DISCOVERY OF NATURAL LANGUAGE CONCEPTS IN INDIVIDUAL UNITS OF CNNs

Seil Na, Yo Joong Choe, Dong-Hyun Lee, Gunhee Kim
ICLR 2019

February 14, 2020

Presenter: Rishab Bamrara

https://qdata.github.io/deep2Read/

Motivation:

- Although deep convolutional networks have achieved improved performance in many natural language tasks, they have been treated as black boxes because they are difficult to interpret.

- Especially, little is known about how they represent language in their intermediate layers.

- Because of their lack of interpretability, deep models are often regarded as hard to debug and unreliable for deployment.

- They also prevent the user from learning about how to make better decisions based on the model's outputs.

Related Work:

1. Zhou et al. (2015): Object Detectors Emerge in Deep Scene CNNs.

1. Erhan et al. (2009): Visualizing Higher-layer Features of a Deep Network.

1. Olah et al. (2017): Feature Visualization.

1. Simonyan et al. (2013): Visualising Image Classification Models and Saliency maps.

1. Radford et al. (2017): Concept of sentiment aligned to a particular unit.

Background:

- **Character level CNN:** Represent each character as a one-hot encoded vector.
- **1D Convolution:** Convolution takes place in only one direction. In NLP the direction is the time axis.
- **Unit:** Each channel in convolutional representation.
- **Natural Language Concepts:** Grammatical units of natural language that preserve meanings; i.e. morphemes, words, and phrases.

Claim / Target Task:

The units of deep CNNs learned in NLP tasks could act as a natural language concept detector.

Top K Activated Sentences Per Unit:

- Given a layer and sentence s ∈ S, let $A^l_u(s)$ denote the activation of unit u at spatial location l.

- Then, for unit u, average activations over all spatial locations as:
  $a_u(s) = \frac{1}{Z} \sum_l A^l_u(s)$, where Z is a normalizer.

- Retrieve top K training sentences per unit with the highest mean activation a_u.

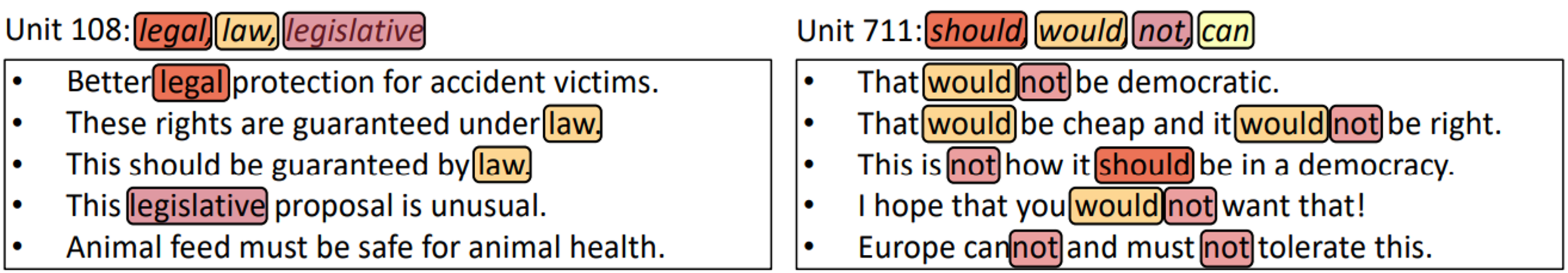


Figure 1: We discover the most activated sentences and aligned concepts to the units in hidden representations of deep convolutional networks. Aligned concepts appear frequently in most activated sentences, implying that those units respond selectively to specific natural language concepts.

Identifying Concepts:

1. Parse each of top K sentences with a constituency parser (Kitaev & Klein, 2018).

1. From sentence “John hit the balls”, we obtain candidate concepts as {John, hit, the, balls, the balls, hit the balls, John hit the balls}.

1. Also break each word into morphemes using a morphological analysis tool (Virpioja et al., 2013) and add them to candidate concepts (e.g. from word “balls”, we obtain morphemes {ball, s}).

$C_u = \{c_1, \ldots, c_N\}$, where N is the number of candidate concepts of the unit.

Measuring Contribution of each Concept:

1. For normalizing, create a synthetic sentence by replicating each candidate concept so that its length is identical to the average length of all training sentences.

