Advanced NLP

Summer 2023

Anoop Sarkar

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ICLR 2015

https://arxiv.org/abs/1409.0473

Machine Translation

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

Target language

Source language

Neural Machine Translation

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c),$$

$$\mathbf{y} = (y_1, \cdots, y_{T_y}).$$

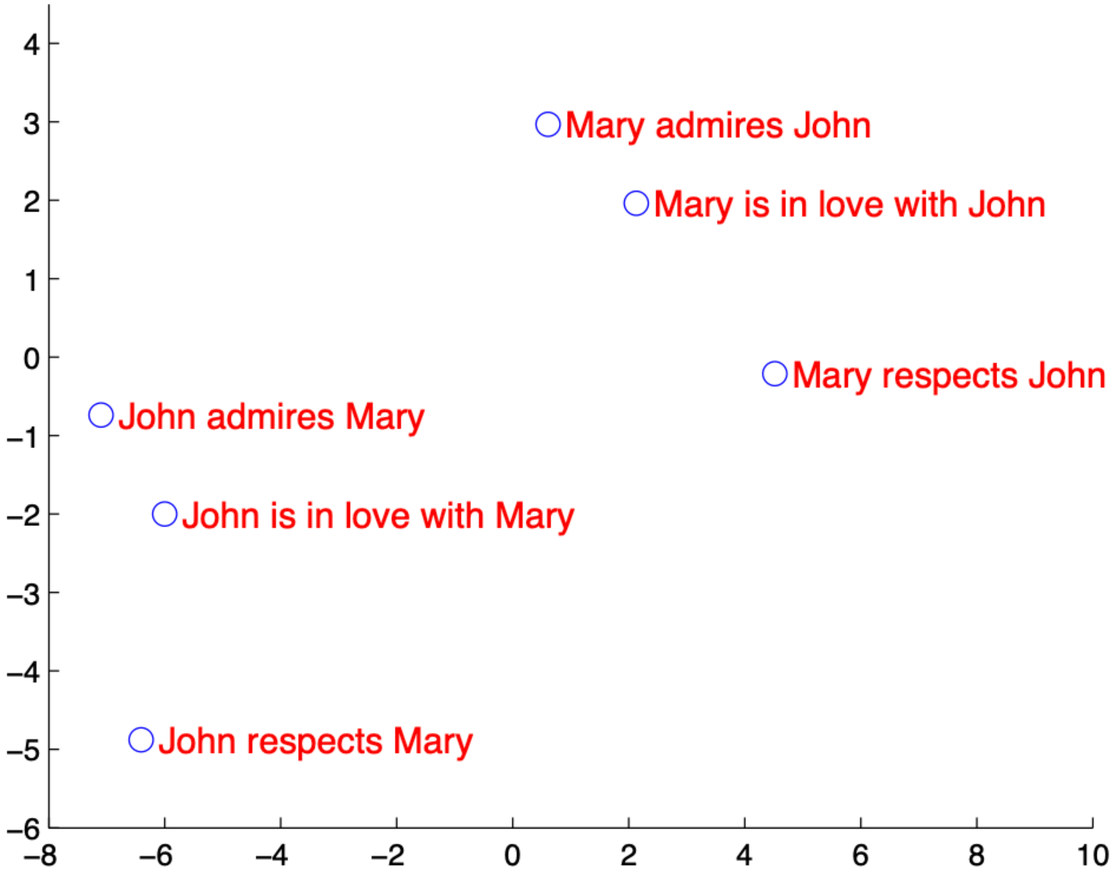
$$\mathbf{x} = (x_1, \cdots, x_{T_x}).$$

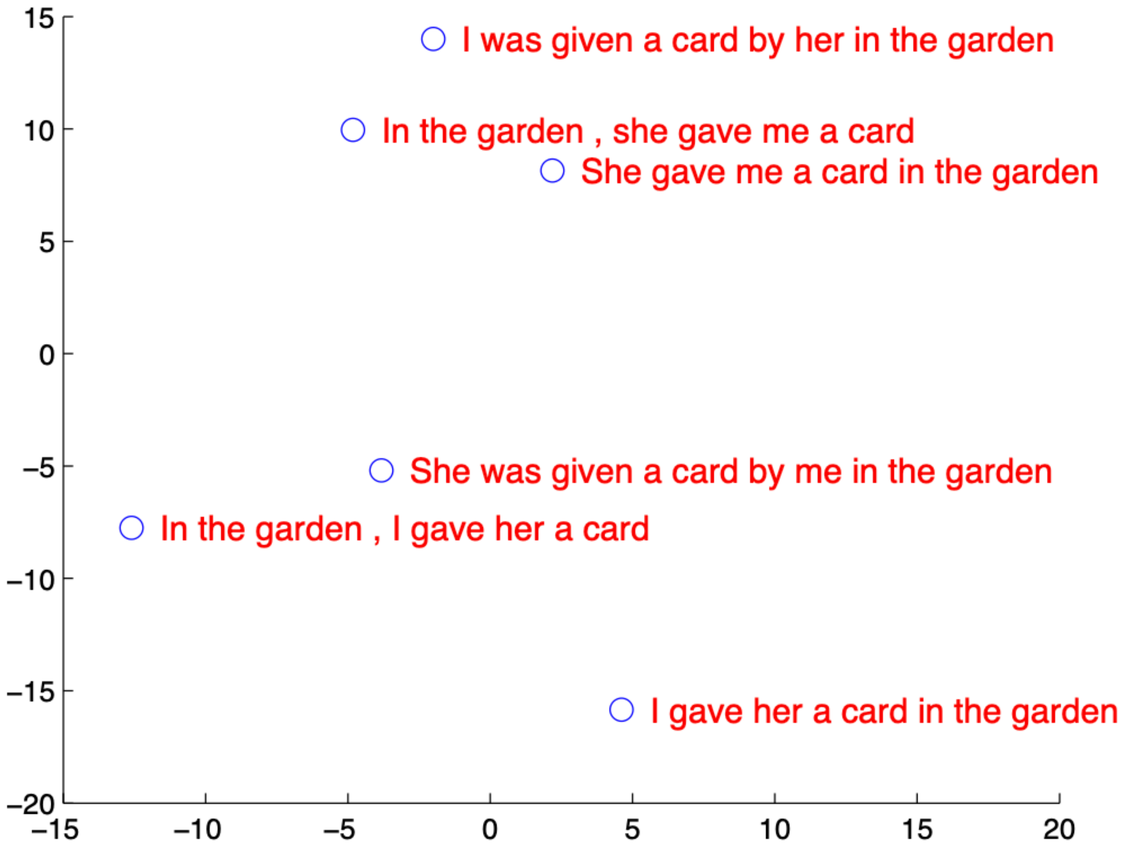
$$c = q\left(\{h_1, \cdots, h_{T_x}\}\right)$$

h_i is a vector representation of x_i

c is a vector assembled from all the h_i vectors

s_i is a vector representation of y_i





Sequence to Sequence Learning with Neural Networks https://arxiv.org/abs/1409.3215

