer**<sup>♥</sup>


Allen Institute for Artificial Intelligence<sup>†</sup>    University of Washington<sup>♥</sup>  
Stanford University<sup>◇</sup>    UMass Amherst<sup>♣</sup>

`{eunsol,yejin,lsz}@cs.washington.edu`  
`{hehe,pliang}@cs.stanford.edu`  
`{mohiti,marky,scotttyih}@allenai.org`


# More Ideas

## Section: Daffy Duck, Origin & History


STUDENT: **What is the origin of Daffy Duck?**

TEACHER:  first appeared in Porky's Duck Hunt


STUDENT: **What was he like in that episode?**

TEACHER:  assertive, unrestrained, combative


STUDENT: **Was he the star?**

TEACHER:  No, barely more than an unnamed bit player in this short


STUDENT: **Who was the star?**

TEACHER:  No answer


STUDENT: **Did he change a lot from that first episode in future episodes?**

TEACHER:  Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc


STUDENT: **How has he changed?**

TEACHER:  Daffy was less anthropomorphic


STUDENT: **In what other ways did he change?**

TEACHER:  Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.




STUDENT: **Why did they add the lisp?**

TEACHER:  One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.

STUDENT: **Is there an "unofficial" story?**

TEACHER:  Yes, Mel Blanc (...) contradicts that conventional belief

...

Figure 1: An example dialog about a Wikipedia section. The student, who does not see the section text, asks questions. The teacher provides a response in the form of a text span (or **No answer**), optionally yes or no (**Yes / No**), and encouragement about continuing a line of questioning (should, , could , or should not  ask a follow-up question).

# More Ideas

	Train	Dev.	Test	Overall
questions	83,568	7,354	7,353	98,407
dialogs	11,567	1,000	1,002	13,594
unique sections	6,843	1,000	1,002	8,854
tokens / section	396.8	440.0	445.8	401.0
tokens / question	6.5	6.5	6.5	6.5
tokens / answer	15.1	12.3	12.3	14.6
questions / dialog	7.2	7.4	7.3	7.2
% yes/no	26.4	22.1	23.4	25.8
% unanswerable	20.2	20.2	20.1	20.2

Table 2: Statistics summarizing the 🦉 dataset.

# More Ideas


Dataset	Multi turn	Text- based	Dialog Acts	Simple Evaluation	Unanswerable Questions	Asker Can't See Evidence
 QuAC	✓	✓	✓	✓	✓	✓
CoQA (Reddy et al., 2018)	✓	✓	✗	✓	✓	✗
CSQA (Saha et al., 2018)	✓	✗	✗	✗	✓	✗
CQA (Talmor and Berant, 2018)	✓	✓	✗	✓	✗	✓
SQA (Iyyer et al., 2017)	✓	✗	✗	✓	✗	✗
NarrativeQA (Kociský et al., 2017)	✗	✓	✗	✗	✗	✓
TriviaQA (Joshi et al., 2017)	✗	✓	✗	✓	✗	✓
SQuAD 2.0 (Rajpurkar et al., 2018)	✗	✓	✗	✓	✓	✗
MS Marco (Nguyen et al., 2016)	✗	✓	✗	✗	✓	✓
NewsQA (Trischler et al., 2016)	✗	✓	✗	✓	✓	✓

Table 1: Comparison of the QUAC dataset to other question answering datasets.



# More Ideas

## Natural Questions: A Benchmark for Question Answering Research

Tom Kwiatkowski♣♦♠ Jennimaria Palomaki♠ Olivia Redfield♦♠ Michael Collins♣♦♠♥  
Ankur Parikh♥ Chris Alberti♥ Danielle Epstein♠♦ Illia Polosukhin♠♦ Jacob Devlin♠  
Kenton Lee♥ Kristina Toutanova♥ Llion Jones♠ Matthew Kelcey♠♦ Ming-Wei Chang♥  
Andrew M. Dai♠♦ Jakob Uszkoreit♠ Quoc Le♠♦ Slav Petrov♠

Google Research

*natural-questions@google.com*



# More Ideas

## **Abstract**

We present the Natural Questions corpus, a question answering data set. Questions consist of real anonymized, aggregated queries issued to the Google search engine. An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present. The public release consists of 307,373 training examples with single annotations; 7,830 examples with 5-way annotations for development data; and a further 7,842 examples with 5-way annotated sequestered as test data. We present

