LSTM, GRU

Червонцев Сергей

МФТИ ФИВТ

Deep Learning School, 10 марта 2018

План

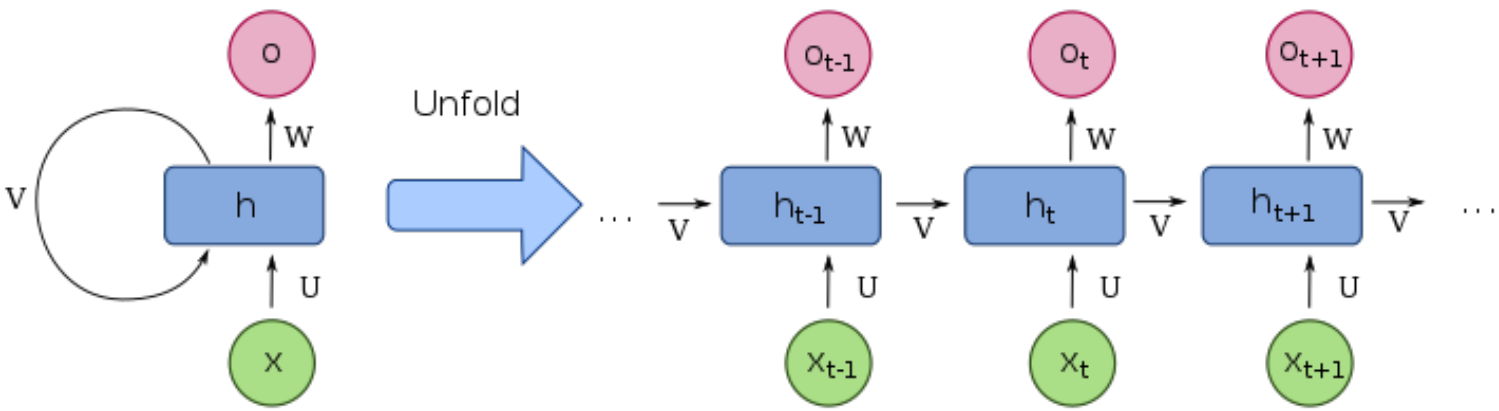
1. Эволюция RNN
   - BasicRNN
   - LSTM
   - GRU
2. Применения RNN
   - Language Modelling
   - Question Answering

