

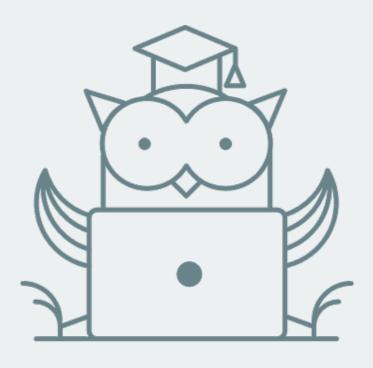
ОНЛАЙН-ОБРАЗОВАНИЕ



Выучивание метрики.

Иногда расстояние мешает подобрать нужные слова

Артур Кадурин



План на сегодня

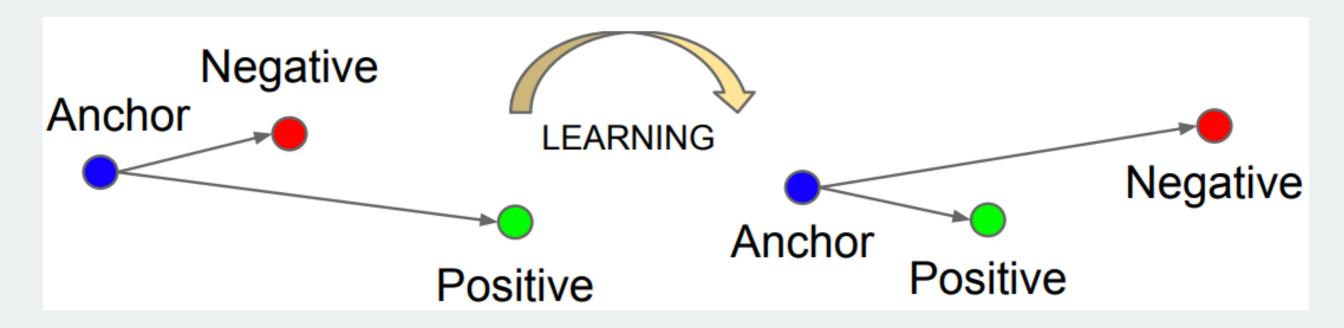


- 1. Опять тройка
- 2. One-shot learning
- 3. Одно представление на всех



Triplet loss





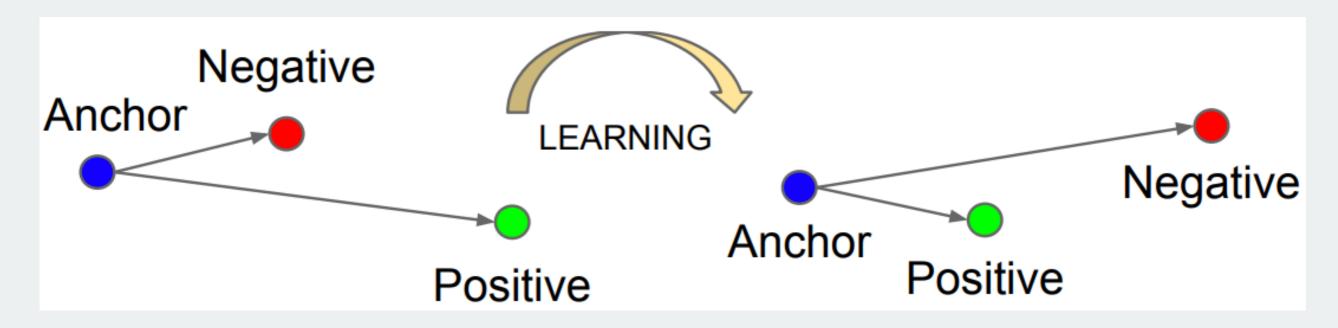
Предположим, что мы способны померять «расстояние» между объектами. Тогда, нам бы хотелось, чтобы объекты с одинаковыми или похожими метками были на небольшом расстоянии друг от друга и наоборот.

Как должно быть устроено такое расстояние?



