# LSTM: A Search Space Odyssey

Authors: Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber
Presenter: Sidhartha Satapathy

# Scientific contributions of the paper:

- The paper aims at evaluating different elements of the most popular LSTM architecture.
- The paper shows the performance of various variants of the vanilla LSTM by making a single change which allows us to isolate the effect of each of these changes on the performance of the architecture.
- The paper also provide insights gained about hyperparameters and their interaction.

# Dataset 1: IAM Online Handwriting Database

- IAM Online Handwriting Database: The IAM Handwriting Database contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments.

In mid-april Anglesey moved his family and entourage from Rome to Naples, there to await the arrival of

Each sequence or line in this case is made up of frames and the task at hand is to classify each of these frames into one of the 82 characters.

Here are the output characters:
abcdefghijklmnopqrstuvwxyz
ABCDEFGHIJKLMNOPQRSTUVWXYZ
0123456789 !"#&\'()*+,-./[]:;? And the empty symbol.

The performance in this case is the character error rate.

# Dataset 2: TIMIT

- TIMIT Speech corpus: **TIMIT** is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects.

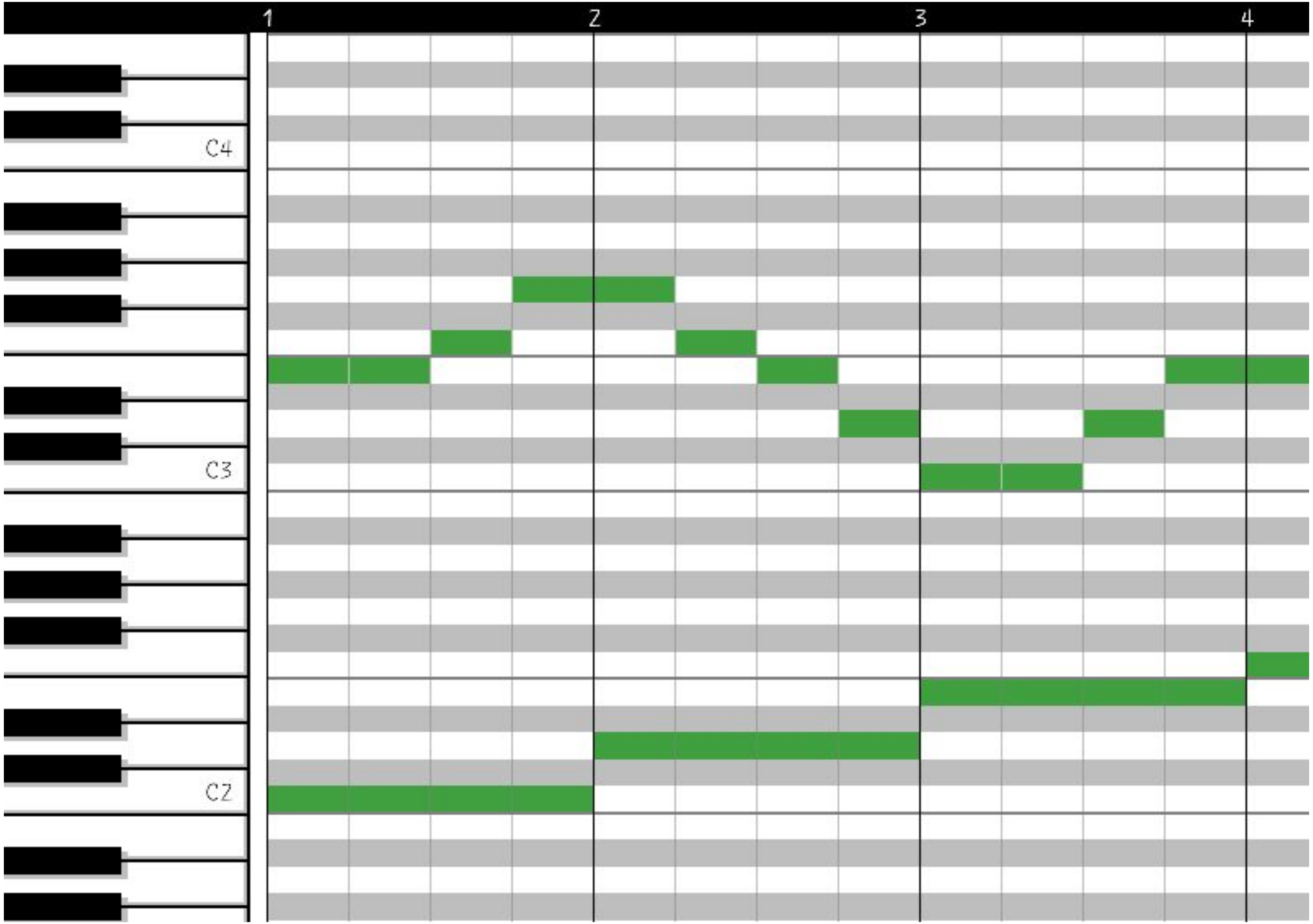
- Our experiments focus on the frame-wise classification task for this dataset, where the objective is to classify each audio-frame as one of 61 phones.
- The performance in this case is the classification error rate.