# More Ideas

## **Example 1**

**Question:** what color was john wilkes booth's hair

**Wikipedia Page:** John\_Wilkes\_Booth

**Long answer:** Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

**Short answer:** jet-black

# More Ideas

## **Example 2**

**Question:** can you make and receive calls in airplane mode

**Wikipedia Page:** Airplane\_mode

**Long answer:** Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

**Short answer:** BOOLEAN:NO

# More Ideas

## **Example 3**

**Question:** why does queen elizabeth sign her name elizabeth r

**Wikipedia Page:** Royal\_sign-manual

**Long answer:** The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

**Short answer:** NULL

# More Ideas

## **BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions**

**Christopher Clark<sup>\*1</sup>, Kenton Lee<sup>†</sup>, Ming-Wei Chang<sup>†</sup>, Tom Kwiatkowski<sup>†</sup>**

**Michael Collins<sup>†2</sup>, Kristina Toutanova<sup>†</sup>**

# More Ideas

---

<b>Q:</b>	Has the UK been hit by a hurricane?
<b>P:</b>	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
<b>A:</b>	Yes. [An example event is given.]
<b>Q:</b>	Does France have a Prime Minister and a President?
<b>P:</b>	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
<b>A:</b>	Yes. [Both are mentioned, so it can be inferred both exist.]
<b>Q:</b>	Have the San Jose Sharks won a Stanley Cup?
<b>P:</b>	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016
	...
<b>A:</b>	No. [They were in the finals once, and lost.]

---

Figure 1: Example yes/no questions from the BoolQ dataset. Each example consists of a question (**Q**), an excerpt from a passage (**P**), and an answer (**A**) with an explanation added for clarity.

# More Ideas

Reasoning Types	Yes/No Question Answering Examples	
<b>Paraphrasing (38.7%)</b> The passage explicitly asserts or refutes what is stated in the question.	<b>Q:</b> Is Tim Brown in the Hall of Fame? <b>P:</b> Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame. <b>A:</b> Yes. ["inducted into" directly implies he is in Hall of Fame.]	
<b>By Example (11.8%)</b> The passage provides an example or counter-example to what is asserted by the question.	<b>Q:</b> Are there any nuclear power plants in Michigan? <b>P:</b> ... three nuclear power plants supply Michigan with about 30% of its electricity. <b>A:</b> Yes. [Since there must be at least three.]	
<b>Factual Reasoning (8.5%)</b> Answering the question requires using world-knowledge to connect what is stated in the passage to the question.	<b>Q:</b> Was designated survivor filmed in the White House? <b>P:</b> The series is... filmed in Toronto, Ontario. <b>A:</b> No. [The White House is not located in Toronto.]	
<b>Implicit (8.5%)</b> The passage mentions or describes entities in the question in way that would not make sense if the answer was not yes/no.	<b>Q:</b> Is static pressure the same as atmospheric pressure? <b>P:</b> The aircraft designer's objective is to ensure the pressure in the aircraft's static pressure system is as close as possible to the atmospheric pressure... <b>A:</b> No. [It would not make sense to bring them "as close as possible" if those terms referred to the same thing.]	

# More Ideas

---

**Missing Mention (6.6%)**

We can conclude the answer is yes or no because, if this was not the case, it would have been mentioned in the passage.

---

**Q:** Did Bonnie Blair's daughter make the Olympic team?**P:** Blair and Cruikshank have two children: a son, Grant, and daughter, Blair.... Blair Cruikshank competed at the 2018 United States Olympic speed skating trials at the 500 meter distance.**A:** No. [The passage describes Blair Cruikshank's daughter's skating accomplishments, so it would have mentioned it if she had qualified.]

---

**Other Inference (25.9%)**

The passage states a fact that can be used to infer whether the answer is true or false, and does not fall into any of the other categories.

---

**Q:** Is the sea snake the most venomous snake?**P:** ... the venom of the inland taipan, drop by drop, is the most toxic among all snakes**A:** No. [If inland taipan is the most venomous snake, the sea snake must not be.]