Attention Networks in Neural Machine Translation

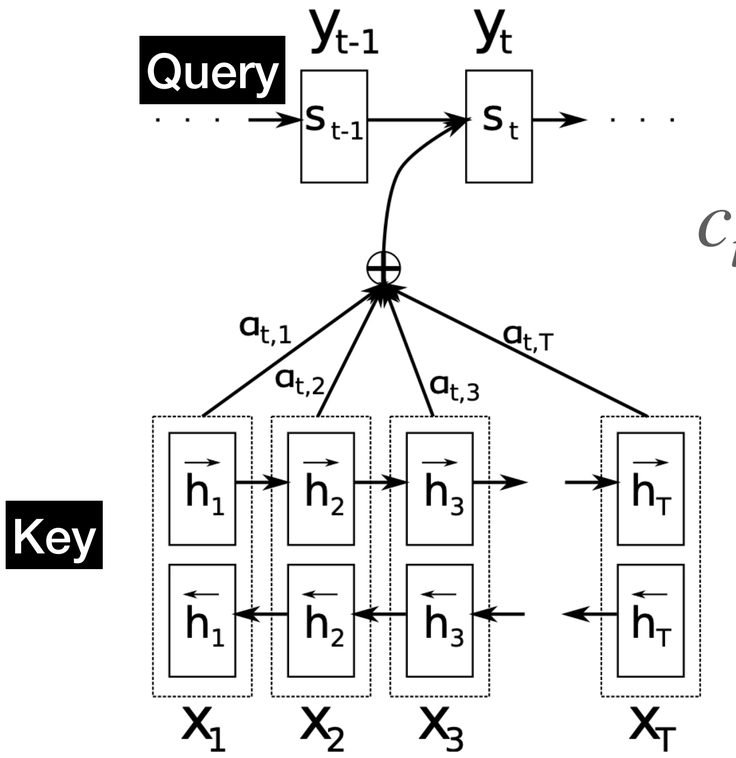


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Value

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j$$

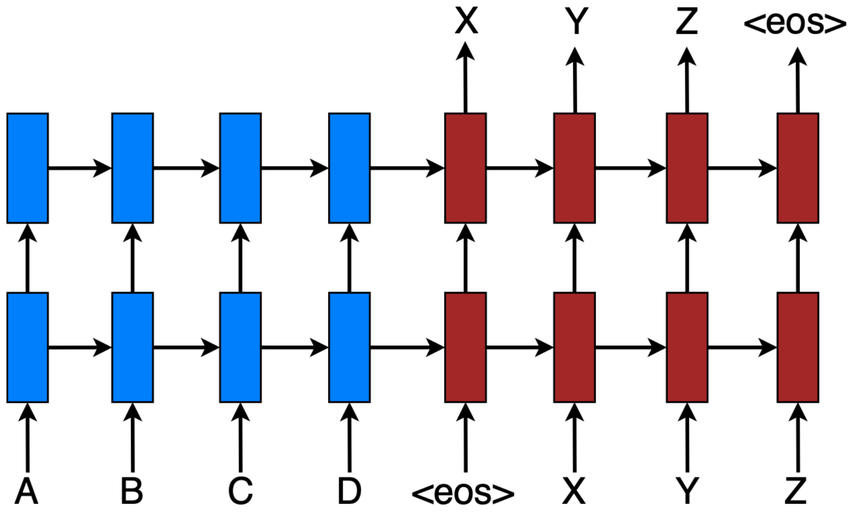
$$e_{tj} = a(s_{t-1}, h_j)$$

Dot Product / MLP

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T_x} \exp(e_{tk})}$$

softmax

Effective Approaches to Attention-based Neural Machine Translation

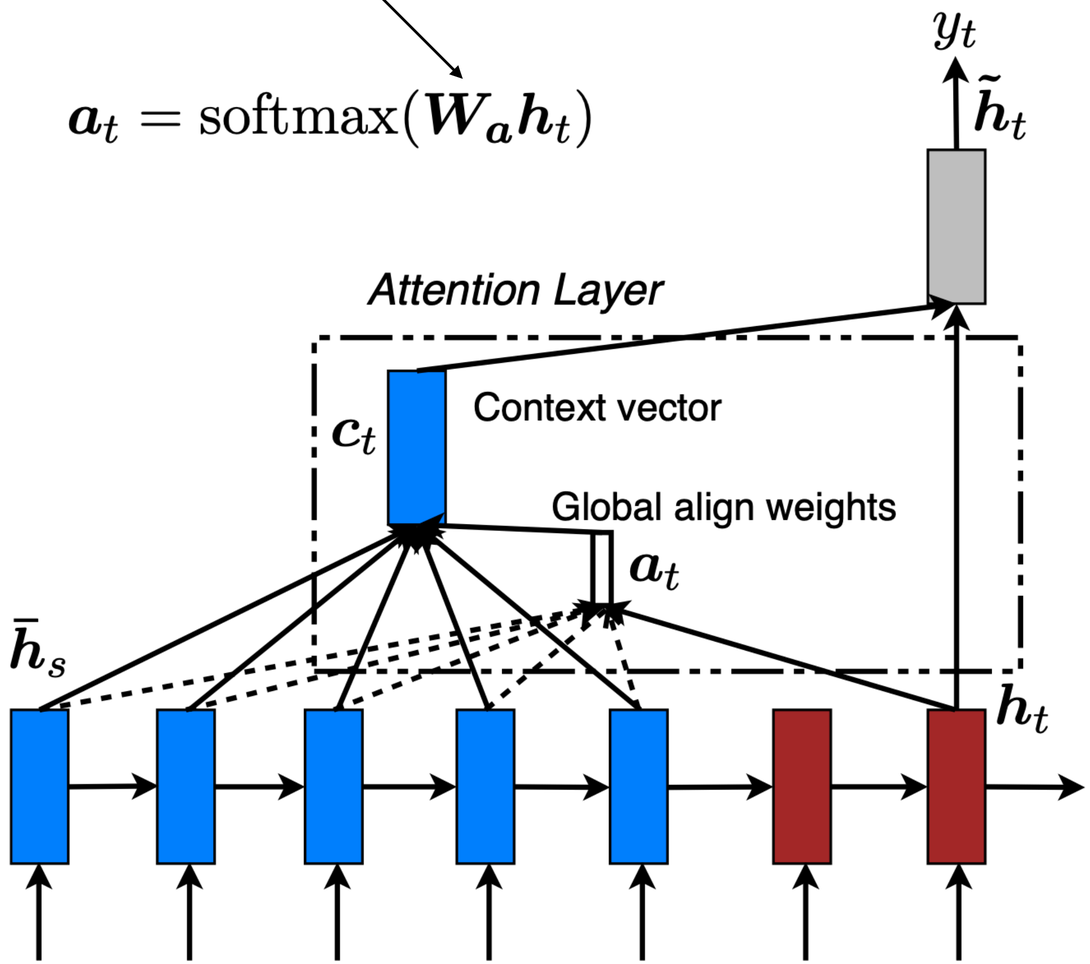


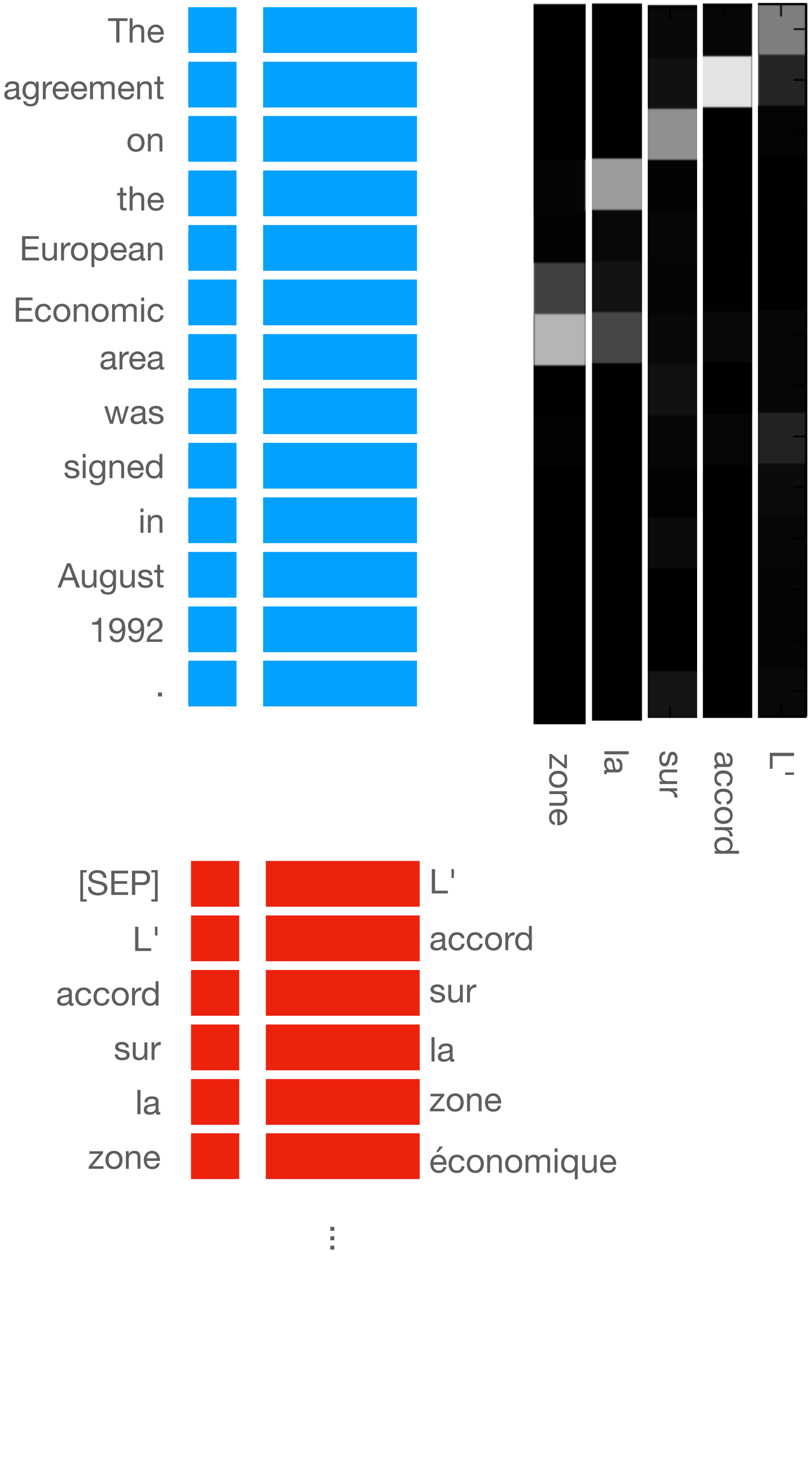
$$\frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

