ML Application with BERT

- **CYsyphus Mission**
  - Identify, extract, and categorize recommendations from reports
  - Require good understanding and representation from Language Models
- **BERT**
  - Bidirectional Encoder Representations from Transformers
  - State-of-the-art Language Model

Fine-Tuning BERT

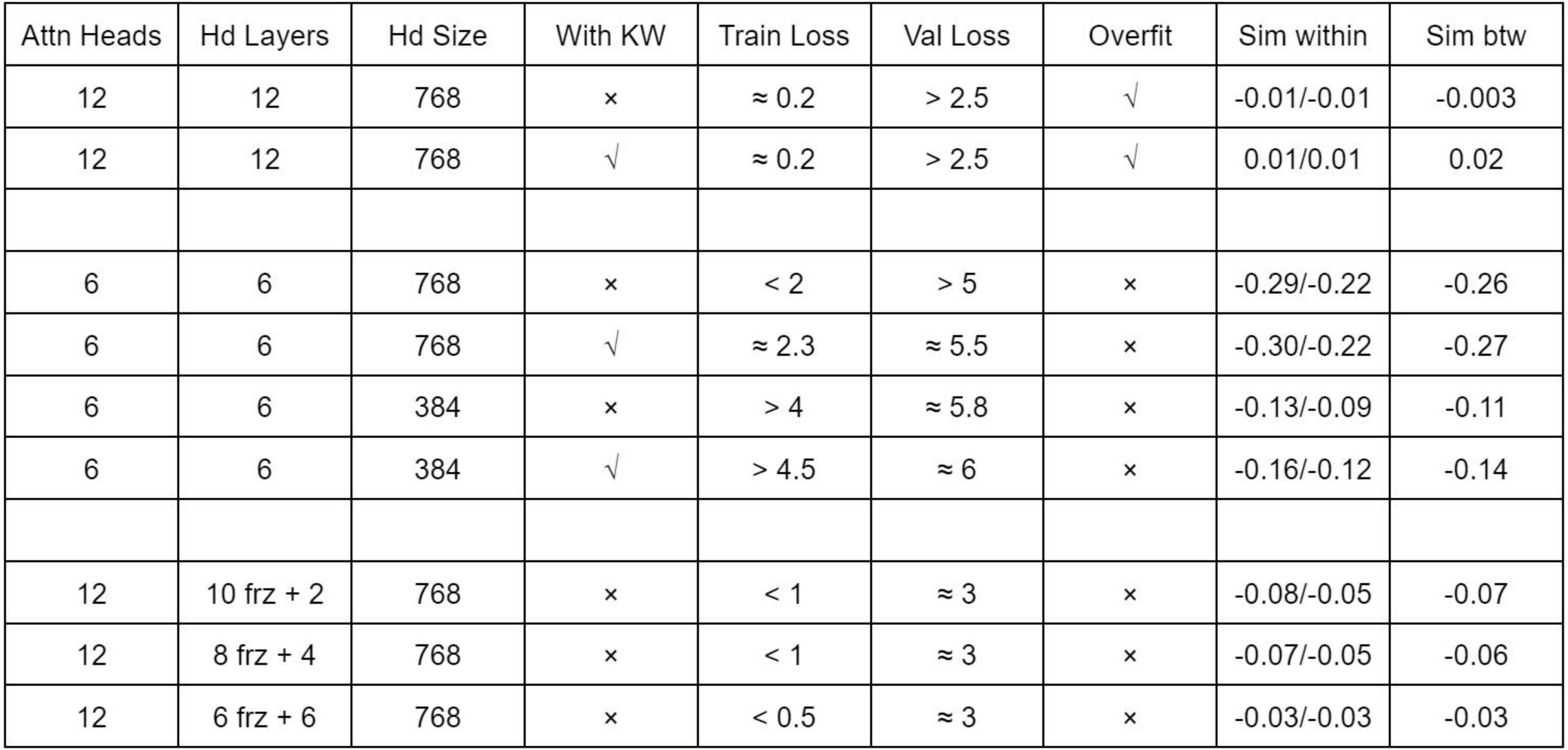
- **Question**
  - Current workflow as well as the original BERT may be too generalized
  - Recommendations in other fields, non-recommendations in cyber-security
- **Solution**
  - Fine-tune a BERT model based on our policy report corpus
  - Capture domain-specific semantic structure
  - Better identify cyber-security recommendations