План

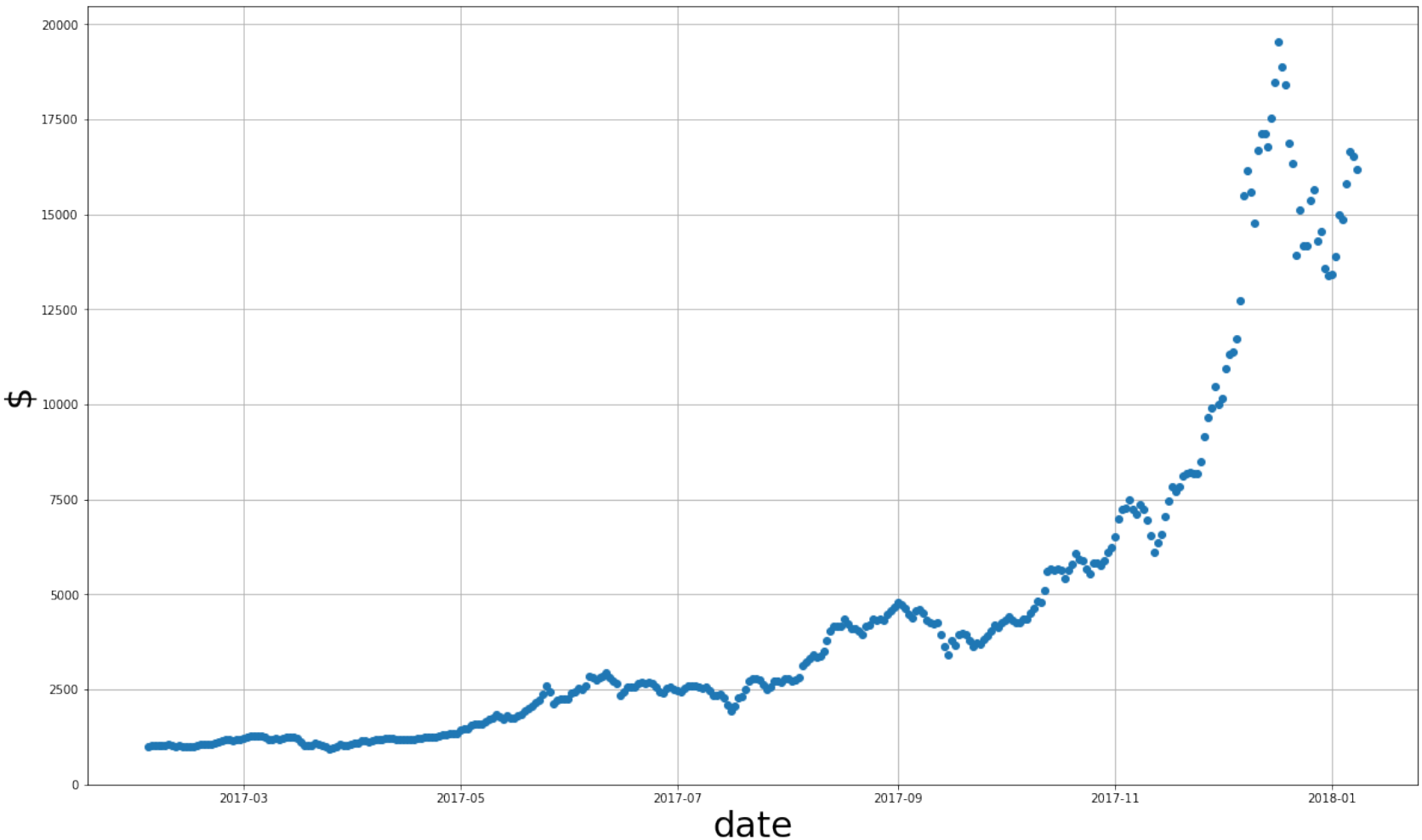
BasicRNN

BasicRNN – это просто

$$\begin{cases} h_t = \sigma_h(W_{hx}x_t + W_{hh}h_{t-1} + b_h), \\ o_t = \sigma_o(W_{oh}h_t + b_{oh}). \end{cases}$$