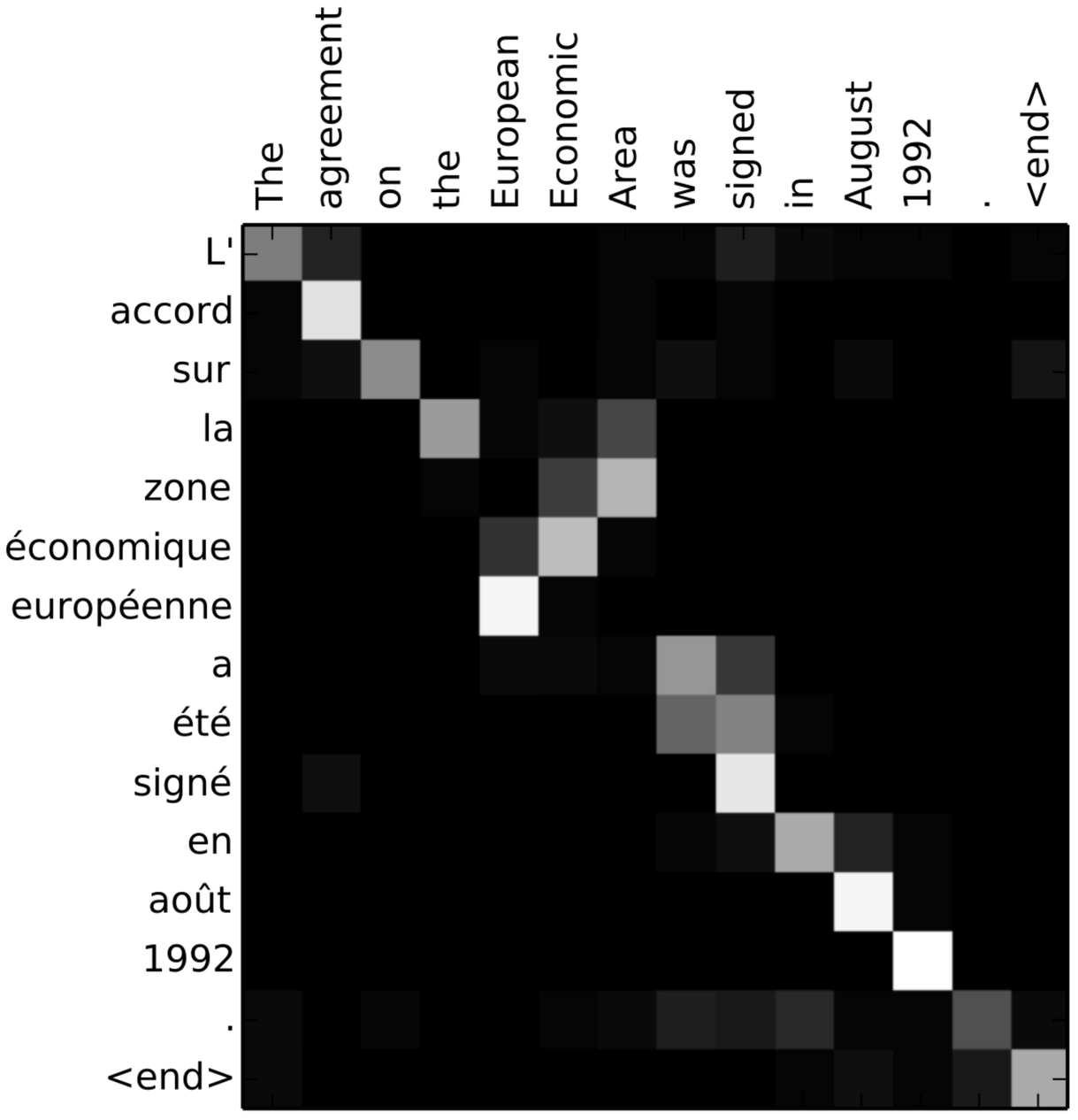
$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{W_a} \boldsymbol{h}_t)$$

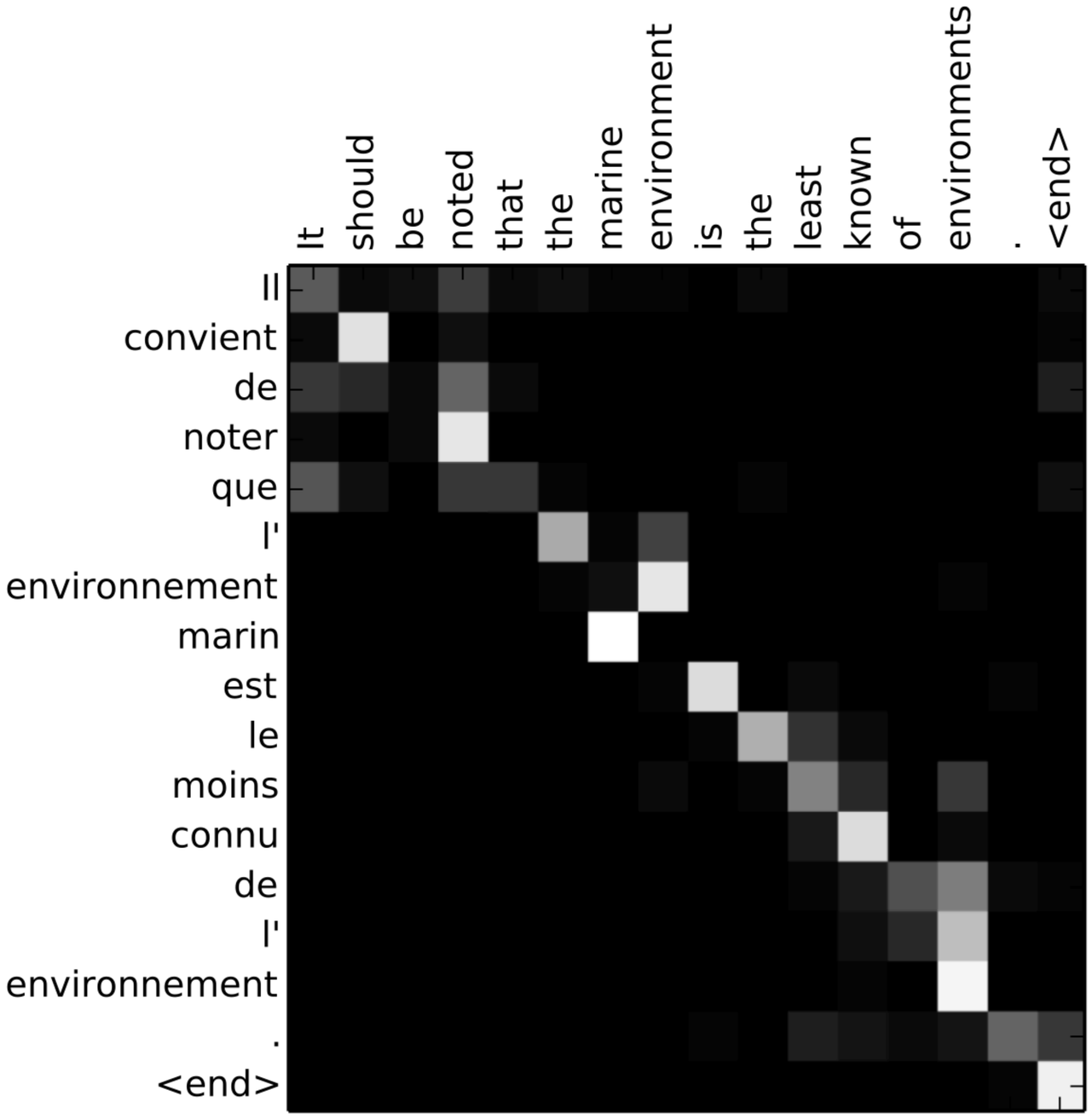
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