Triplet loss





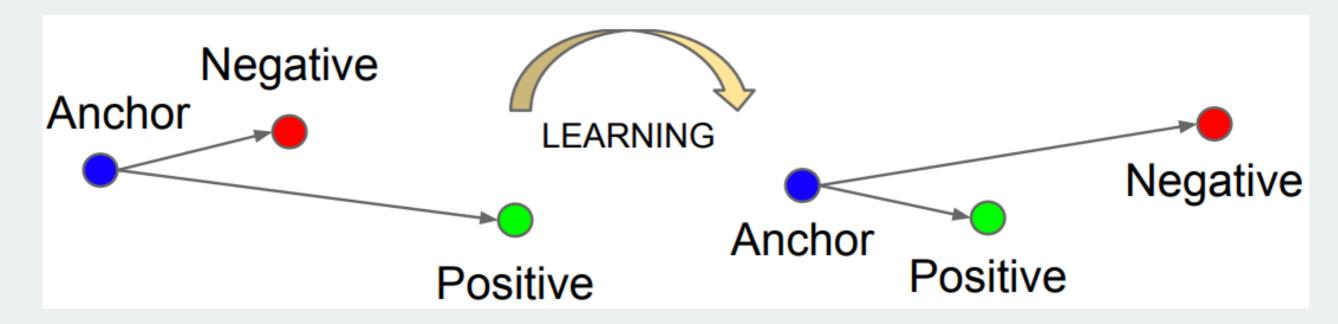
Допустим у нас есть нейросеть отображающая наши примеры в пространство над которым мы умеем мерять расстояние.

Каким должен быть лосс?



Triplet loss





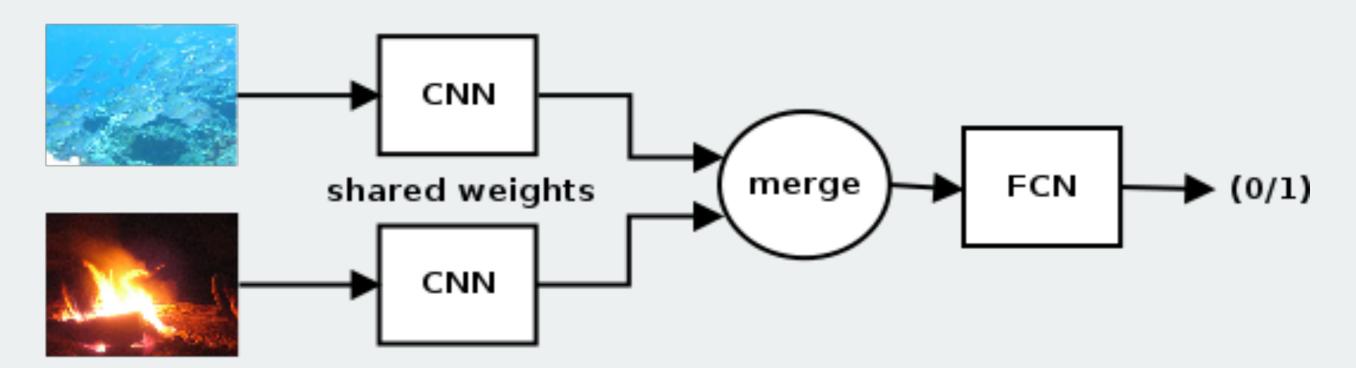
Допустим у нас есть нейросеть $f: x \to R^d$ отображающая наши примеры в пространство над которым мы умеем мерять расстояние D.

$$L = \sum_{i}^{N} \left[D\left(f(x_i^t), f(x_i^p)\right) - D\left(f(x_i^t), f(x_i^n)\right) + \alpha \right]_{+}$$



Сиамские сети





Сеть, которая сначала «сжимает» несколько примеров, а потом вычисляет над ними какую-нибудь функцию обычно называют сиамской.