# More Ideas

## **Temporal Common Sense Acquisition with Minimal Supervision**

**Ben Zhou,<sup>1</sup> Qiang Ning,<sup>2</sup> Daniel Khashabi,<sup>2</sup> Dan Roth<sup>1</sup>**

<sup>1</sup>University of Pennsylvania, <sup>2</sup>Allen Institute for AI

{xyzhou, danroth}@cis.upenn.edu    {qiangn, danielk}@allenai.org

# More Ideas

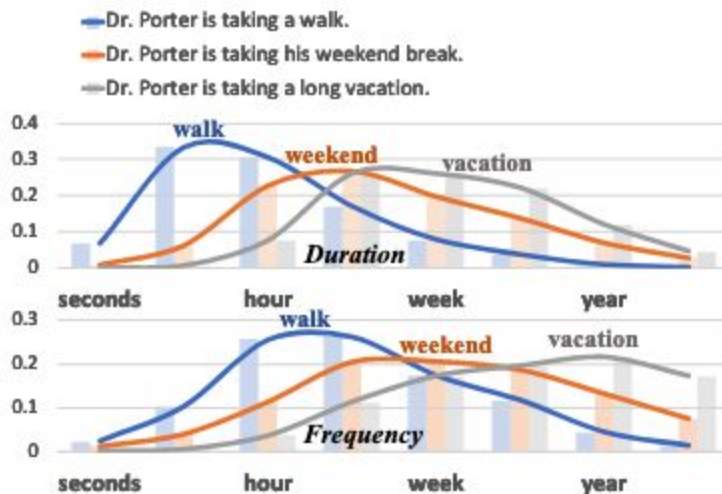


Figure 1: Our model's predicted distributions of event **duration** and **frequency**. The model is able to attend to contextual information and thus produce reasonable estimates.

# More Ideas

- Model Builders vs. Model Breakers
  - Model Builder: build a model to solve a problem
    - success: better results (P, R, F) regardless of the reasons or potential issues in the data you work with
    - I am better if I get better results
  - Model Breaker: figure out when models fail, point out issues either with the model or the benchmark
    - success:
      - show that for some inputs the model gets ridiculously worse or ridiculously better results (P, R, F) [and then fix it]
      - show that simpler models (less parameters, easier to implement, faster to run) get the same or even better results

# More Ideas - A Few Interesting Papers

- Misspelling do break models (surprisingly?)

## **Combating Adversarial Misspellings with Robust Word Recognition**

**Danish Pruthi**

**Bhuwan Dhingra**

**Zachary C. Lipton**

Carnegie Mellon University  
Pittsburgh, USA

`{ddanish, bdhingra}@cs.cmu.edu, zlipton@cmu.edu`

# More Ideas - A Few Interesting Papers

- Misspelling do break models (surprisingly?)

Alteration	Movie Review	Label
Original	A triumph, relentless and beautiful in its downbeat darkness	+
Swap	A triumph, relentless and <b>beuatiful</b> in its downbeat darkness	-
Drop	A triumph, relentless and beautiful in its <b>dwnbeat</b> darkness	-
+ Defense	A triumph, relentless and <b>beautiful</b> in its downbeat darkness	+
+ Defense	A triumph, relentless and beautiful in its <b>downbeat</b> darkness	+

Table 1: Adversarial spelling mistakes inducing sentiment misclassification and word-recognition defenses.

# More Ideas - A Few Interesting Papers

- Sometimes you can solve a problem without feeding to the model the full input
  - meaning that there is something suspicious about the data or the problem definition
  - SNLI corpus (for Natural Language Inference):