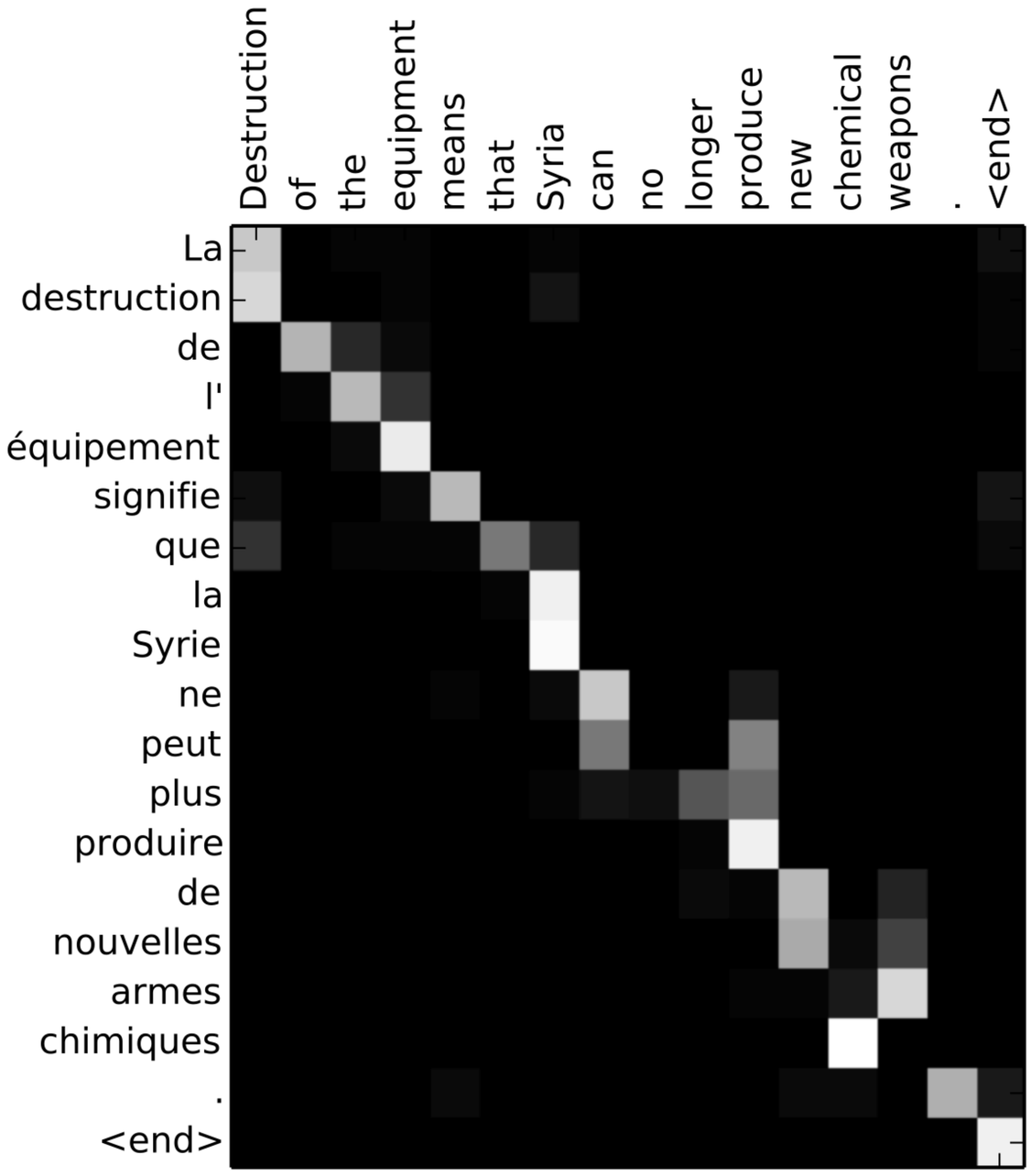
https://arxiv.org/abs/1508.04025

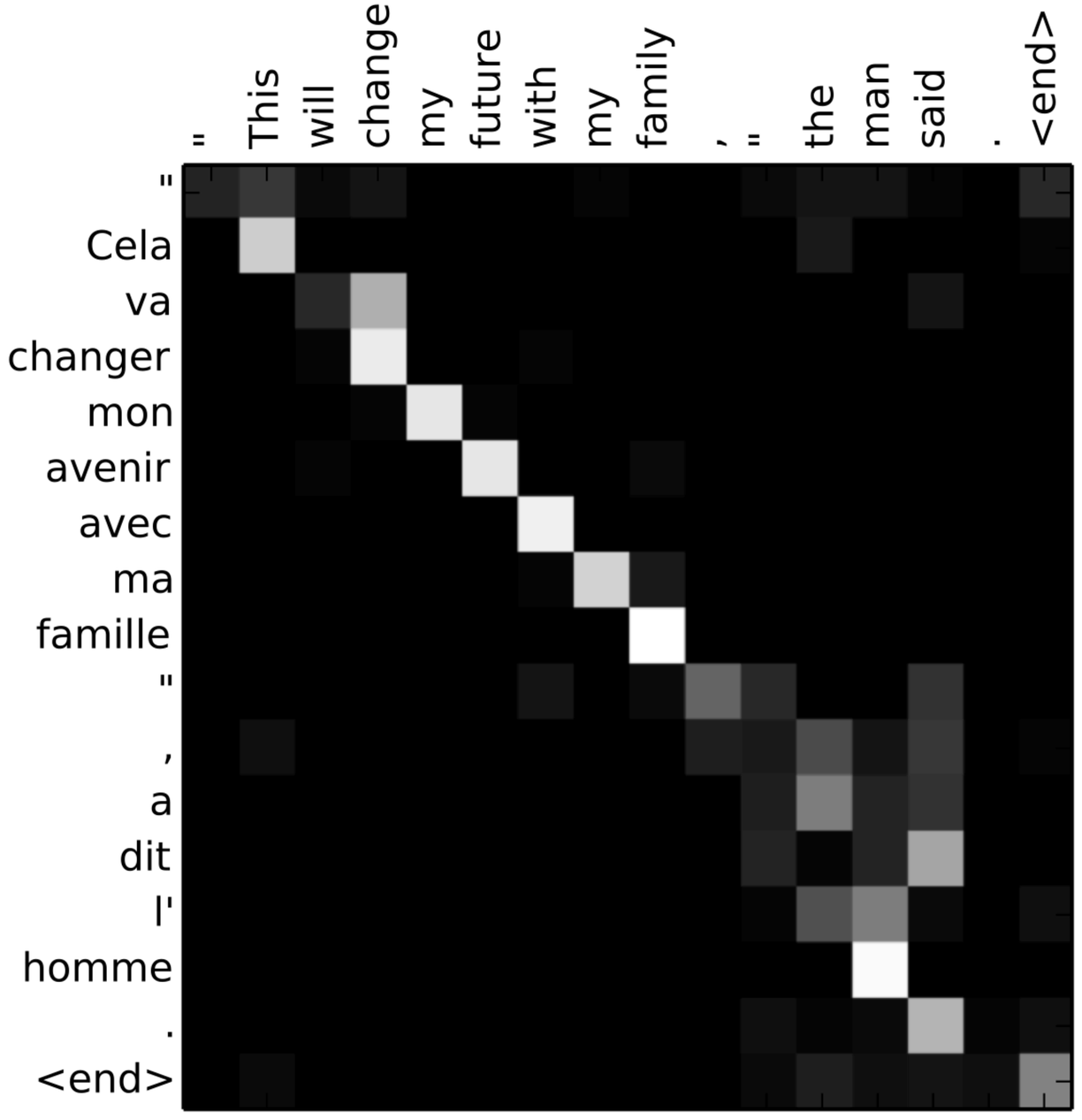



The
agreement
on
the
European
Economic
area
was
signed
in
August
1992
.
zone
la
sur
accord
L'
[SEP]
L'
accord
sur
la
zone
L'
accord
sur
la
zone
économique
⋮


The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>


It
should
be
noted
that
the
marine
environment
is
the
least
known
of
environments
.
<end>
Il
convient
de
noter
que
l'
environnement
marin
est
le
moins
connu
de
l'
environnement
.
<end>


Destruction
of
the
equipment
means
that
Syria
can
no
longer
produce
new
chemical
weapons
.
<end>
La
destruction
de
l'
équipement
signifie
que
la
Syrie
ne
peut
plus
produire
de
nouvelles
armes
chimiques
.
<end>


"
This
will
change
my
future
with
my
family
,
"
the
man
said
.
<end>
"
Cela
va
changer
mon
avenir
avec
ma
famille
"
,
a
dit
l'
homme
.
<end>

Source

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

No attention

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

With attention

Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.

Source

This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.