Current Results

- Fine-tune on 15 policy documents, over 13,000 sentences
- Adjustments to dataset / model
  - Filter sentences with extracted keywords
  - Reduce model capacity
  - Gradually unfreeze BERT hidden layers from the top
- We get
  - Embedding similarity compared to original BERT
  - Extraction quality on an example doc based on clustering: TOP 25 precision

Current Results

- **Original size model:** default size and structure
    - Overfitting (in terms of BERT training obj.)
    - Little similarity difference & 36% precision
- **Reduced size model:** 6 heads, 6 layers, 384 hidden dims, with keyword filtering
    - Significant similarity difference (even within the same class) & 24% precision
- **Model with frozen layers:** 10 freeze + 2 unfreeze
    - Moderate similarity difference & **40% precision**



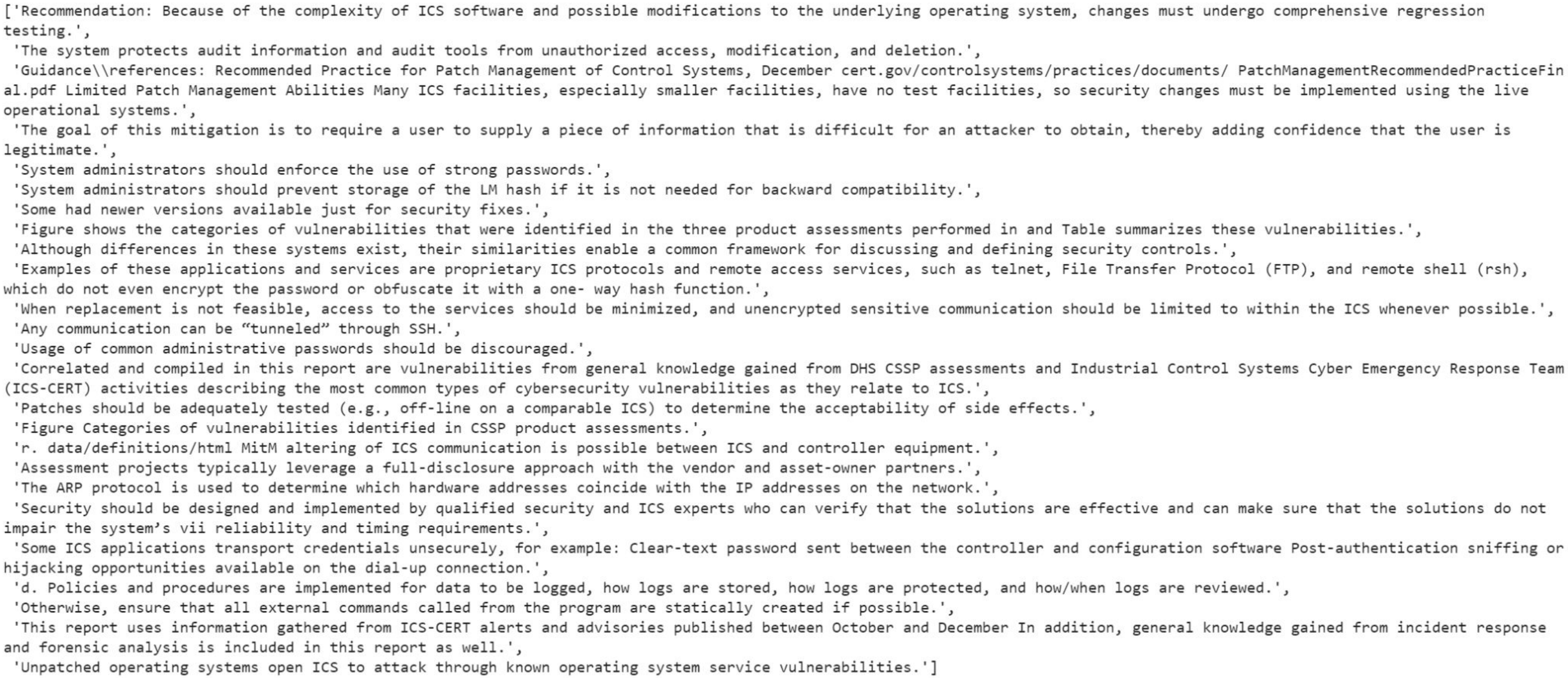
Attn Heads	Hd Layers	Hd Size	With KW	Train Loss	Val Loss	Overfit	Sim within	Sim btw
12	12	768	×	≈ 0.2	> 2.5	√	-0.01/-0.01	-0.003
12	12	768	√	≈ 0.2	> 2.5	√	0.01/0.01	0.02
6	6	768	×	< 2	> 5	×	-0.29/-0.22	-0.26
6	6	768	√	≈ 2.3	≈ 5.5	×	-0.30/-0.22	-0.27
6	6	384	×	> 4	≈ 5.8	×	-0.13/-0.09	-0.11
6	6	384	√	> 4.5	≈ 6	×	-0.16/-0.12	-0.14
12	10 frz + 2	768	×	< 1	≈ 3	×	-0.08/-0.05	-0.07
12	8 frz + 4	768	×	< 1	≈ 3	×	-0.07/-0.05	-0.06
12	6 frz + 6	768	×	< 0.5	≈ 3	×	-0.03/-0.03	-0.03