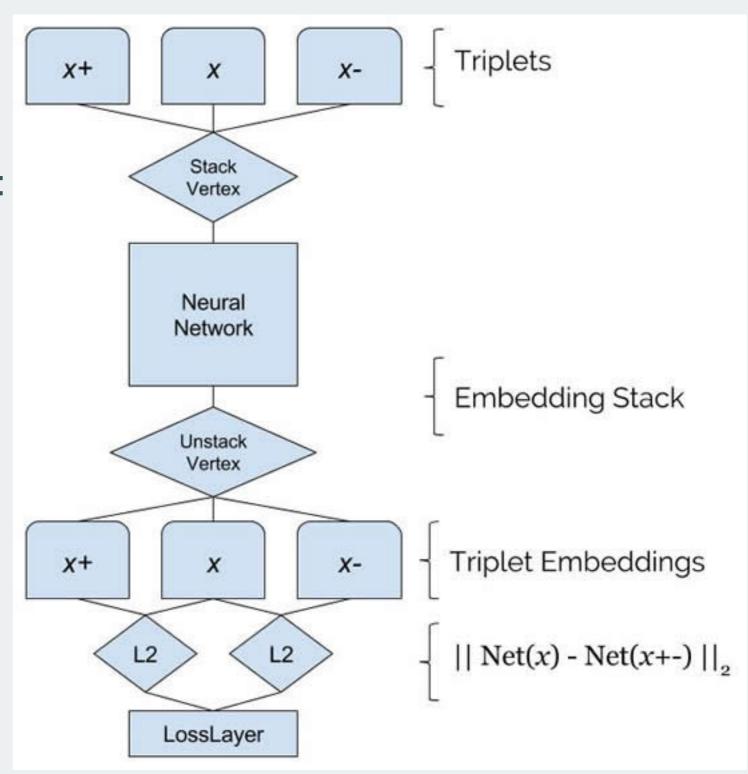


Архитектура



Триплеты собираются из троек: якорь, положительный пример, отрицательный пример.
При этом, есть различные

стратегии выбора положительных и отрицательных примеров.