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

# More Ideas - A Few Interesting Papers

- It turns out that feeding only the hypothesis gets good results

## **Hypothesis Only Baselines in Natural Language Inference**

**Adam Poliak<sup>1</sup> Jason Naradowsky<sup>1</sup> Aparajita Haldar<sup>1,2</sup>**

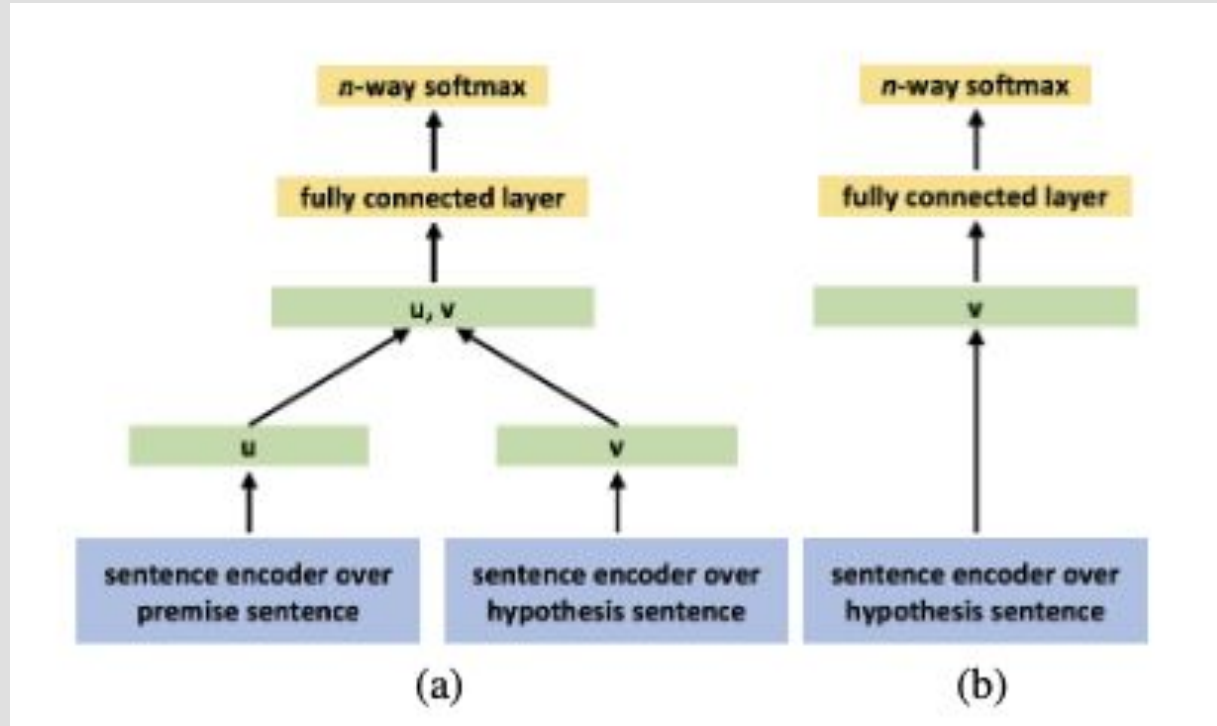
**Rachel Rudinger<sup>1</sup> Benjamin Van Durme<sup>1</sup>**

<sup>1</sup>Johns Hopkins University <sup>2</sup>BITS Pilani, Goa Campus, India

`{azpoliak, vandurme}@cs.jhu.edu {narad, ahaldar1, rudinger}@jhu.edu`

# More Ideas - A Few Interesting Papers

- It turns out that feeding only the hypothesis gets good results





Dataset	DEV				TEST				Baseline	SOTA
	Hyp-Only	MAJ	$ \Delta $	$\Delta\%$	Hyp-Only	MAJ	$ \Delta $	$\Delta\%$		
Recast										
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
Human Judged										
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
Human Elicited										
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	–	–	72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Table 2: NLI accuracies on each dataset. Columns ‘Hyp-Only’ and ‘MAJ’ indicates the accuracy of the hypothesis-only model and the majority baseline.  $|\Delta|$  and  $\Delta\%$  indicate the absolute difference in percentage points and the percentage increase between the Hyp-Only and MAJ. Blue numbers indicate that the hypothesis-model outperforms MAJ. In the right-most section, ‘Baseline’ indicates the original baseline on the test when the dataset was released and ‘SOTA’ indicates current state-of-the-art results. MNLI-1 is the matched version and MNLI-2 is the mismatched for MNLI. The names of datasets are italicized if containing  $\leq 10K$  labeled examples.

Dataset	DEV				TEST				Baseline	SOTA
	Hyp-Only	MAJ	$ \Delta $	$\Delta\%$	Hyp-Only	MAJ	$ \Delta $	$\Delta\%$		
Recast										
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
Human Judged										
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	—	—
Human Elicited										
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+36.01	—	35.6	—	—	72.5	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	—	36.5	—	—	72.1	83.21

Table 2: NLI accuracies on each dataset. Columns ‘Hyp-Only’ and ‘MAJ’ indicates the accuracy of the hypothesis-only model and the majority baseline.  $|\Delta|$  and  $\Delta\%$  indicate the absolute difference in percentage points and the percentage increase between the Hyp-Only and MAJ. Blue numbers indicate that the hypothesis-model outperforms MAJ. In the right-most section, ‘Baseline’ indicates the original baseline on the test when the dataset was released and ‘SOTA’ indicates current state-of-the-art results. MNLI-1 is the matched version and MNLI-2 is the mismatched for MNLI. The names of datasets are italicized if containing  $\leq 10K$  labeled examples.