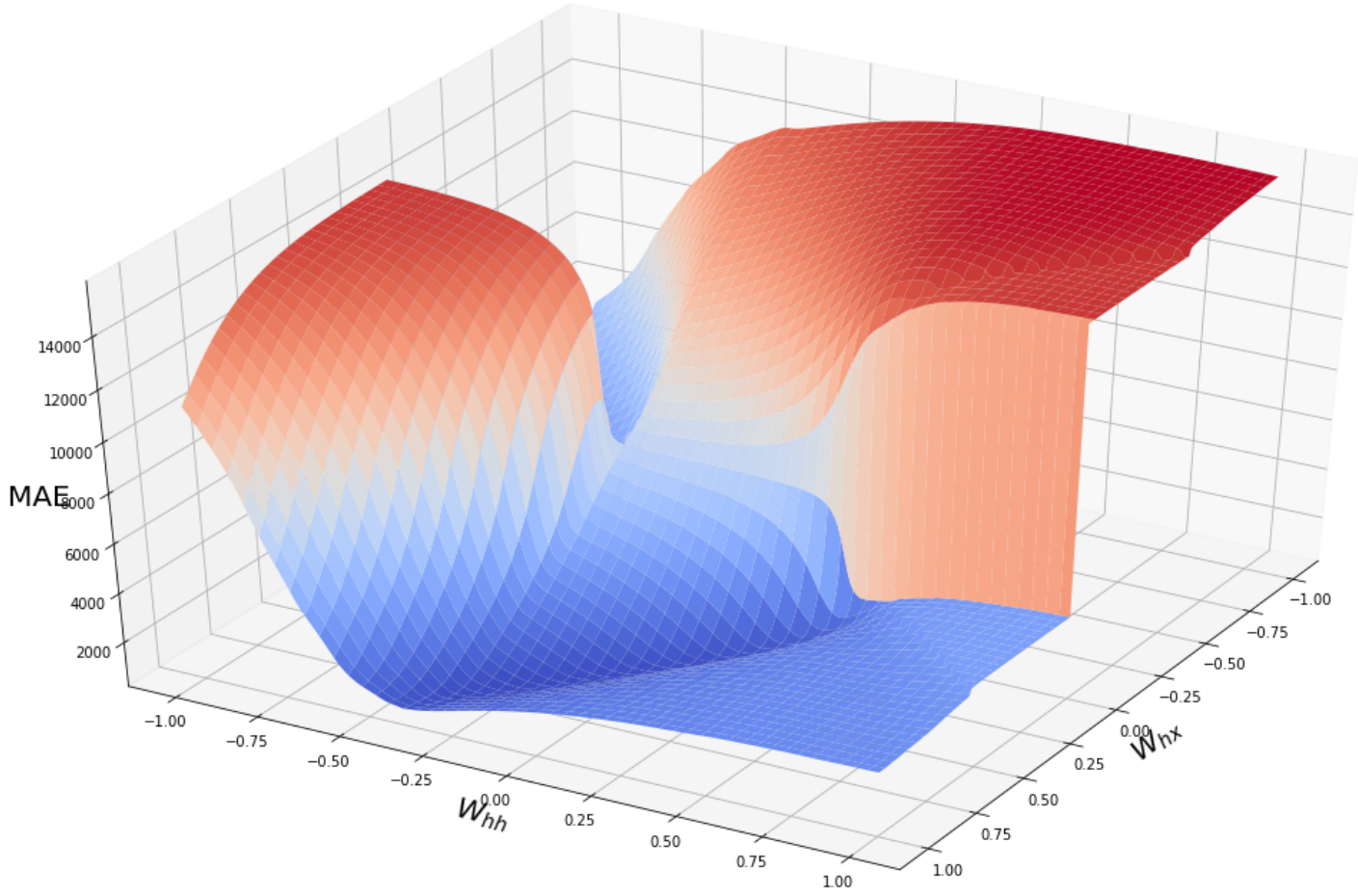
Время конвертировать знания в деньги!