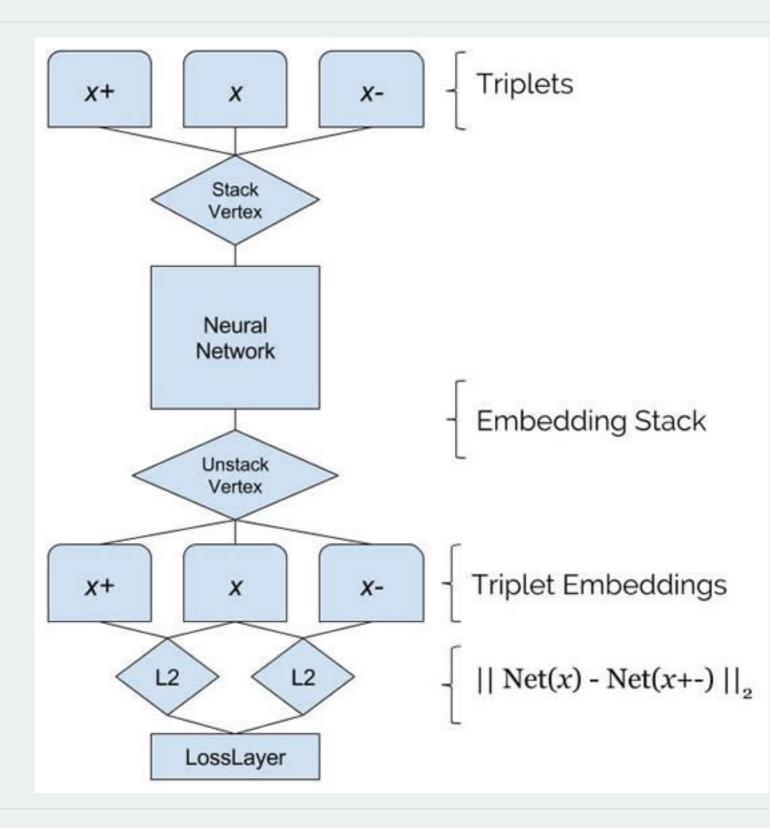




Архитектура



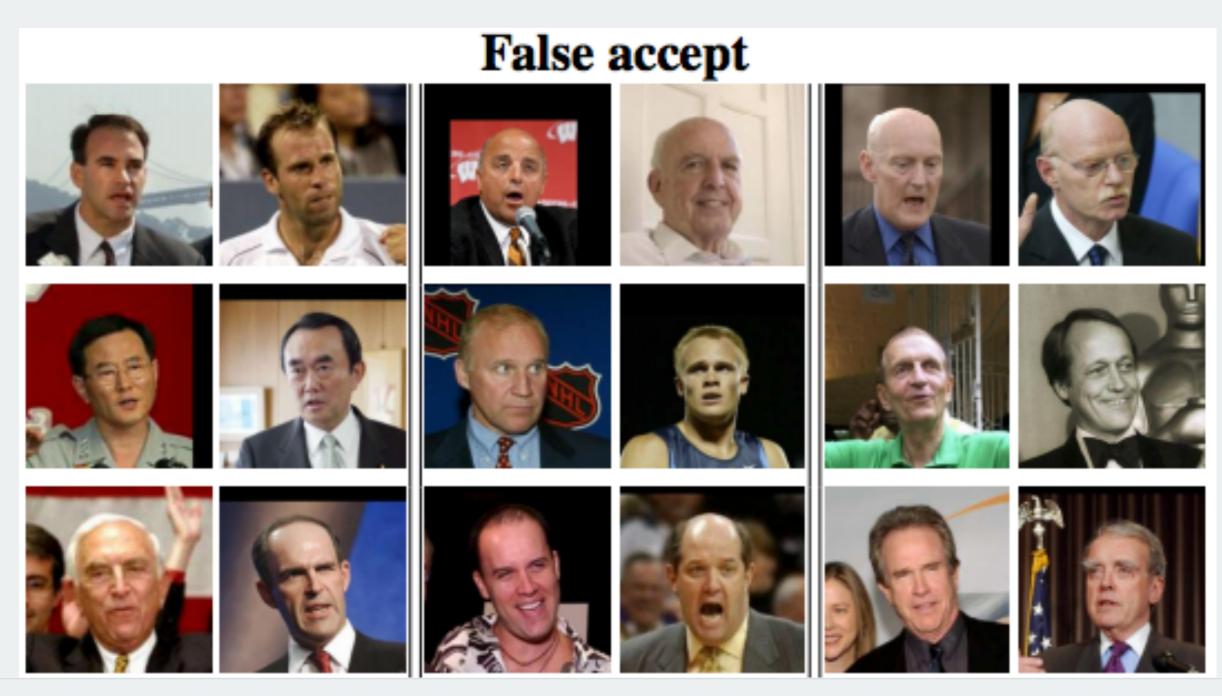
Кроме того, triplet loss может быть регуляризатором латентного слоя, а так же, вычисляться не только в последнем слое, но и после каждого слоя сети.