# More Ideas - A Few Interesting Papers

- Natural Language Inference and negation

## **An Analysis of Natural Language Inference Benchmarks through the Lens of Negation**

**Md Mosharaf Hossain,<sup>°</sup> Venelin Kovatchev,<sup>3</sup> Pranoy Dutta,<sup>°</sup> Tiffany Kao,<sup>°</sup>  
Elizabeth Wei,<sup>°</sup> and Eduardo Blanco<sup>°</sup>**

<sup>°</sup>University of North Texas      <sup>3</sup>University of Barcelona

mdmosharafhossain@my.unt.edu      vkovatchev@ub.edu

{PranoyDutta,TiffanyKao,ElizabethWei}@my.unt.edu      eduardo.blanco@unt.edu

# More Ideas - A Fe

- Natural Language I

	#sents.	% w/ neg.
General English		
Online Reviews		
books	4,845,154	22.64
movies	616,287	28.97
Conversations		
oral	538,973	27.43
written	510,458	29.92
Wikipedia	2,735,930	8.69
Books	1,809,184	28.45
OntoNotes	63,918	17.14
NLI benchmarks		
RTE	16,389	7.16
SNLI	1,138,598	1.19
MNLI	883,436	22.63

Table 1: Percentage of sentences containing negation in general-purpose English corpora (reviews, conversations, Wikipedia, books and OntoNotes) and existing natural language inference benchmarks (also in English). Negation is underrepresented in RTE and SNLI.

# More Ideas - A Few Interesting Papers

- Natural Language Inference and negation

MNLI

7) T: It was summertime the air conditioner was on the door was closed and i couldn't knock because i had to hold the jack with the other hand i finally with my elbow rang the doorbell and mother came to the door.

H: The wintertime is when the air conditioning was on, I couldn't ring the doorbell because it was frozen.

8) T: It runs advertisements for its supporters at the top of shows and strikes business deals with MCI, TCI, and Disney, but still insists it's not commercial.

H: It runs ads for its supporters at shows and strikes business deals, but insists it is not commercial.

Table 2: Examples of the few text-hypothesis pairs that contain negation in the three natural language inference corpora we work with (RTE, SNLI and MNLI). Negation cues are underlined, and we have made minimal edits to some examples so that they fit within the width of the table.

# More Ideas - A Few Interesting Papers

- Natural Language Inference and negation

	Original pair	New pair w/ negation
RTE	<p>T: Tropical Storm Debby is blamed for several deaths across the Caribbean.</p> <p>H: A tropical storm has caused loss of life.</p> <p><i>Judgments:</i> T-H: entailment, <math>T_{neg}</math>-H: no_entailment, T-<math>H_{neg}</math>: no_entailment, <math>T_{neg}</math>-<math>H_{neg}</math>: entailment</p>	<p><math>T_{neg}</math>: Tropical Storm Debby is not blamed for several deaths across the Caribbean.</p> <p><math>H_{neg}</math>: A tropical storm has not caused loss of life.</p>
	<p>T: Dr. Pridi was forced into exile, and Field Marshal Pibul again assumed power.</p> <p>H: Pibul was a field marshal.</p> <p><i>Judgments:</i> T-H: entailment, <math>T_{neg}</math>-H: entailment, T-<math>H_{neg}</math>: no_entailment, <math>T_{neg}</math>-<math>H_{neg}</math>: no_entailment</p>	<p><math>T_{neg}</math>: Dr. Pridi was not forced into exile, and Field Marshal Pibul again assumed power.</p> <p><math>H_{neg}</math>: Pibul was not a field marshal.</p>



# More Ideas - A Few Interesting Papers

- Natural Language Inference and negation

Test pairs	RTE				SNLI				MNLI			
	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]
Original												
dev	52.7	75.8	69.9	66.1	33.8	91.6	90.6	89.9	35.5	87.9	86.7	83.2
dev <sub>neg</sub>	51.2	78.1	73.2	63.4	54.4	91.7	90.3	89.4	50.2	88.0	86.7	83.0
New w/ neg.												
T <sub>neg</sub> -H	80.2	70.8	69.0	65.2	62.0	46.4	39.8	32.6	45.8	66.2	63.8	65.6
T-H <sub>neg</sub>	91.0	51.4	44.2	39.2	41.0	63.6	67.4	58.8	47.6	70.4	69.8	62.4
T <sub>neg</sub> -H <sub>neg</sub>	65.6	65.4	69.6	68.4	69.8	45.8	47.2	41.8	47.0	63.6	65.4	63.6
All	78.9	62.5	60.9	57.6	56.5	51.9	51.5	44.4	39.3	66.7	66.3	63.9

Table 7: Results obtained with state-of-the-art models trained with the original training split for each benchmark and evaluated with (a) the original development split (dev), (b) pairs in the original development split containing negation (dev<sub>neg</sub>), and (c) the new pairs containing negation. MB stands for the majority baseline, [1] for RoBERTa (Liu et al., 2019), [2] for XLNet (Yang et al., 2019) and [3] for BERT (Devlin et al., 2019).