No attention

Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.

With attention

Ce genre d'expérience fait partie des efforts de Disney pour "prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes", a-t-il ajouté.

LayerNorm

https://arxiv.org/abs/1607.06450

also see: https://arxiv.org/abs/1911.07013

$$\mu = \frac{1}{H}\sum_{i=1}^{H} x_i \qquad \sigma^2 = \frac{1}{H}\sum_{i=1}^{H} (x_i - \mu)^2$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_H)$$

$$N(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma + \epsilon}$$

ϵ avoids div by zero

$$\mathbf{h} = \mathbf{g} \cdot N(\mathbf{x}) + \mathbf{b}$$

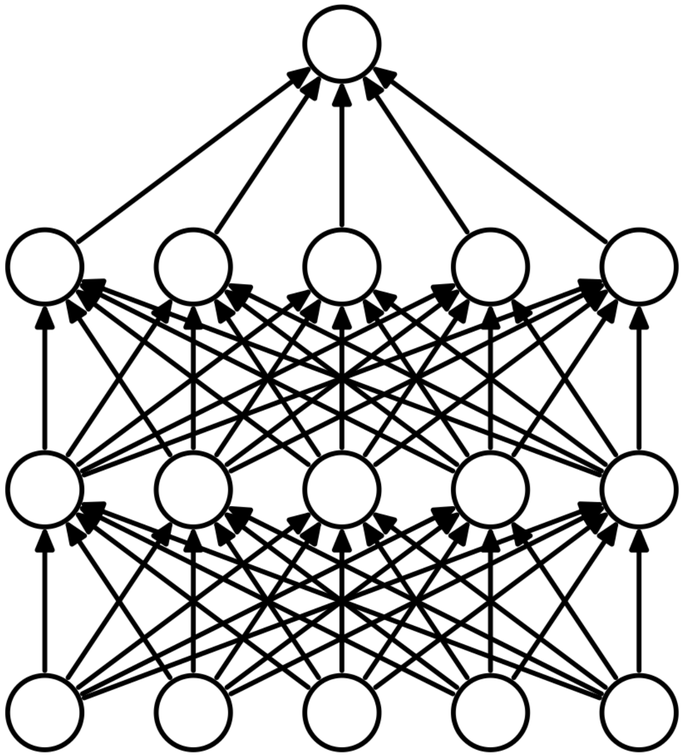
g and **b** are hyperparameters with dimension H

In PyTorch

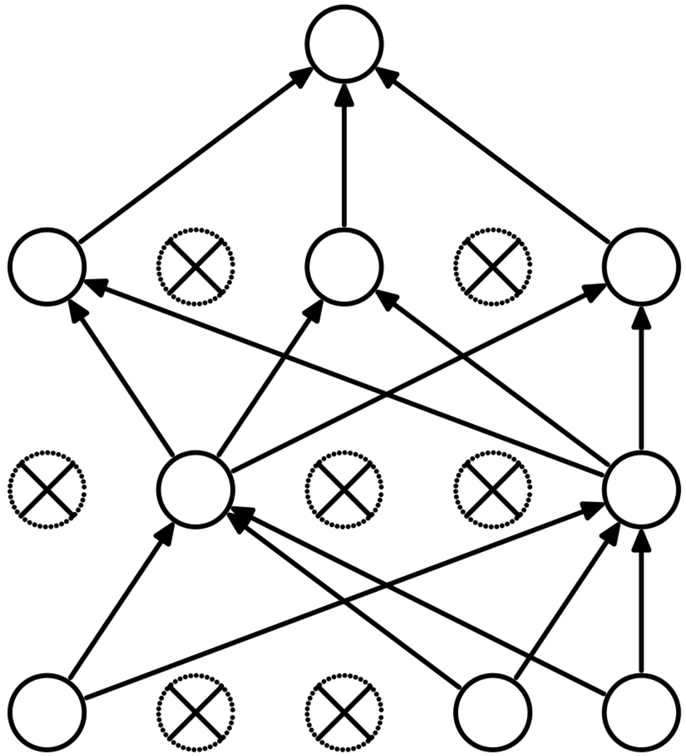
```
>>> # NLP Example
>>> batch, sentence_length, embedding_dim = 20, 5, 10
>>> embedding = torch.randn(batch, sentence_length, embedding_dim)
>>> layer_norm = nn.LayerNorm(embedding_dim)
>>> # Activate module
>>> layer_norm(embedding)
```

Dropout

https://jmlr.org/papers/v15/srivastava14a.html



(a) Standard Neural Net



(b) After applying dropout.