FaceNet

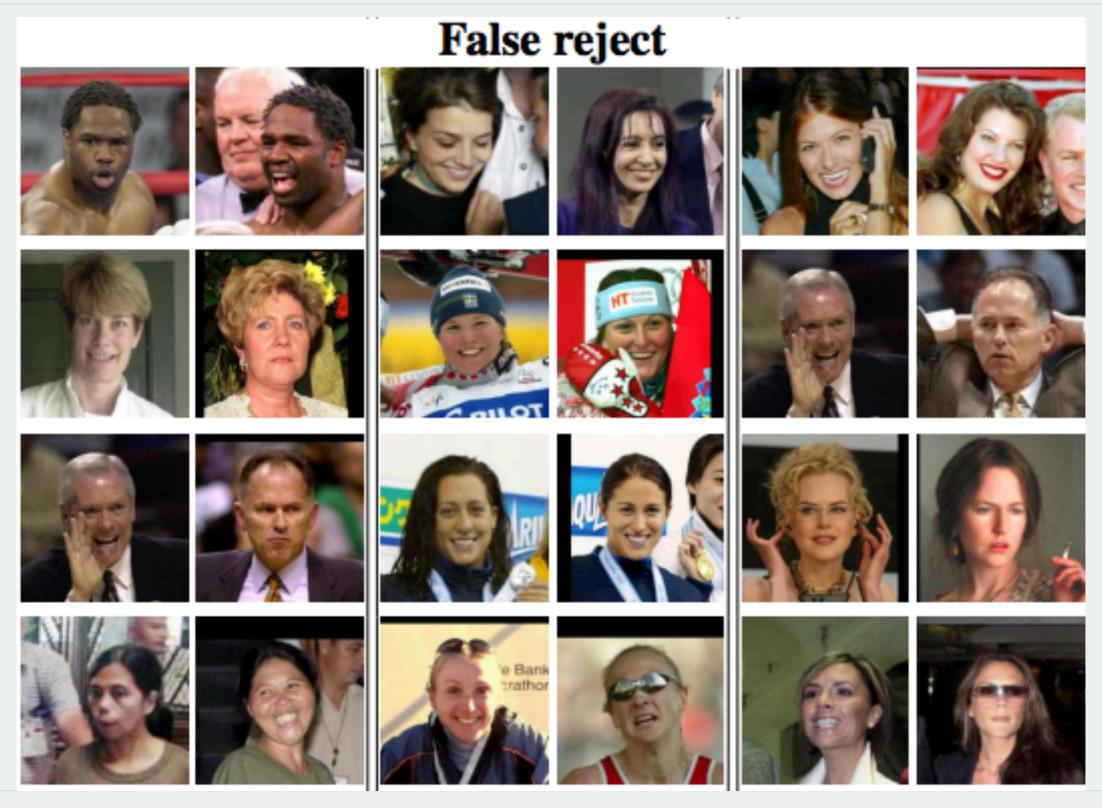






FaceNet



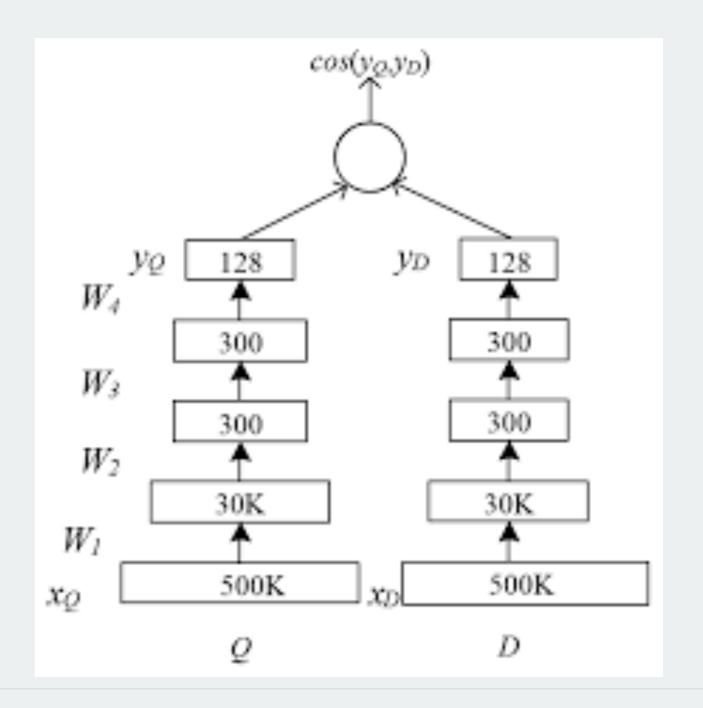




Deep structured semantic model



В DSSM в качестве якоря используется запрос, а положительными и отрицательными примерами служат документы. При этом обе головы все так же используют одинаковые веса.





План на сегодня



- 1. Опять тройка
- 2. One-shot learning
- 3. Одно представление на всех



Omniglot



В датасете Омниглот представлены 50 алфавитов, каждый символ в которых написан 20 раз различными людьми.