# More Ideas - A Few Interesting Papers

- Natural Language Inference and negation

Test pairs	RTE				SNLI				MNLI			
	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]
Original												
dev	52.7	75.8	69.9	66.1	33.8	91.6	90.6	89.9	35.5	87.9	86.7	83.2
dev <sub>neg</sub>	51.2	78.1	73.2	63.4	54.4	91.7	90.3	89.4	50.2	88.0	86.7	83.0
New w/ neg.												
T <sub>neg</sub> -H	80.2	70.8	69.0	65.2	62.0	46.4	39.8	32.6	45.8	66.2	63.8	65.6
T-H <sub>neg</sub>	91.0	51.4	44.2	39.2	41.0	63.6	67.4	58.8	47.6	70.4	69.8	62.4
T <sub>neg</sub> -H <sub>neg</sub>	65.6	65.4	69.6	68.4	69.8	45.8	47.2	41.8	47.0	63.6	65.4	63.6
All	78.9	62.5	60.9	57.6	56.5	51.9	51.5	44.4	39.3	66.7	66.3	63.9

Table 7: Results obtained with state-of-the-art models trained with the original training split for each benchmark and evaluated with (a) the original development split (dev), (b) pairs in the original development split containing negation (dev<sub>neg</sub>), and (c) the new pairs containing negation. MB stands for the majority baseline, [1] for RoBERTa (Liu et al., 2019), [2] for XLNet (Yang et al., 2019) and [3] for BERT (Devlin et al., 2019).



# More Ideas - A Few Interesting Papers

- You do not need complicated networks (or at least not always)

## **Deep Unordered Composition Rivals Syntactic Methods for Text Classification**

**Mohit Iyyer,<sup>1</sup> Varun Manjunatha,<sup>1</sup> Jordan Boyd-Graber,<sup>2</sup> Hal Daumé III<sup>1</sup>**

<sup>1</sup>University of Maryland, Department of Computer Science and UMIACS

<sup>2</sup>University of Colorado, Department of Computer Science

`{miyyer, varunm, hal}@umiacs.umd.edu, Jordan.Boyd.Grabber@colorado.edu`

## Abstract

Many existing deep learning models for natural language processing tasks focus on learning the *compositionality* of their inputs, which requires many expensive com-

putations. We present a simple deep neural network that competes with and, in some cases, outperforms such models on sentiment analysis and factoid question answering tasks while taking only a fraction of the training time. While our model is syntactically-ignorant, we show significant improvements over previous bag-of-words models by deepening our network and applying a novel variant of dropout. Moreover, our model performs better than syntactic models on datasets with high syntactic variance. We show that our model

makes similar errors to syntactically-aware models, indicating that for the tasks we consider, nonlinearly transforming the input is more important than tailoring a network to incorporate word order and syntax.

## Interesting Papers

and networks (or at least not always)

This model, the deep averaging network (DAN), works in three simple steps:

1. take the vector average of the embeddings associated with an input sequence of tokens
2. pass that average through one or more feed-forward layers
3. perform (linear) classification on the final layer's representation