Before dropout

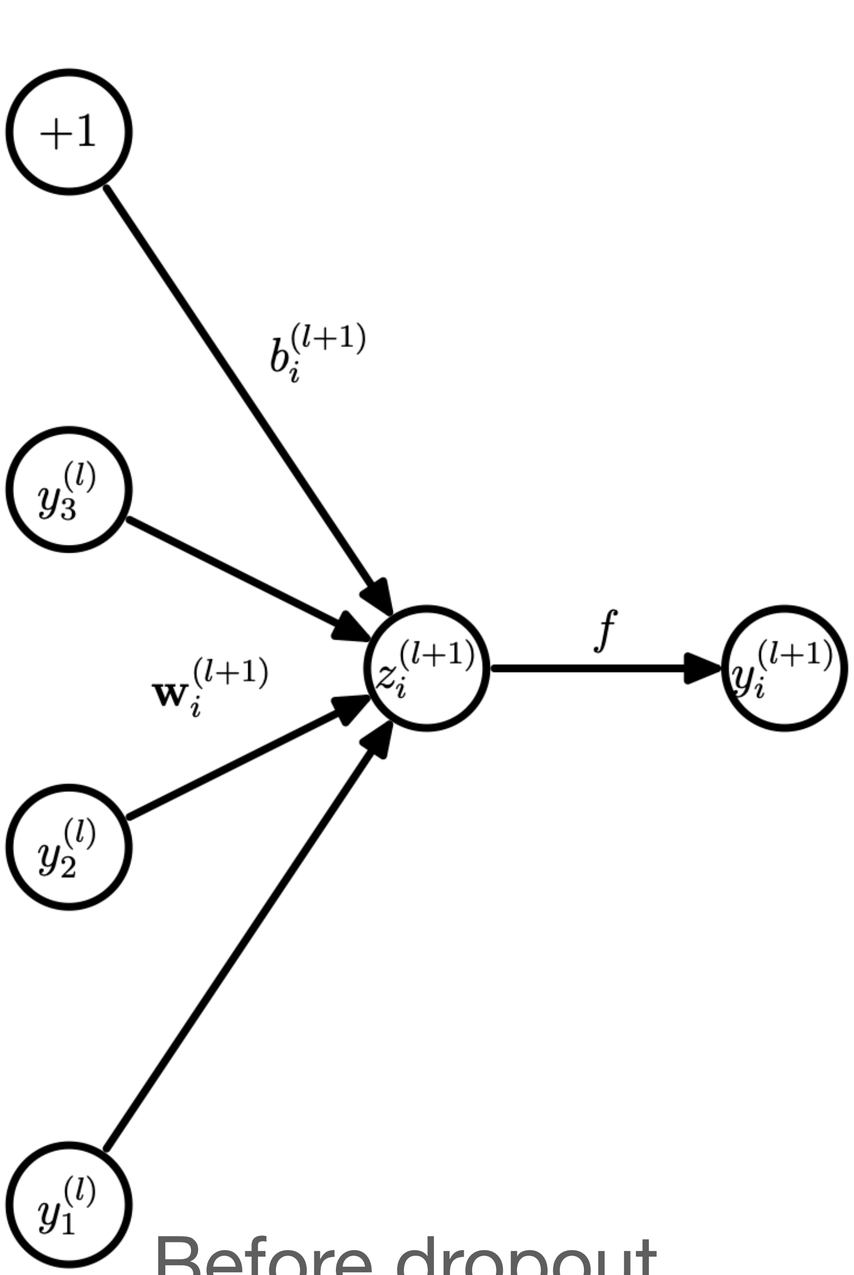
$$
\begin{aligned}
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)} \mathbf{y}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}),
\end{aligned}
$$

After dropout

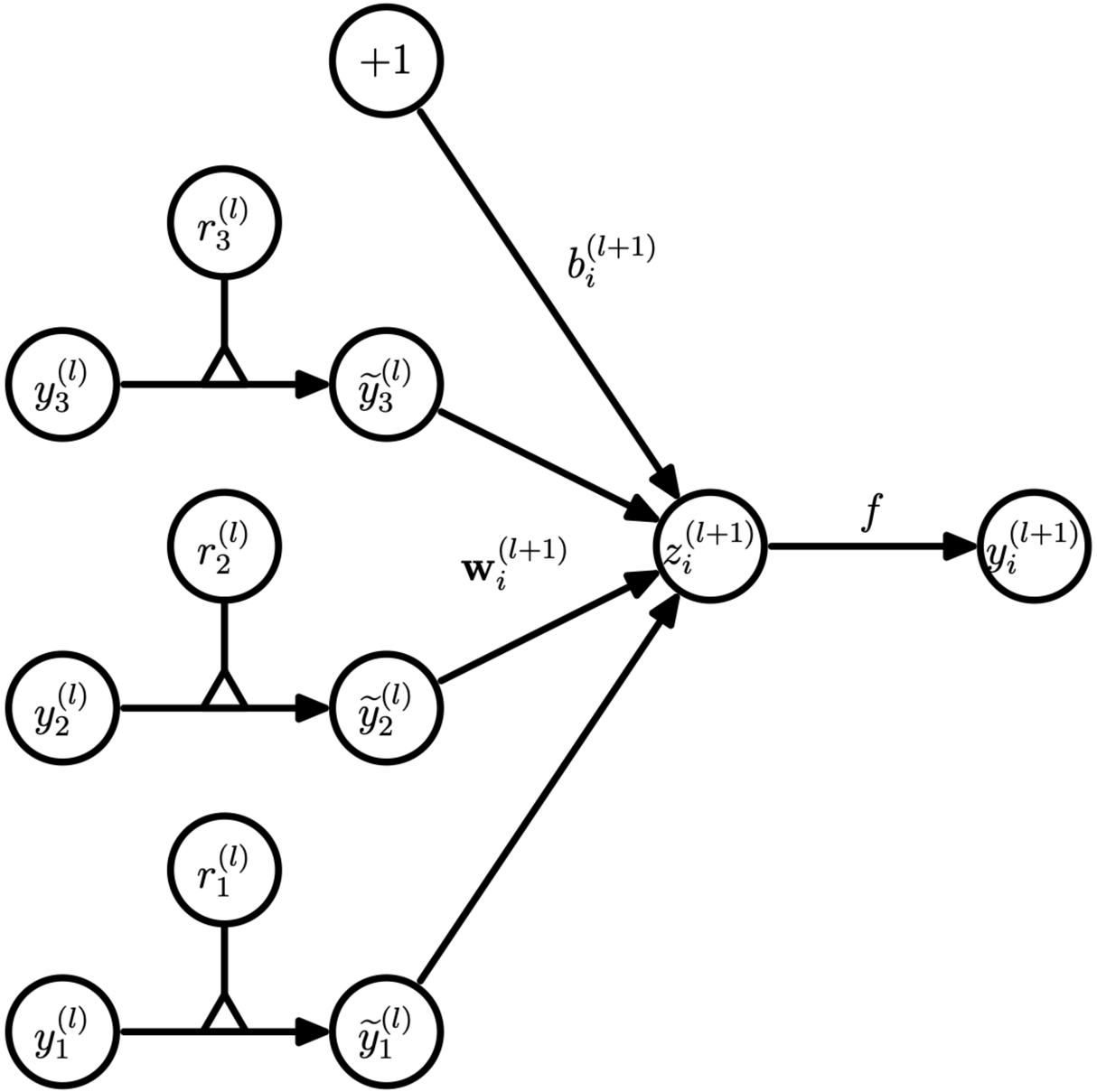
$$
\begin{aligned}
r_j^{(l)} &\sim \text{Bernoulli}(p), \\
\widetilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} * \mathbf{y}^{(l)}, \\
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)} \widetilde{\mathbf{y}}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}).
\end{aligned}
$$

In PyTorch

```
>>> m = nn.Dropout(p=0.2)
>>> input = torch.randn(20, 16)
>>> output = m(input)
```