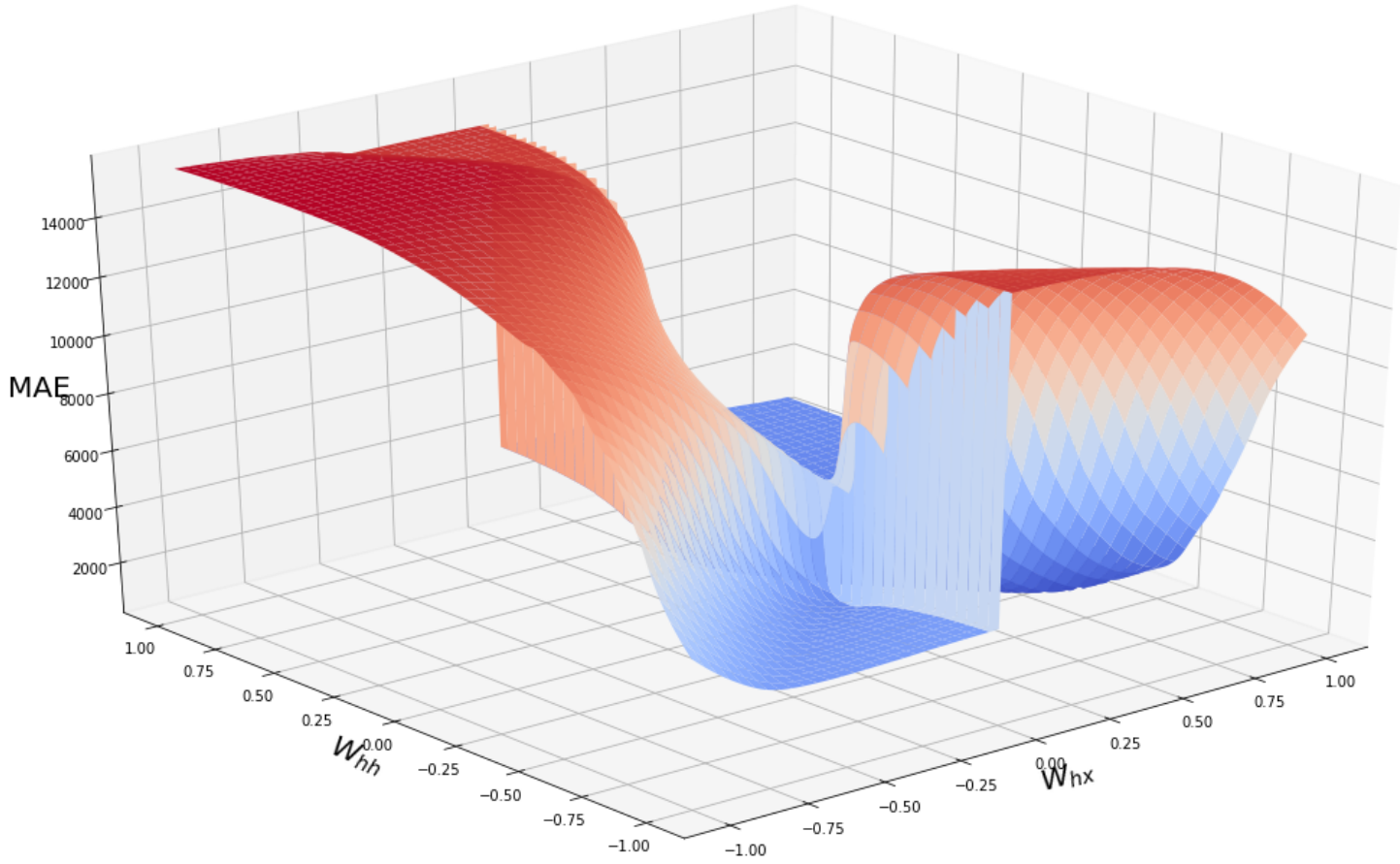
BasicRNN – это коротко

```
def batch_wise_dummy_RNN(X, W_hx, W_hh, b_h):
    batch_size = X.shape[0]
    sequence_len = X.shape[1]

    h = np.zeros((batch_size, sequence_len+1))
    o = np.zeros((batch_size, sequence_len+1))

    for t in range(0, sequence_len):
        h[:, t + 1] = 3 * np.tanh(W_hx * X[:, t] + W_hh * h[:, t] + b_h * np.ones((batch_size,)))
        o[:, t + 1] = h[:, t + 1]
    return o[:, -1]
```

Но учить это сложно



BasicRNN


MAE
14000
12000
10000
8000
6000
4000
2000
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
W_{hh}
W_{hx}

BasicRNN

Проблемы BasicRNN:

- Долго учится
- Взрывающиеся градиенты
- Затухающие градиенты
- Короткая память

BasicRNN

Проблемы BasicRNN:

- Долго учится
- Взрывающиеся градиенты
- Затухающие градиенты
- Короткая память

LSTM и GRU решают последние две.

LSTM

Нужна инвариантность к "сжатию-растяжению"времени [1].

Unwarped task example:
Input: `All human beings are born free and equal`
Output: `All human beings are born free and equa`

Uniform warping example (warping $\times 4$):
Input: `AAAAlllllll hhhhuuuummmmaaaannnn`
Output: `AAAAlllllll hhhhuuuummmmaaaa`

Variable warping example (random warping $\times 1$–$\times 4$):
Input: `Allllll hhhummmmaannn bbbbeeiiingssss`
Output: `AAAllllll huuuummaaan bbeeeinggg`

Figure 2: A task involving pure warping.