   (e.g. candidate concept "the ball" is replicated as "the ball the ball the ball…")

1. Degree of alignment (DoA) between a candidate concept "c_n" and a unit "u" :

$$\mathbf{DoA_{u,cn} = a_u(r_n)}$$

1. DoA measures the extent to unit u's activation is sensitive to the presence of candidate concept c_n.

1. Larger the value suggests that candidate concept c_n is strongly aligned to unit u.

1. For each unit u, define a set of its aligned concepts $\mathbf{C^*_u = \{c^*_1 , \ldots, c^*_M\}}$ as M candidate concepts with the largest DoA values in C_u.

Datasets and Model Descriptions:

Dataset	Task	Model	# of Layers	# of Units
AG News	Ontology Classification	VDCNN	4	[64, 128, 256, 512]
DBpedia	Topic Classification	VDCNN	4	[64, 128, 256, 512]
Yelp Review	Polarity Classification	VDCNN	4	[64, 128, 256, 512]
WMT17' EN-DE	Translation	ByteNet	15	[1024] for all
WMT14' EN-FR	Translation	ByteNet	15	[1024] for all
WMT14' EN-CS	Translation	ByteNet	15	[1024] for all
EN-DE Europarl-v7	Translation	ByteNet	15	[1024] for all

Table 1: Datasets and model descriptions used in our analysis.

Evaluation of Concept Alignment:

Define the concept selectivity of a unit u, to a set of concepts C^*_u as follows:

$$\mathsf{Sel}_u = \frac{\mu_+ - \mu_-}{\max_{s \in \mathcal{S}} a_u(s) - \min_{s \in \mathcal{S}} a_u(s)} \tag{2}$$

where $\mathcal{S}$ denotes all sentences in training set, and $\mu_+ = \frac{1}{|\mathcal{S}_+|} \sum_{s \in \mathcal{S}_+} a_u(s)$ is the average value of unit activation when forwarding a set of sentences $\mathcal{S}_+$, which is defined as one of the following:

replicate: S+ contains the sentences created by replicating each concept in C^*_u.

one instance: S+ contains just one instance of each concept in C^*_u.

inclusion: S+ contains the training sentences that include at least one concept in C^*_u.

random: S+ contains randomly sampled sentences from the training data.

In contrast, **μ–** is the average value of unit activation when forwarding S–, which consists of training sentences that do not include any concept in $C * u$.

Evaluation of Concept Alignment:

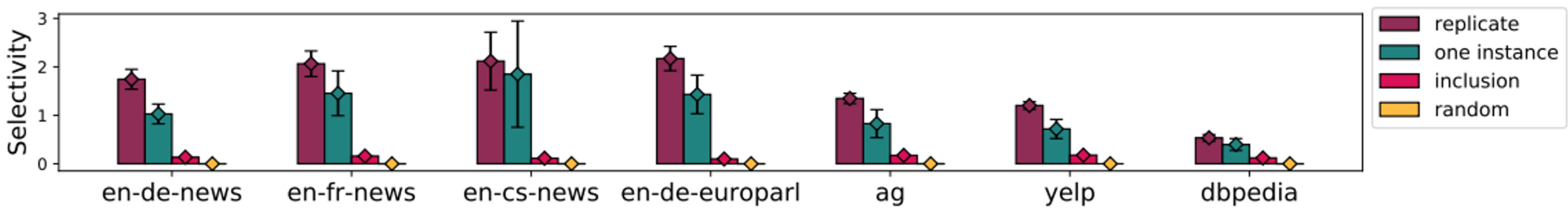
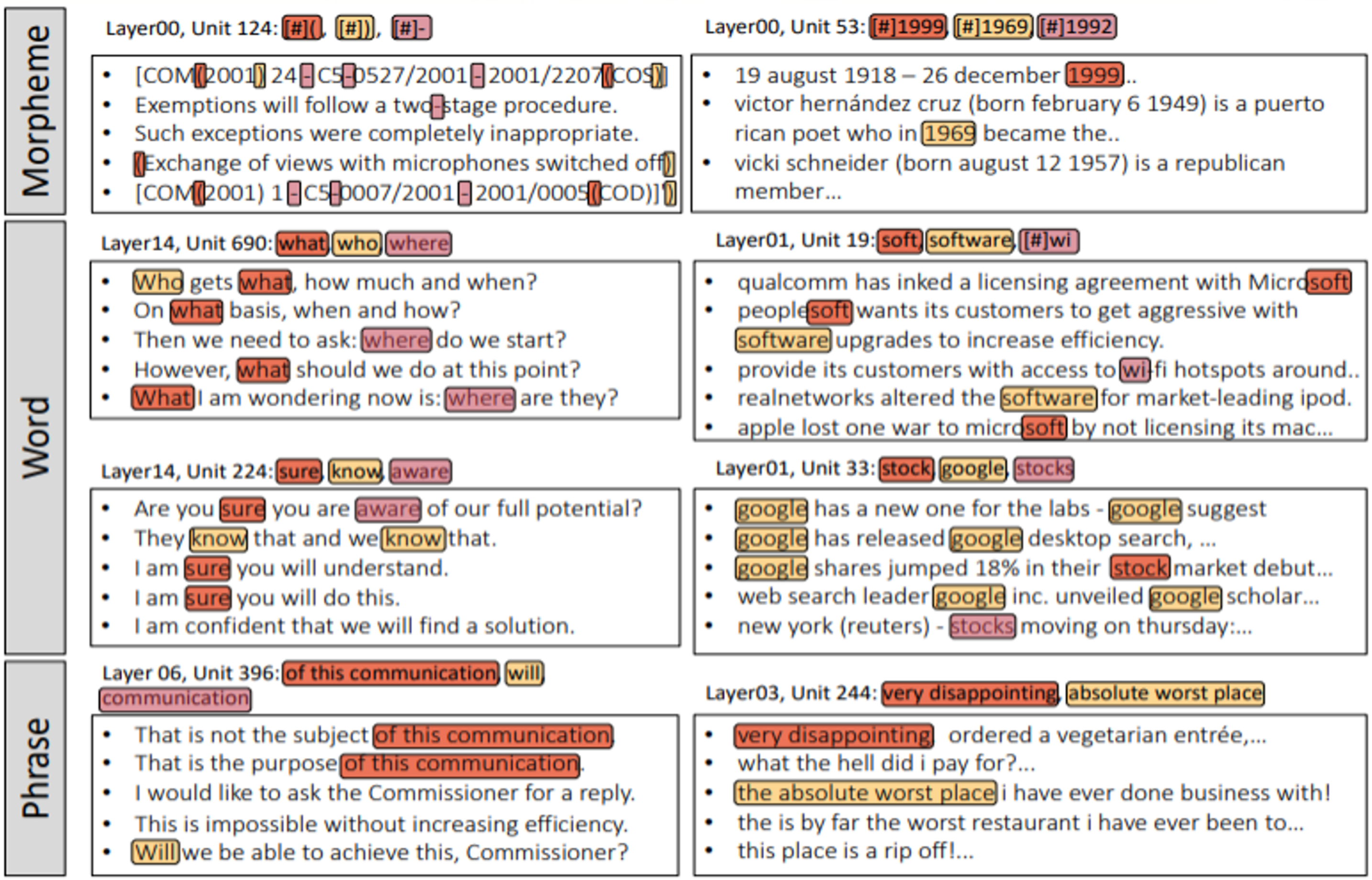


Figure 2: Mean and variance of selectivity values over all units in the learned representation for each dataset. Sentences including the concepts that our alignment method discovers always activate units significantly more than random sentences. See section 4.2 for details.

- Mean selectivity of the replicate set is the highest with a significant margin.
- Mean selectivity of the replicate set is higher than that of the one instance set, which implies that a unit's activation increases as its concepts appear more often in the input text.

Concept Alignment of Units:



(a) Translation (ByteNet) (b) Classification (VDCNN)

Figure 3: Examples of top activated sentences and aligned concepts to some units in several encoding layers of ByteNet and VDCNN. For each unit, concepts and their presence in top K sentences are shown in the same color. [#] symbol denotes morpheme concepts. See section 4.3 for details.