أمنيغلوت Arabic	Օմնիգլօթ Armenian	なみななのかでた Atlantean	ည်ရေးဗျာကျေ (Balinese
অমিগ্লট্		WVN VY WVV	ჀჅჼჄჅჅ
Bengali	Blissymbolics	Buginese	Buhid
ဩပ်နိဂ်လောတ် Burmese	SY hEGW Cherokee	全量文学網至 Chinese/Bopomofo	Ommeyor
⊳ړ ۵۹⊃۵	Омниглот	ไวหเตเ เรา	ऑम्निग्लॉट्
0	Cyrillic	Deseret	De∨anagari
Cree	cyc	2000.00	20141149411
Smile Land	16×754	አምኒ <i>ግ</i> ሎት	ომნიგლოტ
	•	_	- *
SMILENUS	16×7554	አምኒ <i>ግ</i> ሎት	ომნიგლოტ
J'MJEELUV D'Ni	イレメモダケキ Early Aramaic	አ ምኒ <i>ግ</i> ሎት ^{Ge′ez}	ომნიგლოტ _{Georgian}
<i>ΟνωΣΕνω</i> ν _{D'Ni} Ομνιγλοτ	1 6 ¥ 7 5 5 4 Early Aramaic ઑમ્નિગ્લૉટ્	አምኒ গ ሎት ^{Ge'ez} ਓਮ੍ਰਿਗ੍ਲੋਟ੍ Gurmukhi (Punjabi)	ომნიგლოტ Georgian אמני- גלוט
Ονωσενών D'Ni Ομνιγλοτ Greek	16 🗴 રે ઇ ઇ ઇ Early Aramaic ઑમ્નિગ્લૉટ્ Gujarati	አምኒ ግሎት ^{©©™} ਓਮ੍ਹਿਗ੍ਲੋਟ੍	ომნიგლოტ Georgian אמני- גלוט Hebrew
Umacessalv D'Ni Oμνιγλοτ Greek 음니글롯	イレメ <i>そう</i> サキ Early Aramaic 砂ルーンご Gujarati おむにぐろっと	አምኒግሎት ^{Ge'ez} ਓਮ੍ਰਿਗ੍ਲੋਟ੍ ^{Gurmukhi (Punjabi)} ಒಮ್ನಿಗ್	ომნიგლოტ Georgian אמני- גלוט Hebrew オムニグロット



Omniglot



В эксперименте 2015г командой из Университета Торонто для обучения были выбраны 30/50 алфавитов и 12/20 «писателей». Тест проводился на новом алфавите для новой пары «писателей» каждая из букв первого классифицировалась вторым. Итоговая точность: 92%

أمنيغلوت	Օմնիգլօթ	<i>የያ</i> ያውያልያው	ည်ရေါ့ဗ္ဗ ^ည ည်ချှ
Arabic	Armenian	Atlantean	Balinese
অমিগ্লট্	$\boxtimes \emptyset$	$\sim \sim $	ჀჅჼჄჅჅ
Bengali	Blissymbolics	Buginese	Buhid
ဩပ်နိဂ်လောတ်	Ճℋℎ EGW	全氧文英網系	Ommeyor
Burmese	Cherokee	Chinese / Bopomofo	Coptic
⊳ړ مه م	Омниглот	ไวหรด เหา	ऑम्निग्लॉट्
Cree	Cyrillic	Deseret	Dovernogori
Ciee	Cyllinc	Deselec	Devanagari
Smiller	16×7554	አምኒ <i>ግ</i> ሎት	ომნიგლოტ
	,	_	- *
SMILENS	16×7554	አምኒ <i>ግ</i> ሎት	ომნიგლოტ
JMJEETUS D'Ni	イレメモダケキ Early Aramaic	አምኒ <i>ግ</i> ሎት ^{Ge′ez}	ომნიგლოტ Georgian
<i>ΟνωΣΕνω</i> ν _{D'Ni} Ομνιγλοτ	1 6 ¥ 7 5 5 4 Early Aramaic ઑમ્નિગ્લૉટ્	አምኒ ግሎት [©] ਉਂਦ ਓਮ੍ਹਿਗ੍ਲੋਟ੍	ომნიგლოტ ^{Georgian} אמני- גלוט
Ονωσενών D'Ni Ομνιγλοτ Greek	16 🗴 રે ઇ ઇ ‡ Early Aramaic ઑમ્નિગ્લૉટ્ Gujarati	አምኒ গ ሎት ^{Ge'ez} ਓਮ੍ਰਿਗ੍ਲੋਟ੍ Gurmukhi (Punjabi)	ომნიგლოტ ^{Georgian} אמני- גלוט Hebrew

Klingon

Limbu

Khmer

Malayalam

Zero-shot тегирование



Предсказание одного или нескольких тегов можно представить задачей классификации, однако, когда тегов сотни тысяч, собрать трейнсет для классификации практически невозможно.

В статье от Paralleldots 2018 года задача классификации переформулирована в стиле выучивания метрик.