[1] Tallec, Ollivier. 2018. Can Recurrent Neural Networks warp time?

LSTM

Для этого используется input gate:

$$\begin{cases} i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\ \tilde{h}_t = \tanh\left(W_{hx}x_t + W_{hh}h_{t-1} + b_h\right), \\ h_t = (1 - i_t) \odot h_{t-1} + i_t \odot \tilde{h}_t. \end{cases}$$

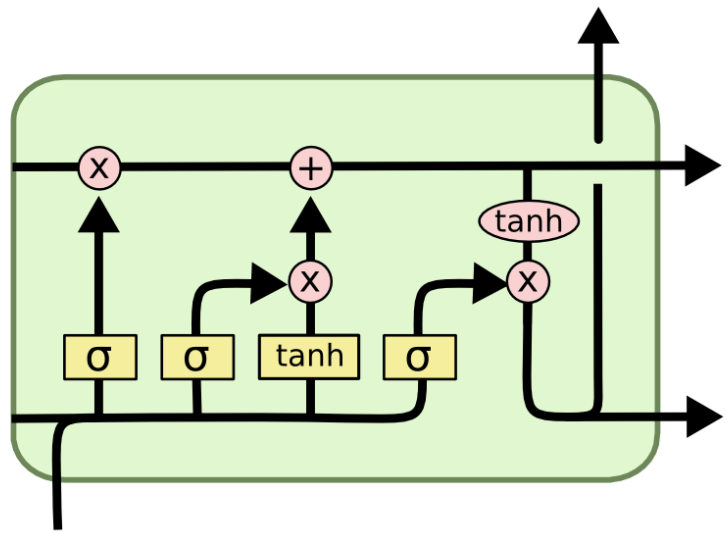
LSTM

Но некоторые входы могут быть важны в дальнейшем, но не так важны сейчас. Чтобы учитывать это, предложили два вида памяти — долгосрочную и короткую:

$$\begin{cases} i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\ \tilde{c_t} = \tanh\left(W_{hx}x_t + W_{hh}h_{t-1} + b_h\right), \\ c_t = (1 - i_t) \odot c_{t-1} + i_t \odot \tilde{c_t}, \\ o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \\ h_t = o_t \odot \tanh\left(c_t\right). \end{cases}$$

LSTM

$$
\begin{cases}
i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\
f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \\
\tilde{c}_t = \tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h), \\
c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\
o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \\
h_t = o_t \odot \tanh(c_t).
\end{cases}
$$



GRU

Существенный недостаток LSTM — огромное количество параметров.

GRU

Существенный недостаток LSTM — огромное количество параметров.

"Лайт" версия LSTM — GRU:

$$\begin{cases} i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\ f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \\ \tilde{h_t} = \tanh\left(W_{hx}x_t + W_{hh}(f_t \odot h_{t-1}) + b_h\right), \\ h_t = (1 - i_t) \odot h_{t-1} + i_t \odot \tilde{h_t}. \end{cases}$$

План

Language Modelling

Будем делать one-hot символов и учить LSTM выдавать следующий [2].

Для уверености можно настакать несколько LSTM, подавая hidden state нижних как input в верхние...

[2] Karpathy. 2015. The Unreasonable Effectiveness of Recurrent Neural Networks

Language Modelling

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Language Modelling

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariccoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M^\bullet} = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition* **??** *(without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S*. Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S, $E = \mathcal{F}_{X/S}$. *Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$. *Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation* **??**. *Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (**??**). On the other hand, by Lemma **??** we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