Concept Distribution In Layers:

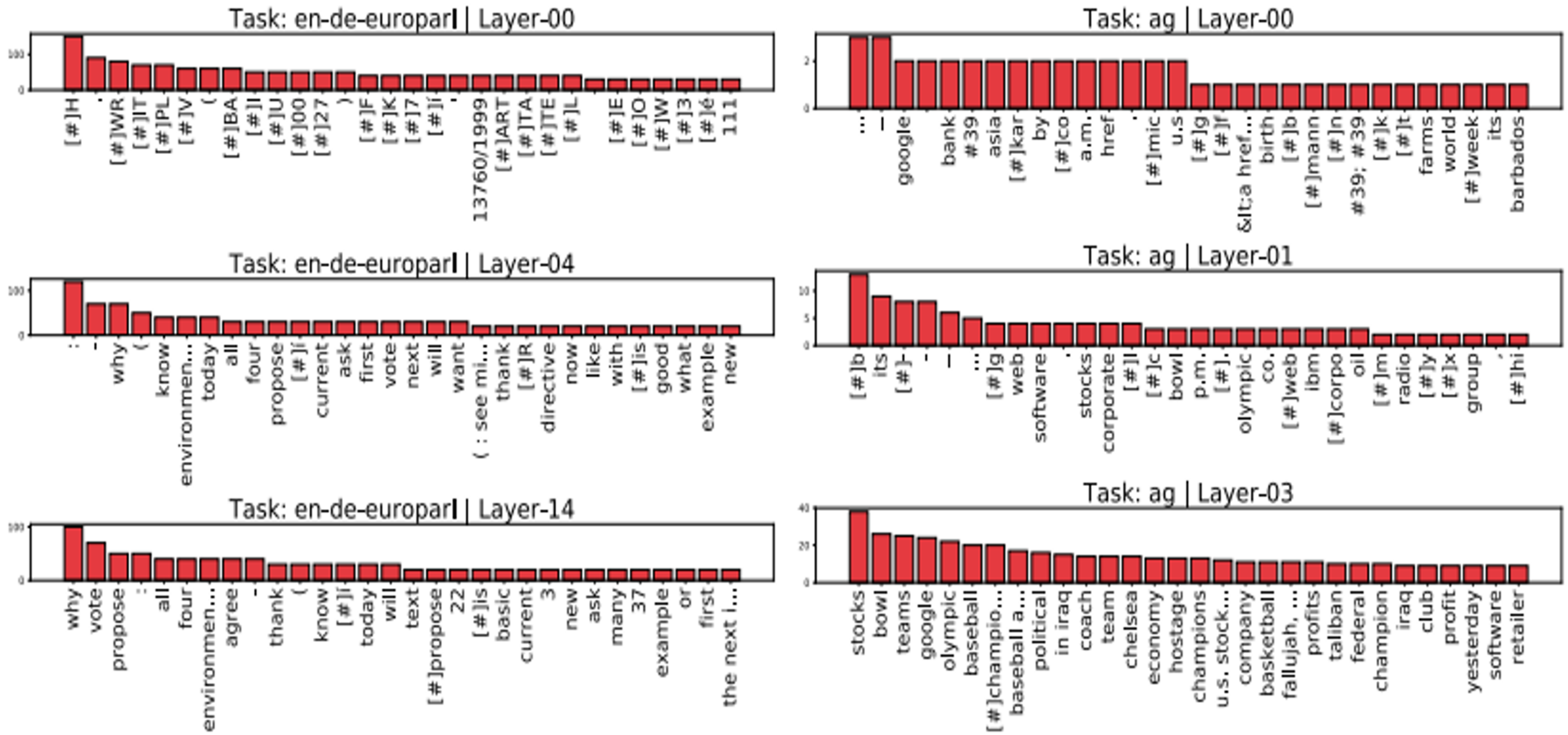


Figure 4: 30 concepts selected by the number of aligned units in three encoding layers of ByteNet learned on the Europarl translation dataset (left column) and VDCNN learned on AG-News (right column). [#] symbol denotes morpheme concepts. See section 4.4 for details.

- Data and Task-specific concepts are likely to be aligned to many units

Concept Granularity Evolution with Layers:

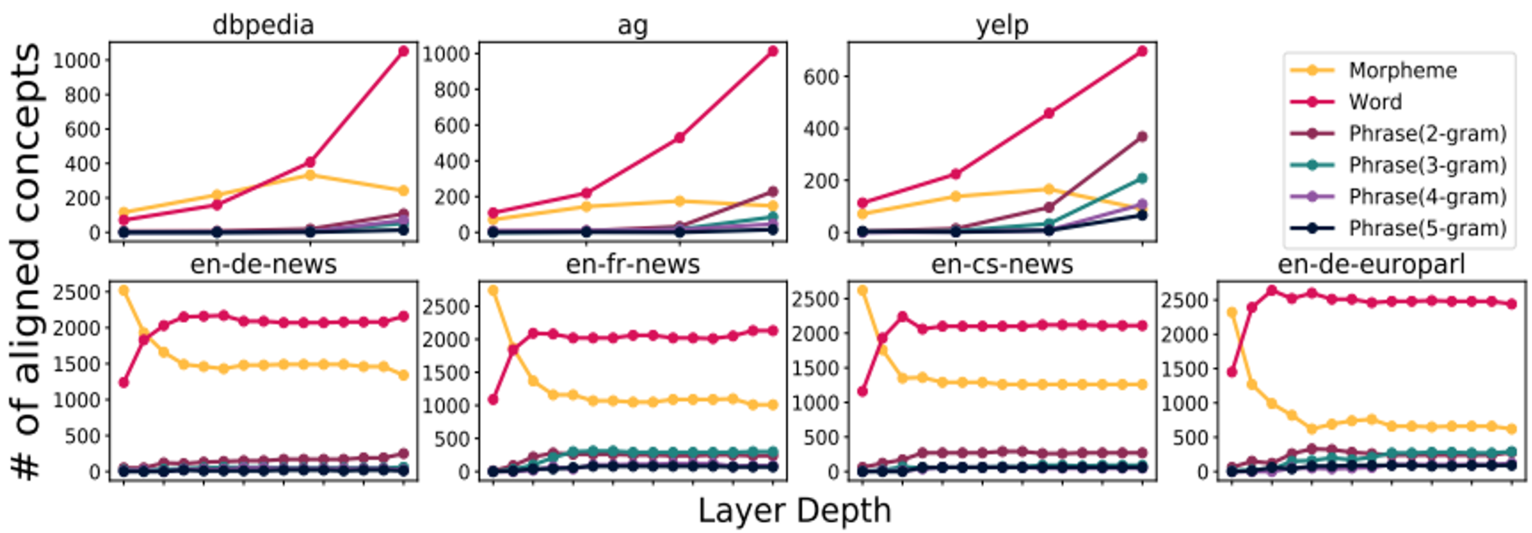


Figure 5: Aligned concepts are divided into six different levels of granularity: morphemes, words and N-gram phrases ($N = 2, 3, 4, 5$) and shown layerwise across multiple datasets and tasks. The number of units increases with layers in the classification models (*i.e.* $[64, 128, 256, 512]$), but in translation the number is constant (*i.e.* 1024) across all layers.

- In lower layers fewer phrase concepts but more morphemes and words are detected
- Concepts significantly change in shallower layers, but do not change much from middle to deeper layers.

Concept Granularity Evolution with Layers:

Why does concept granularity not evolve much in deeper layers?

1. Network is large enough so that the representations in the middle layers could be sufficiently informative to solve the task.

   Retrained ByteNet from scratch while varying only layer depth of the encoder.

- Unlike in computer vision where deeper layers are usually more useful and discriminative.

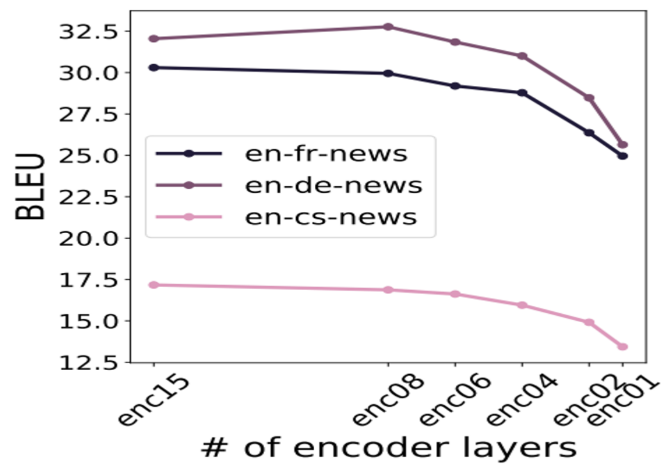


Figure 6: BLEU scores on the validation data for three translation models. We train ByteNet from scratch on each dataset by varying the number of encoder layers.

What Makes Certain Concepts Emerge More Than Others?

Why does concept granularity not evolve much in deeper layers?