# More Ideas - A Few Interesting Papers

- 

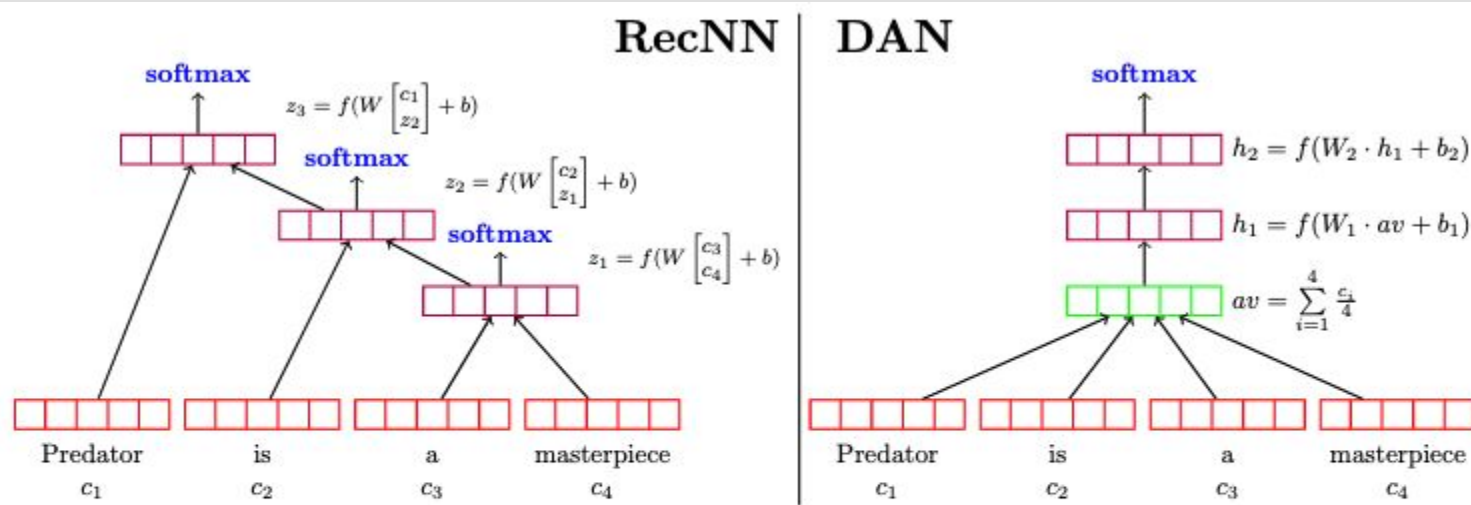


Figure 1: On the left, a **RecNN** is given an input sentence for sentiment classification. Softmax layers are placed above every internal node to avoid vanishing gradient issues. On the right is a two-layer **DAN** taking the same input. While the **RecNN** has to compute a nonlinear representation (purple vectors) for every node in the parse tree of its input, this **DAN** only computes two nonlinear layers for every possible input.

## More Ideas - A Few

- You do not need co

Model	RT	SST fine	SST bin	IMDB	Time (s)
DAN-ROOT	—	46.9	85.7	—	<b>31</b>
DAN-RAND	77.3	45.4	83.2	88.8	136
DAN	80.3	47.7	86.3	89.4	136
NBOW-RAND	76.2	42.3	81.4	88.9	91
NBOW	79.0	43.6	83.6	89.0	91
BiNB	—	41.9	83.1	—	—
NBSVM-bi	79.4	—	—	91.2	—
RecNN*	77.7	43.2	82.4	—	—
RecNTN*	—	45.7	85.4	—	—
DRecNN	—	49.8	86.6	—	431
TreeLSTM	—	<b>50.6</b>	86.9	—	—
DCNN*	—	48.5	86.9	89.4	—
PVEC*	—	48.7	87.8	<b>92.6</b>	—
CNN-MC	<b>81.1</b>	47.4	<b>88.1</b>	—	2,452
WRRBM*	—	—	—	89.2	—

Table 1: DANs achieve comparable sentiment accuracies to syntactic functions (bottom third of table) but require much less training time (measured as time of a single epoch on the SST fine-grained task). Asterisked models are initialized either with different pretrained embeddings or randomly.

ways)



# Looking at the output of the models

-

# Error Analysis

- Regardless of the problem you work with, you need to report more than a table with numbers
  - quantitative vs. qualitative analyses, error analyses
- You cannot make up the numbers, but you can check that they make sense
  - When does the model make the most errors?

# Error Analysis

- Fashion MNIST: 10,000 images in the test set
  - 28x28 grayscale image, 10 classes:
    - 0 T-shirt/top
    - 1 Trouser
    - 2 Pullover
    - 3 Dress
    - 4 Coat
    - 5 Sandal
    - 6 Shirt
    - 7 Sneaker
    - 8 Bag
    - 9 Ankle boot