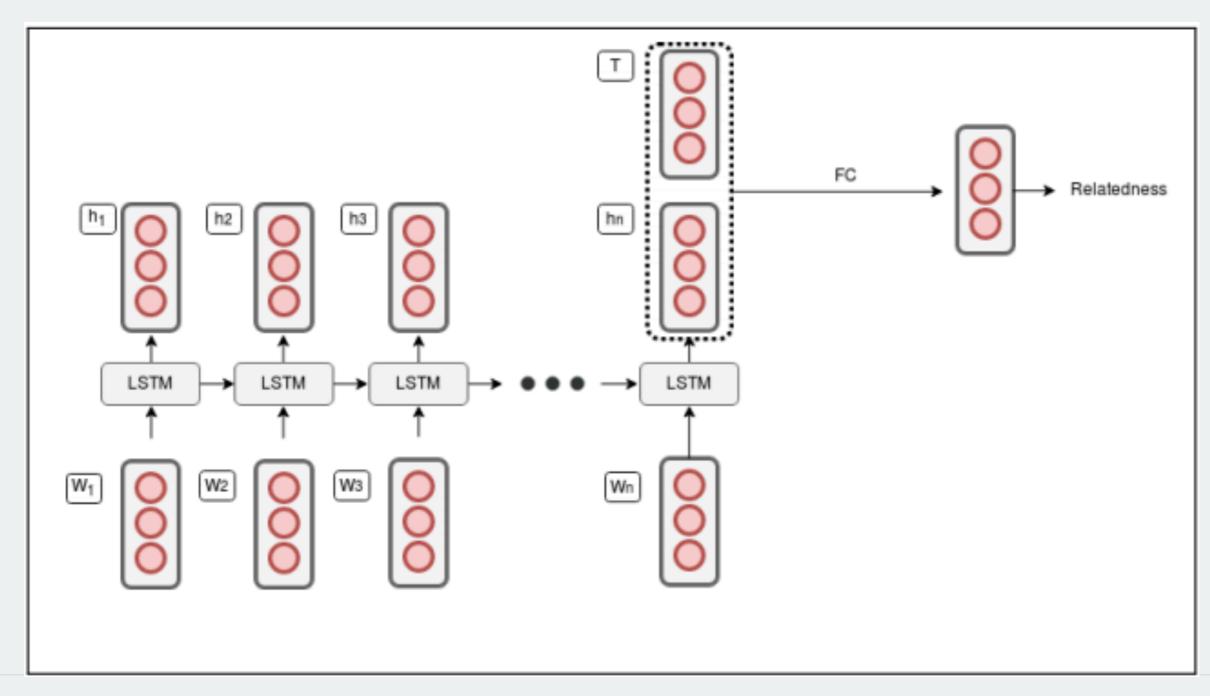
В качестве эмбеддингов слов авторы использовали Google News