Before dropout



After dropout

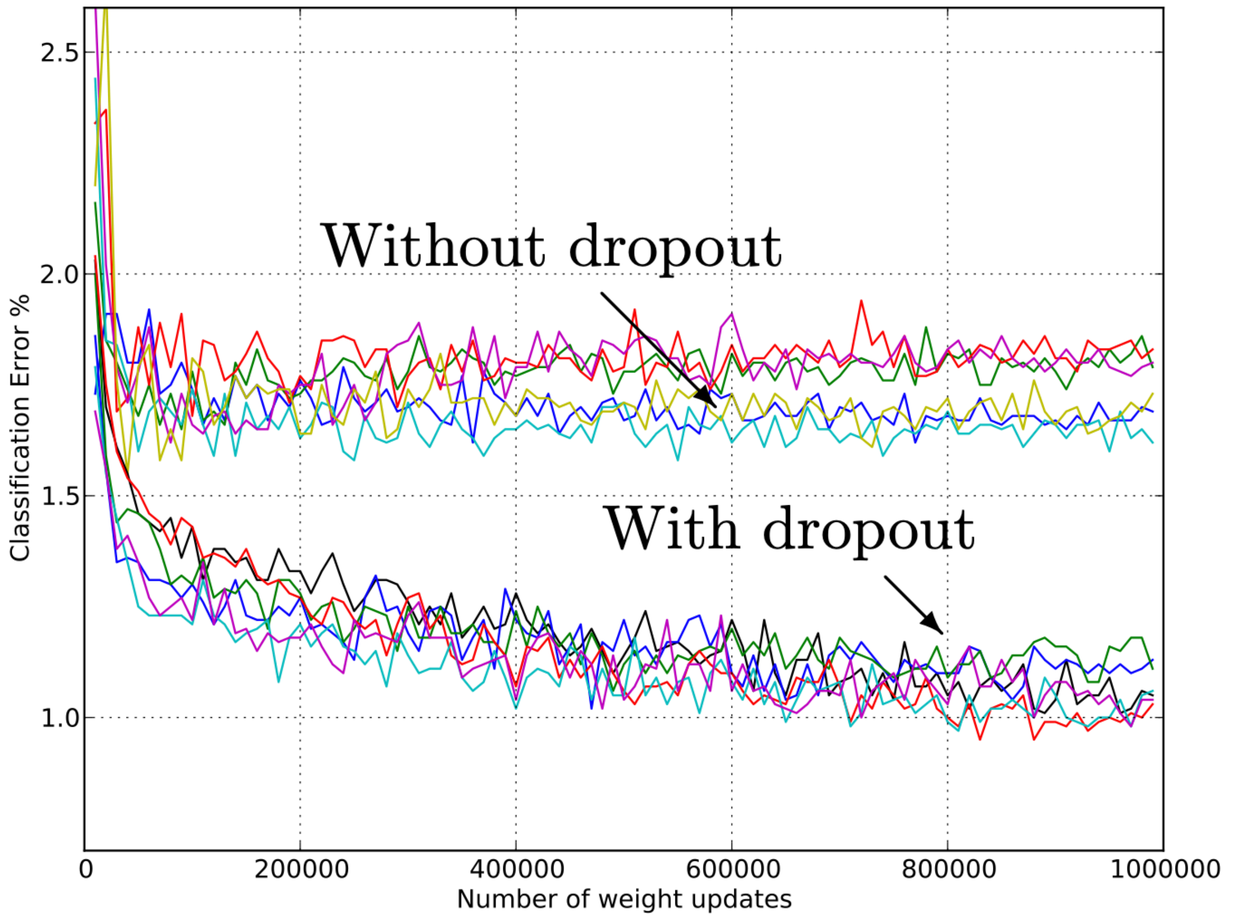
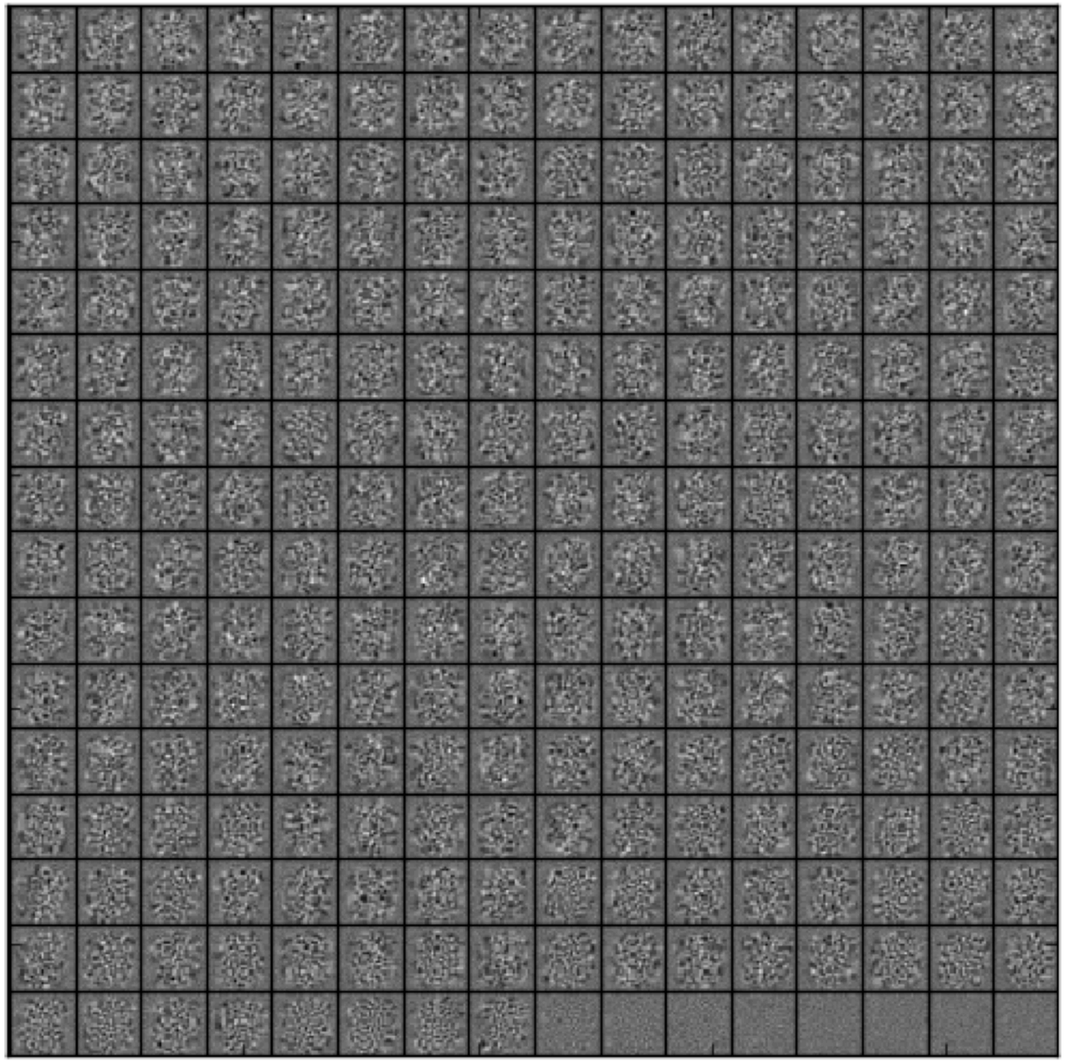
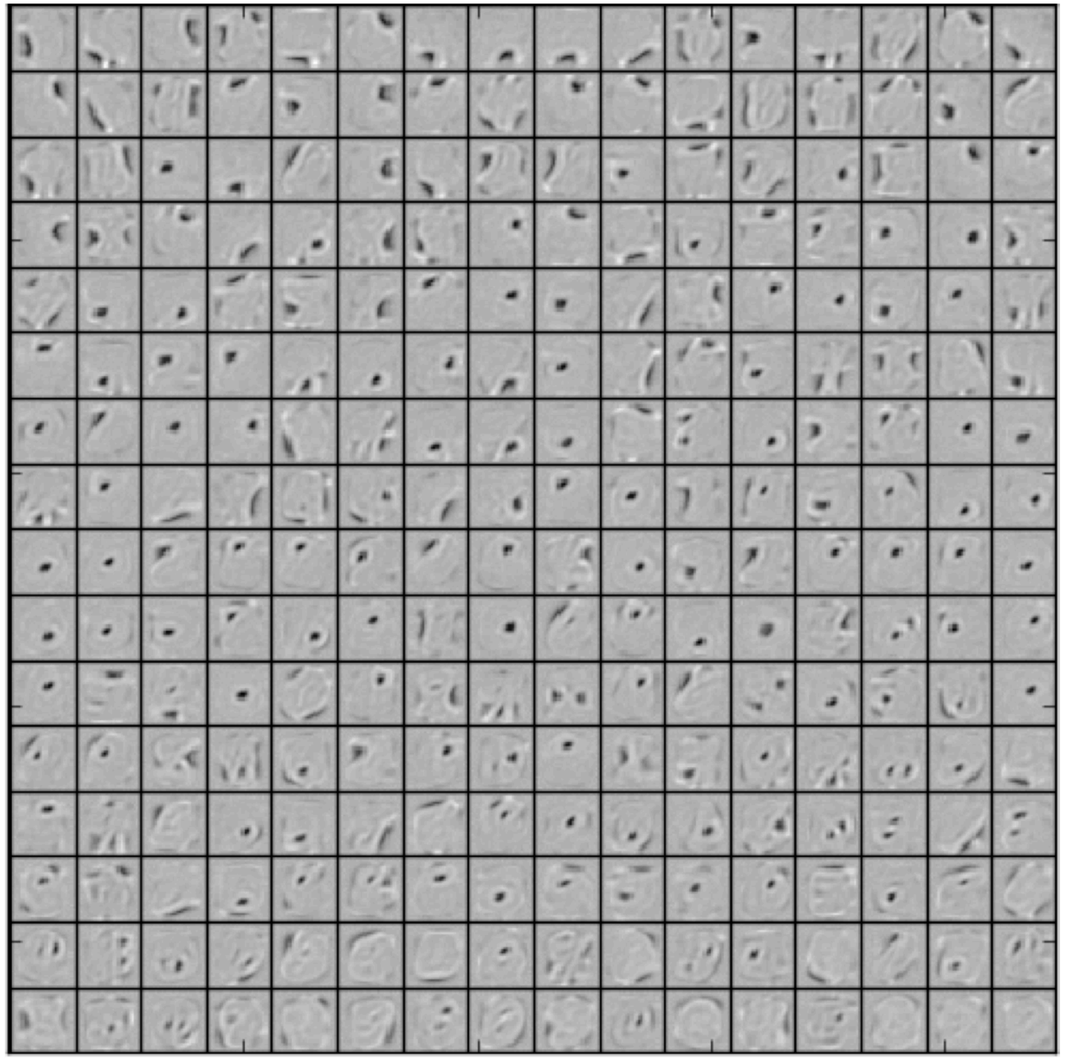