Extraction Exemplars: 1-5

- (√) Recommendation: Because of the complexity of ICS software and possible modifications to the underlying operating system, changes must undergo comprehensive regression testing.
- (×) The system protects audit information and audit tools from unauthorized access, modification, and deletion.
- (√) Guidance\\references: Recommended Practice for Patch Management of Control Systems, December cert.gov/controlsystems/practices/documents/ PatchManagementRecommendedPracticeFin al.pdf Limited Patch Management Abilities Many ICS facilities, especially smaller facilities, have no test facilities, so security changes must be implemented using the live operational systems.
- (×) The goal of this mitigation is to require a user to supply a piece of information that is difficult for an attacker to obtain, thereby adding confidence that the user is legitimate.
- (√) System administrators should enforce the use of strong passwords.

Extraction Exemplars: 6-10

- (√) System administrators should prevent storage of the LM hash if it is not needed for backward compatibility.
- (×) Some had newer versions available just for security fixes.
- (×) Figure shows the categories of vulnerabilities that were identified in the three product assessments performed in and Table summarizes these vulnerabilities.
- (×) Although differences in these systems exist, their similarities enable a common framework for discussing and defining security controls.
- (×) Examples of these applications and services are proprietary ICS protocols and remote access services, such as telnet, File Transfer Protocol (FTP), and remote shell (rsh), which do not even encrypt the password or obfuscate it with a one- way hash function.



```
→ ['Recommendation: Because of the complexity of ICS software and possible modifications to the underlying operating system, changes must undergo comprehensive regression
  testing.',
   'The system protects audit information and audit tools from unauthorized access, modification, and deletion.',
→  'Guidance\\references: Recommended Practice for Patch Management of Control Systems, December cert.gov/controlsystems/practices/documents/ PatchManagementRecommendedPracticeFin
  al.pdf Limited Patch Management Abilities Many ICS facilities, especially smaller facilities, have no test facilities, so security changes must be implemented using the live
  operational systems.',
   'The goal of this mitigation is to require a user to supply a piece of information that is difficult for an attacker to obtain, thereby adding confidence that the user is
  legitimate.',
→  'System administrators should enforce the use of strong passwords.',
→  'System administrators should prevent storage of the LM hash if it is not needed for backward compatibility.',
   'Some had newer versions available just for security fixes.',
   'Figure shows the categories of vulnerabilities that were identified in the three product assessments performed in and Table summarizes these vulnerabilities.',
   'Although differences in these systems exist, their similarities enable a common framework for discussing and defining security controls.',
   'Examples of these applications and services are proprietary ICS protocols and remote access services, such as telnet, File Transfer Protocol (FTP), and remote shell (rsh),
  which do not even encrypt the password or obfuscate it with a one- way hash function.',
→  'When replacement is not feasible, access to the services should be minimized, and unencrypted sensitive communication should be limited to within the ICS whenever possible.',
   'Any communication can be “tunneled” through SSH.',
→  'Usage of common administrative passwords should be discouraged.',
   'Correlated and compiled in this report are vulnerabilities from general knowledge gained from DHS CSSP assessments and Industrial Control Systems Cyber Emergency Response Team
  (ICS-CERT) activities describing the most common types of cybersecurity vulnerabilities as they relate to ICS.',
→  'Patches should be adequately tested (e.g., off-line on a comparable ICS) to determine the acceptability of side effects.',
   'Figure Categories of vulnerabilities identified in CSSP product assessments.',
   'r. data/definitions/html MitM altering of ICS communication is possible between ICS and controller equipment.',
→  'Assessment projects typically leverage a full-disclosure approach with the vendor and asset-owner partners.',
   'The ARP protocol is used to determine which hardware addresses coincide with the IP addresses on the network.',
→  'Security should be designed and implemented by qualified security and ICS experts who can verify that the solutions are effective and can make sure that the solutions do not
  impair the system’s vii reliability and timing requirements.',
   'Some ICS applications transport credentials unsecurely, for example: Clear-text password sent between the controller and configuration software Post-authentication sniffing or
  hijacking opportunities available on the dial-up connection.',
   'd. Policies and procedures are implemented for data to be logged, how logs are stored, how logs are protected, and how/when logs are reviewed.',
→  'Otherwise, ensure that all external commands called from the program are statically created if possible.',
   'This report uses information gathered from ICS-CERT alerts and advisories published between October and December In addition, general knowledge gained from incident response
  and forensic analysis is included in this report as well.',
   'Unpatched operating systems open ICS to attack through known operating system service vulnerabilities.']
```