Embedding отсюда: https://code.google.com/archive/p/word2vec/



Zero-shot тегирование

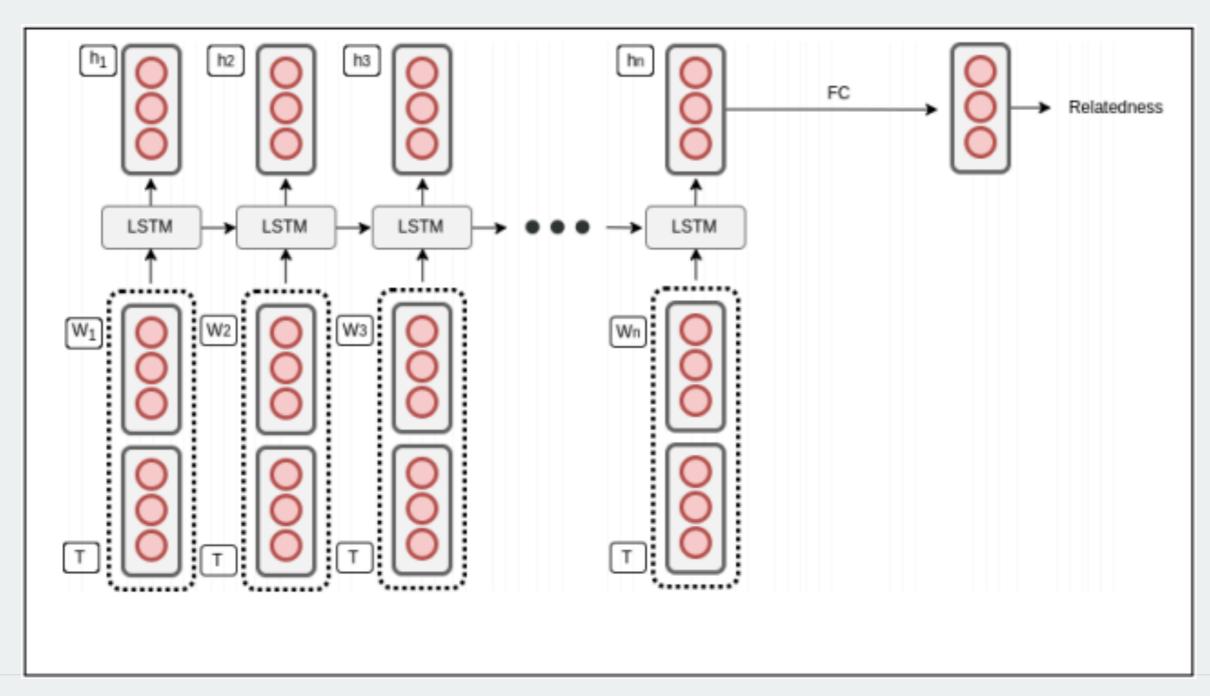






Zero-shot тегирование







План на сегодня

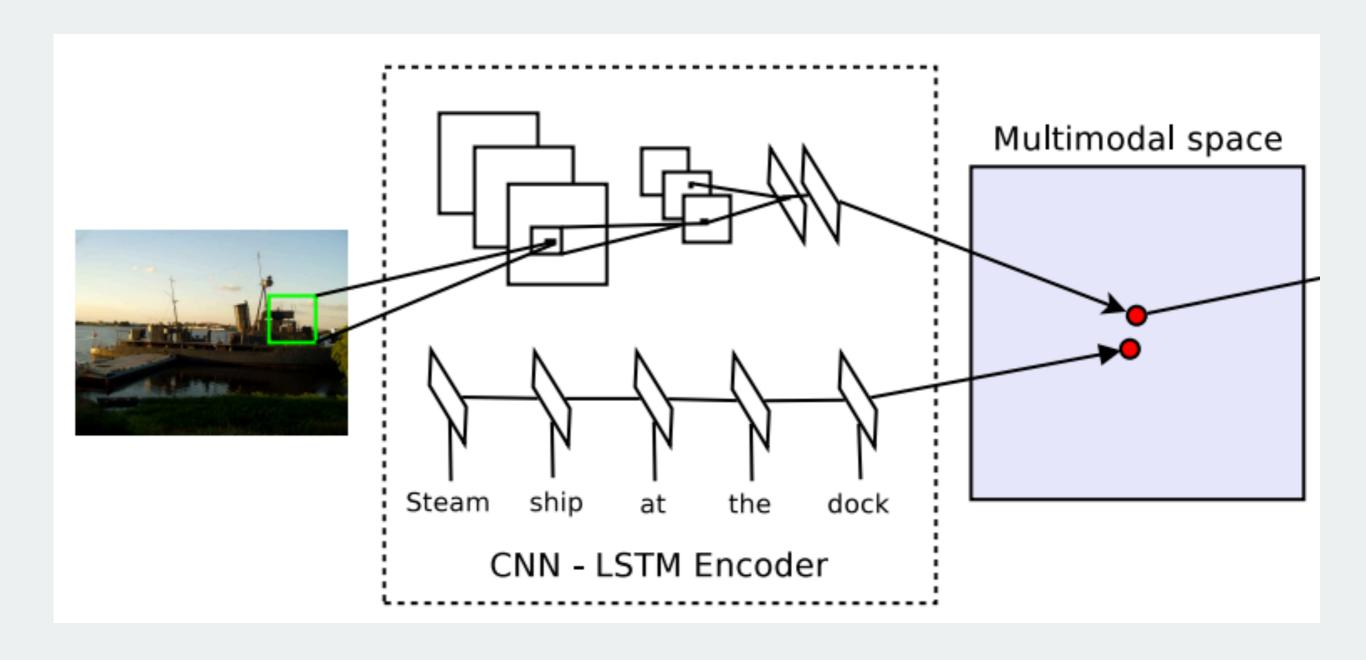


- 1. Опять тройка
- 2. One-shot learning
- 3. Одно представление на всех



Мультимодальные данные











Спасибо за внимание!