Figure 4: Test error for different architectures with and without dropout. The networks have 2 to 4 hidden layers each with 1024 to 2048 units.

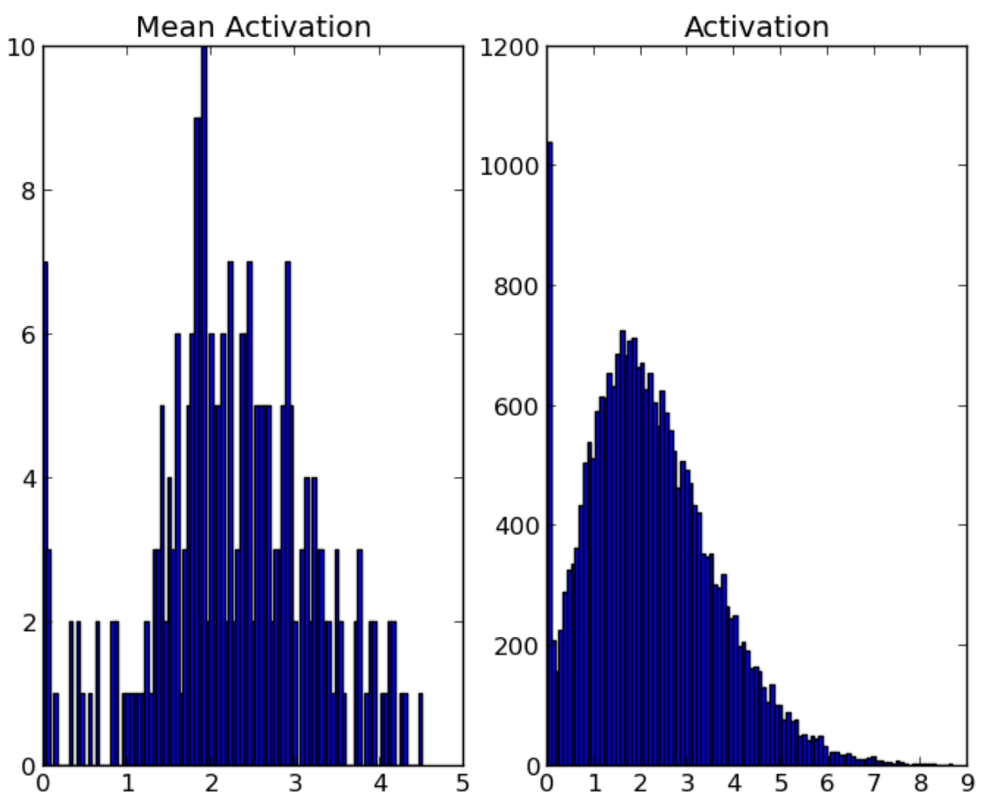


(a) Without dropout

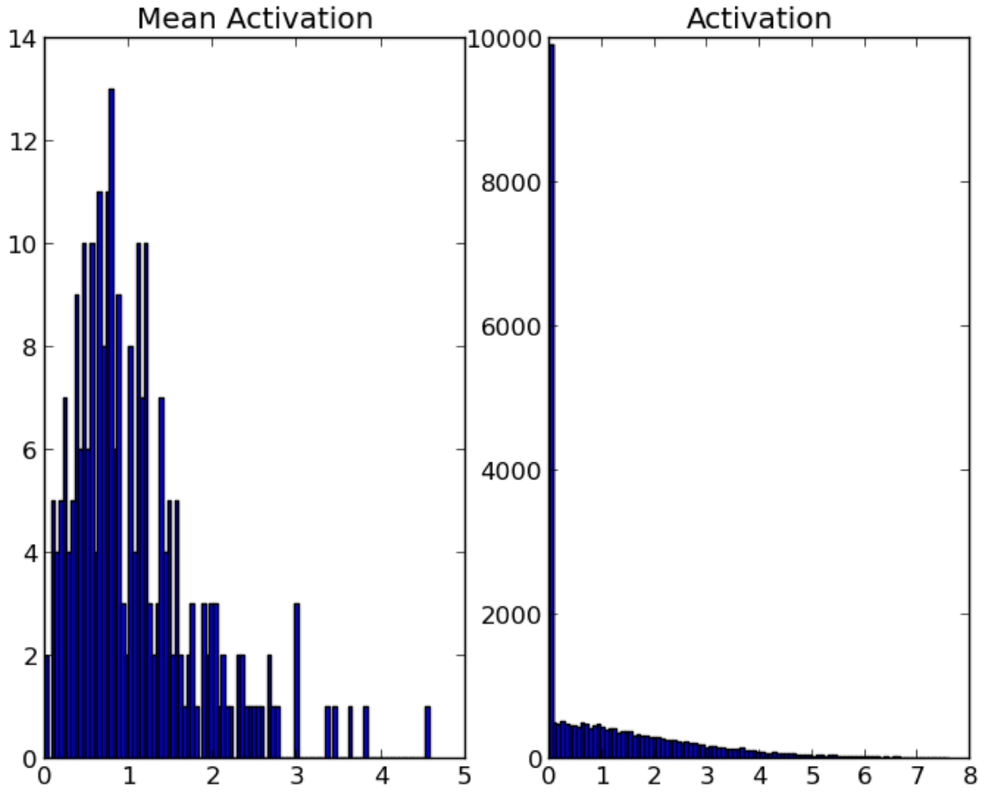


(b) Dropout with $p = 0.5$.

Figure 7: Features learned on MNIST with one hidden layer autoencoders having 256 rectified linear units.



(a) Without dropout



(b) Dropout with $p = 0.5$.

Figure 8: Effect of dropout on sparsity. ReLUs were used for both models. **Left**: The histogram of mean activations shows that most units have a mean activation of about 2.0. The histogram of activations shows a huge mode away from zero. Clearly, a large fraction of units have high activation. **Right**: The histogram of mean activations shows that most units have a smaller mean mean activation of about 0.7. The histogram of activations shows a sharp peak at zero. Very few units have high activation.