Next Steps

- **Unsupervised clustering** (so far)
    - A reasonable number of clusters
- **Further studying unfreezing approach**
- **Supervised modeling**
    - Classify recommendations with RNN-type models
- **(Semantic) structure search**
    - [ORG] recommend [noun] should [verb] …

Workflow 1

- Get sentence embeddings for all recommendations from training documents
- Calculate the mean recommendation embedding (MRE)
- Calculate the distance of every sentence in test documents to MRE
    - Distance metric: cosine similarity
- Objective: maximize mean cosine similarity of test recommendations with MRE
    - Mean test cosine similarity: 0.7971

Workflow 2*

- Get sentence embeddings for all recommendations from training documents
- Find centroids of **n** (= 2) clusters
- Use modal verbs to filter sentences from test documents
  - can, could, may, might, must, shall, should, will, would, ensure
  - about 28% of recommendations in training set do NOT have these verbs
- Calculate the distance of candidate sentences to each centroid
  - distance metric: cosine similarity
  - consider the larger of cosine similarities with each centroid
- Objective: maximize mean cosine similarity of test recommendations with either centroids
  - Mean test cosine similarity: **0.8069**

* from last summer

Some predictions from Workflow 2 (missed by us)

- Federal departments and agencies should give substantial consideration to the acquisition and use of security-related IT products and services that are compatible with the CVE vulnerability naming scheme.
- While CVE compatibility should be an important consideration in IT security product and service acquisition, federal departments and agencies should foremost consider their overall requirements (functionality, cost, performance, architecture, etc.)
- It could provide the basis for meaningful dialogues on principles of responsibility and accountability to manage what Joseph Nye Jr. has called their “cooperative rivalry,” and then on how to deal with potential cyber operations against capabilities that could affect NC3.
- Agencies should use CVE in their internal reports of vulnerability scans, notifications to system owners of observed vulnerabilities, and alerts identifying the vulnerabilities targeted by active exploits.

Work in Progress

- Using Distilbert to train a binary classifier
- Data Augmentation to improve performance
    - Back-translation with Russian and German
    - TF-IDF Word Replacement
- Will also use some Semi-supervised Learning techniques
    - UDA
    - MixText

Evaluation Metrics

- We need a way to compare different modifications to the base workflow to understand which one is best (or good enough)
- Some ideas (likely will want to look at combination of these)
  - Mean cosine similarity of test exemplars to cluster centroids
  - Embedding similarity scores within and between classes
  - Model accuracy for supervised models
  - Hand evaluated extractions (person power intensive)
  - Other??