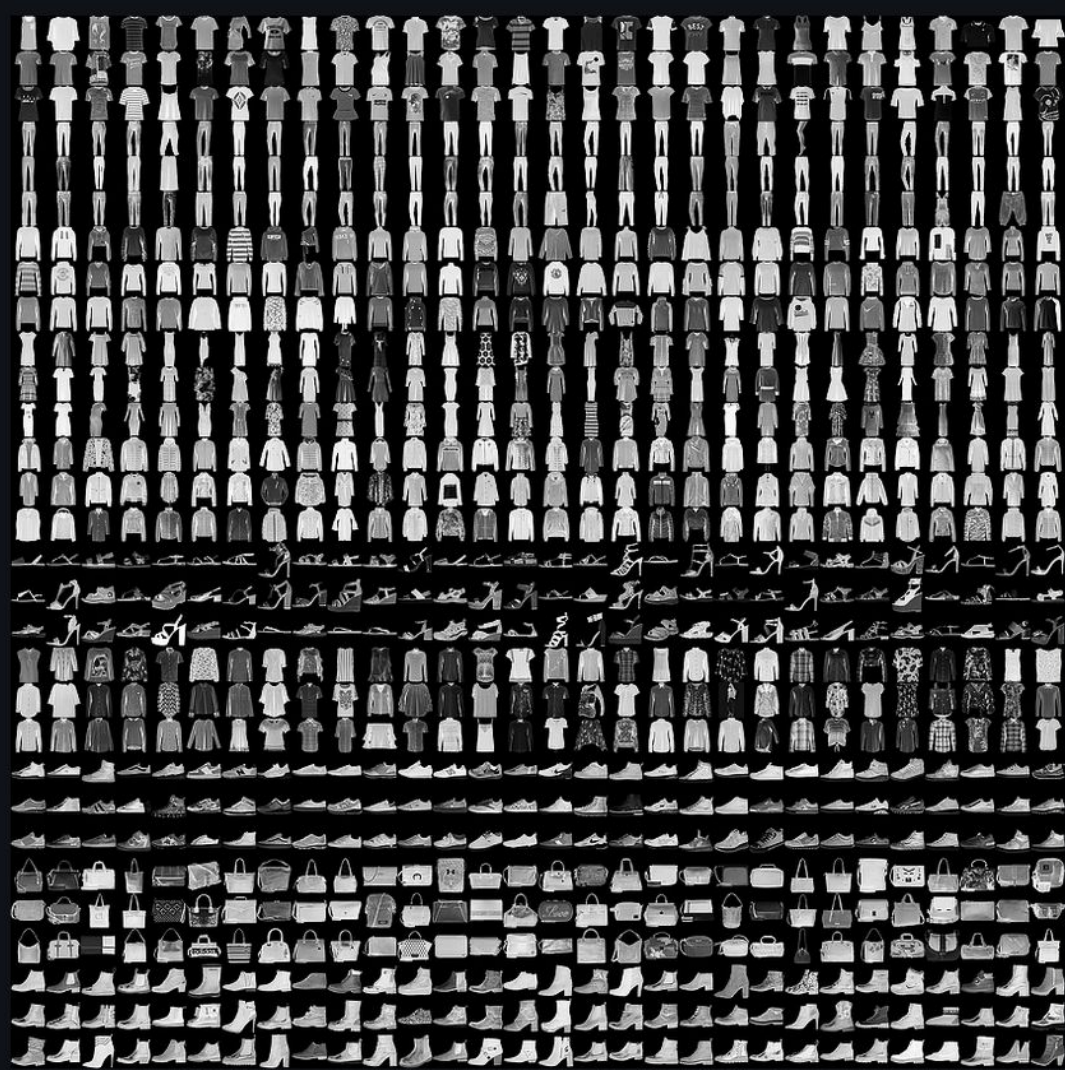


# Which classes do you think are harder?

- 0 T-shirt/top
- 1 Trouser
- 2 Pullover
- 3 Dress
- 4 Coat
- 5 Sandal
- 6 Shirt
- 7 Sneaker
- 8 Bag
- 9 Ankle boot

# Which classes do you thi

- 0 T-shirt/top
- 1 Trouser
- 2 Pullover
- 3 Dress
- 4 Coat
- 5 Sandal
- 6 Shirt
- 7 Sneaker
- 8 Bag
- 9 Ankle boot



# Quantitative Results

- c&p quantitative results is not needed but not enough.
  - ok with a baseline for the proposal

	precision	recall	f1-score	support
T-shirt/top	0.86	0.83	0.85	1000
Trouser	1.00	0.97	0.99	1000
Pullover	0.83	0.88	0.85	1000
Dress	0.85	0.95	0.90	1000
Coat	0.84	0.83	0.84	1000
Sandal	0.94	0.99	0.96	1000
Shirt	0.78	0.68	0.73	1000
Sneaker	0.90	0.97	0.94	1000
Bag	0.99	0.97	0.98	1000
Ankle boot	1.00	0.88	0.93	1000
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000

# Qualitative Results

- Shirts are hard
- Ankle boots are hard
- Sandals are easy??

