

DW Poznań - KERAS PODSTAWY #7

2020-02-26

KERAS Co to jest i jak użyć do nauki NLP

Agenda

- Preprocessing NLP - **Basia**
- Basic KERAS - **Krzysztof**
- Pierwsza sieć Kerasa, czyli podstawowa sieć - **Mateusz**
- Convolutional network - **Paweł**
- LSTM Network - czyli sieci rekurencyjne które mogą pomóc - **Arek**
- Dropout - **Marcin Sola**
- BatchNormalization - **Marcin Sola / Arek**

Preprocessing NLP

- Tokenizacja
- Usuwanie stopwords
- Usuwanie interpunkcji
- Lematyzacja/stemming

<i>Word</i>	<i>Lemmatization</i>	<i>Stemming</i>
was	be	wa
studies	study	studi
studying	study	study

- Standaryzacja wielkości liter

Dropout

- Technika regularyzacji sieci neuronowych zaproponowana w 2014 roku przez Srivastava, et al. w publikacji pt. [Dropout: A Simple Way to Prevent Neural Networks from Overfitting](#). (opatentowana przez Google)
- Neurony składające się na sieć neuronową z czasem zaczynają “specjalizować się” w wykrywaniu pewnych zależności. W przypadku zbyt dużej specjalizacji, sąsiadujące ze sobą neurony mogą być zbyt mocno uzależnione od tej specjalizacji, przez co model łatwiej przeuczyć.
- Dropout jest sposobem na ograniczenie ryzyka overfittingu oraz przyspieszenia procesu trenowania sieci neuronowych.

Dropout

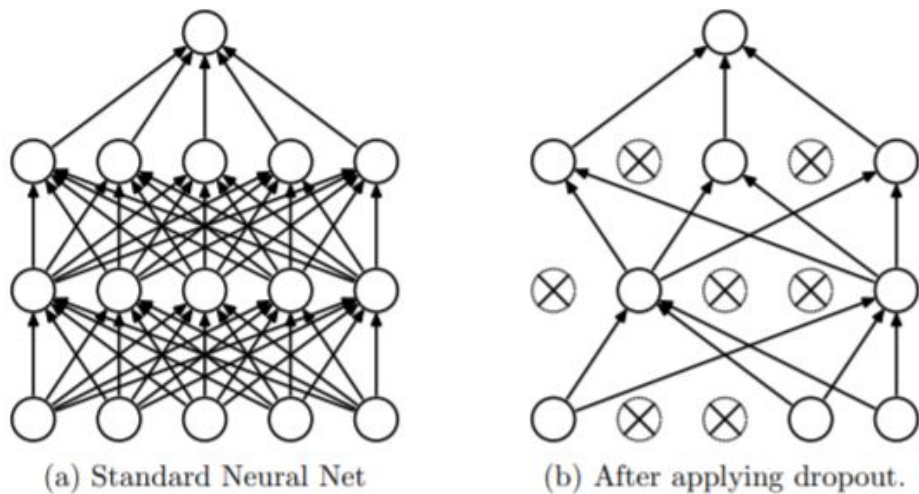


Figure 1: Dropout Neural Net Model. **Left:** A standard neural net with 2 hidden layers. **Right:** An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

Dropout

- Z jednej sieci neuronowej o liczbie elementów n można za pomocą dropoutu utworzyć 2^n “odchudzonych” (ang. *thinned*) sieci treningowych
- Następnie kolekcja tych sieci jest trenowana (z czego wiele z tych “podsieci” nie jest trenowana w ogóle, lub bardzo rzadko, ze względu na ich wykładniczą naturę ich generowania)
- Do testowania używana jest “pełna” sieć neuronowa o n elementach ze złagodzonymi wagami.

Batch Normalization

- Technika optymalizacji pozwalająca na szybsze trenowanie sieci neuronowych
- Zaproponowana w 2015 roku przez inżynierów z Google ([link](#))
- Polega na normalizacji wartości na wejściu i wyjściu poszczególnych warstw w sieci, dzięki czemu unika się zbyt wysokich lub zbyt niskich wartości w neuronach sieci.
- Autorzy publikacji sugerują, że BatchNorm może zastępować dropout, ale w praktyce spotyka się architektury wykorzystujące jedno i drugie rozwiązanie jednocześnie.

Podsumowanie



<https://forms.gle/LDKuSTJeqnzSQLbe9>

<https://github.com/dataworkshop/dw-poznan-project>

Dziękuję