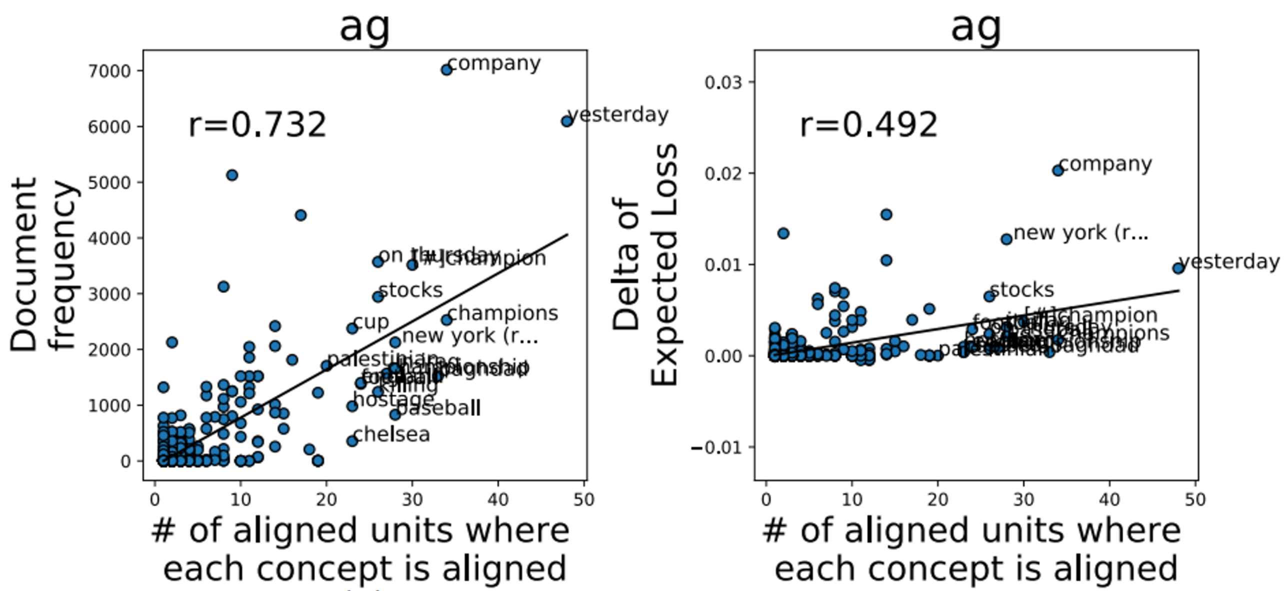
1. The concepts with a higher frequency in training data may be aligned to more units.
1. The concepts that have more influence on the objective function (expected loss) may be aligned to more units.

Delta of Expected Loss (DEL (c)) = $\mathbb{E}_{s\in\mathcal{S},y\in\mathcal{Y}}[\mathcal{L}(s,y)] - \mathbb{E}_{s\in\mathcal{S},y\in\mathcal{Y}}[\mathcal{L}(\mathsf{Occ}_c(s),y)]$

where, S is a set of training sentences, and Y is the set of ground-truths, and L(s, y) is the loss function for the input sentence s and label y.

$Occ_c(s)$ is an occlusion of concept c in sentence s: replace concept c by dummy character tokens that have no meaning.

What Makes Certain Concepts Emerge More Than Others?



(a) (b)

Figure 7: Correlations between the number of units per concept and (a) document frequency and (b) delta of expected loss. Pearson correlation coefficients are measured at the final layer of the AG News sentence classification task. Some concepts aligned to many units are annotated. Results on other tasks are available in Appendix F.

Conclusion and Future Work:

- Proposed a simple but highly effective concept alignment method for character-level CNNs to confirm that each unit of the hidden layers serves as detectors of natural language concepts.
- Consequently, authors shed light on how deep representations capture the natural language, and how they vary with various conditions.
- An interesting future direction is to extend the concept coverage from natural language to more abstract forms such as sentence structure, nuance, and tone.
- Combining definition of concepts with the attention mechanism.

References:

- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. ICLR, 2017.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In CVPR, 2017.
- Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing Higher-layer Features of a Deep Network. University of Montreal, 2009.
- Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and Understanding Recurrent Networks.
- Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. How Transferable are Neural Networks in NLP Applications? In EMNLP, 2016.
- Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. Distill, 2017.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object Detectors Emerge in Deep Scene CNNs. In ICLR, 2015.