Language Modelling

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

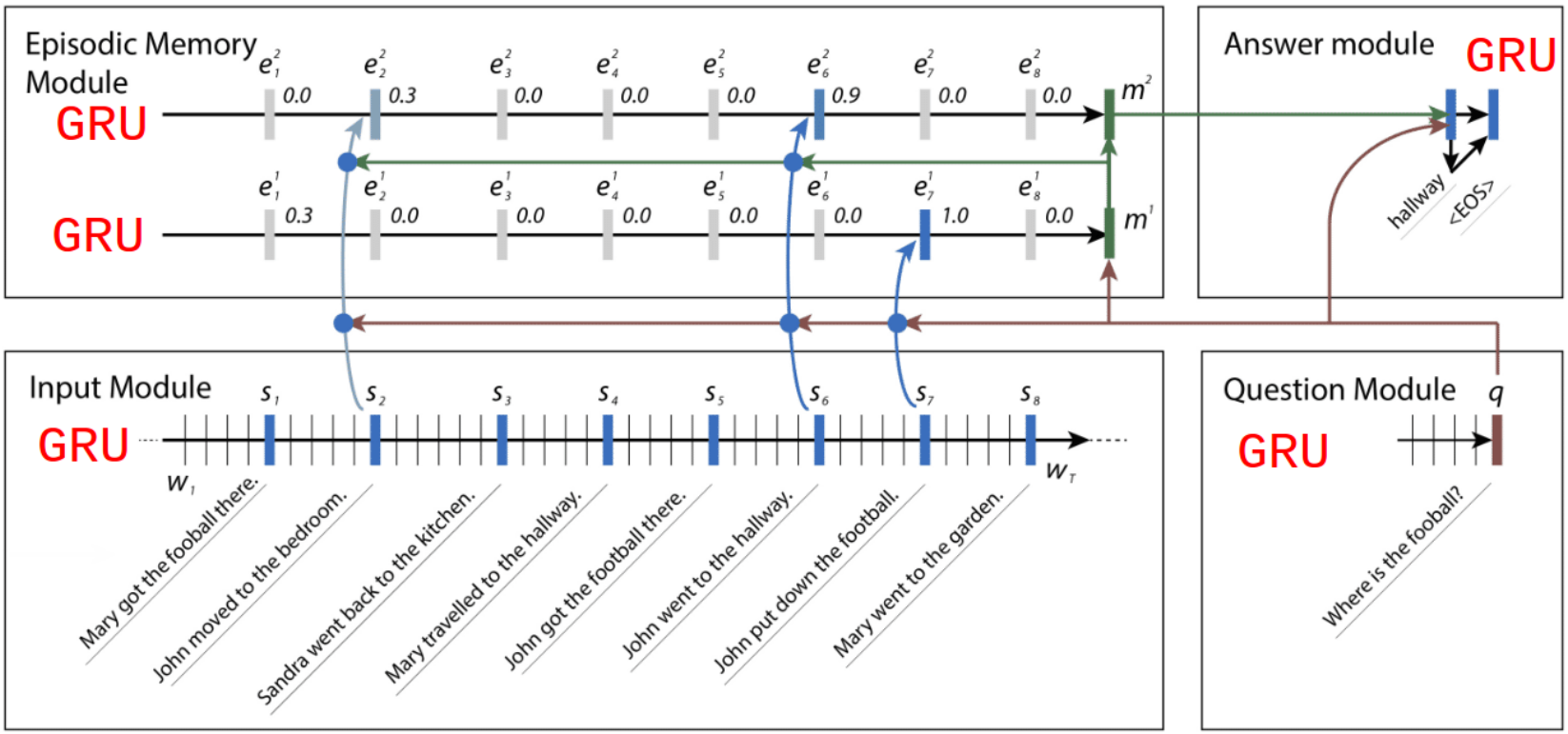
Question Answering

У нас есть текст (I), вопрос (Q) и ответ (A). Учим сетку отвечать на вопросы.

I: Jane went to the hallway.
I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden
I: It started boring, but then it got interesting.
Q: What's the sentiment?
A: positive
Q: POS tags?
A: PRP VBD JJ , CC RB PRP VBD JJ .

Question Answering

Немного тяжеловесный, но естественный подход - Dynamical Memory Network [3].



[3] Kumar, Ondruska, Socher, ... 2016. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Question Answering