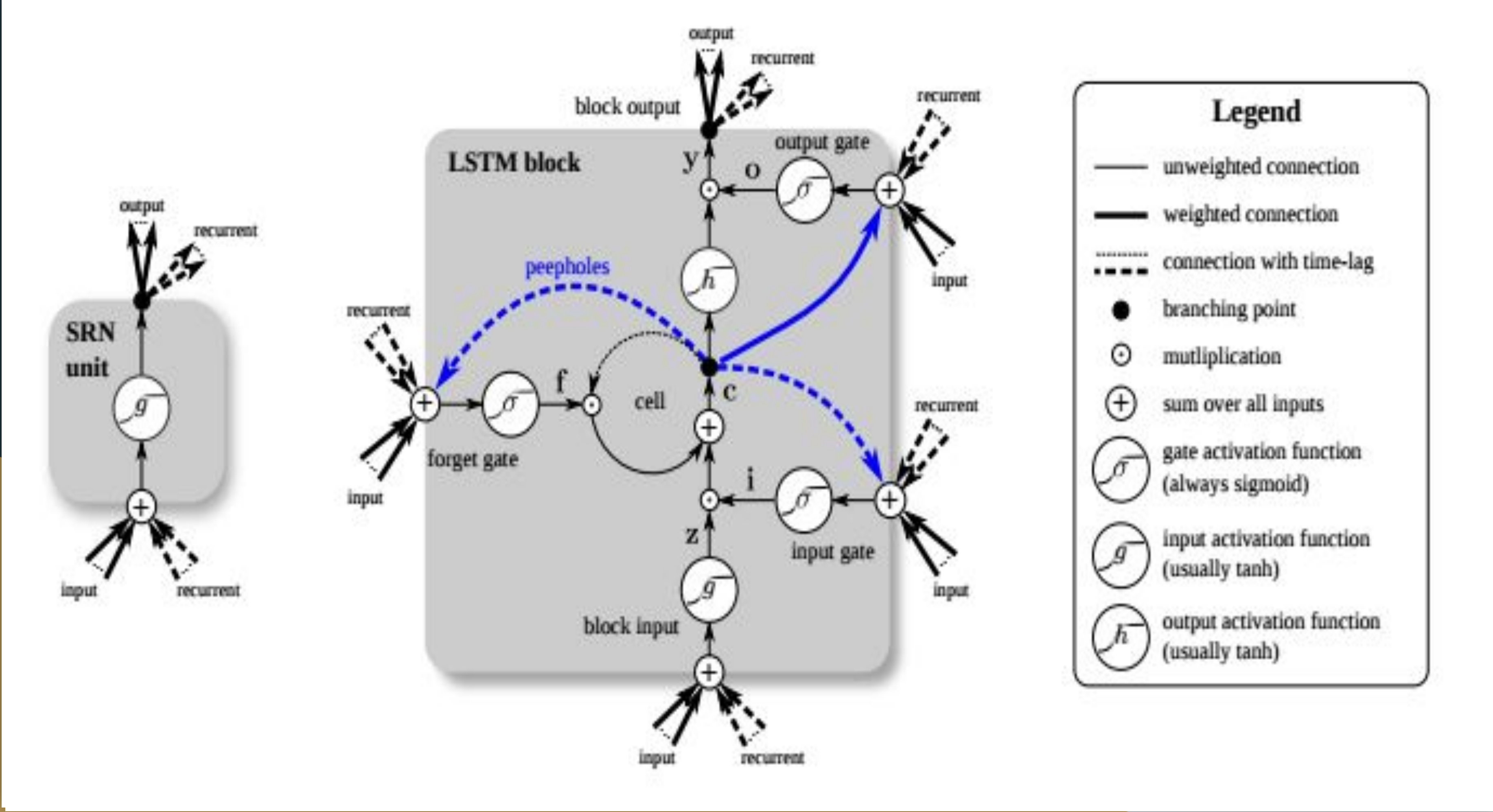
| | Phone Label | Example | | Phone Label | Example | | Phone Label | Example |
|---|---|---|---|---|---|---|---|---|
| 1 | iy | beet | 22 | ch | *choke* | 43 | en | butt*on* |
| 2 | ih | bit | 23 | b | *bee* | 44 | eng | Washi*ng*ton |
| 3 | eh | bet | 24 | d | *day* | 45 | l | *lay* |
| 4 | ey | bait | 25 | g | *gay* | 46 | r | *ray* |
| 5 | ae | bat | 26 | p | *pea* | 47 | w | *way* |
| 6 | aa | bob | 27 | t | *tea* | 48 | y | *yacht* |
| 7 | aw | bout | 28 | k | *key* | 49 | hh | *hay* |
| 8 | ay | bite | 29 | dx | mu*dd*y | 50 | hv | a*h*ead |
| 9 | ah | but | 30 | s | *sea* | 51 | el | bott*le* |
| 10 | ao | bought | 31 | sh | *she* | 52 | bcl | b closure |
| 11 | oy | boy | 32 | z | *zone* | 53 | dcl | d closure |
| 12 | ow | boat | 33 | zh | azure | 54 | gcl | g closure |
| 13 | uh | book | 34 | f | *fin* | 55 | pcl | p closure |
| 14 | uw | boot | 35 | th | *thin* | 56 | tcl | t closure |
| 15 | ux | toot | 36 | v | *van* | 57 | kcl | k closure |
| 16 | er | bird | 37 | dh | *then* | 58 | q | glotal stop |
| 17 | ax | *a*bout | 38 | m | *mom* | 59 | pau | pause |
| 18 | ix | debit | 39 | n | *noon* | 60 | epi | epenthetic silence |
| 19 | axr | butt*er* | 40 | ng | si*ng* | 61 | h# | begin/end marker |
| 20 | ax-h | s*u*spect | 41 | em | bott*om* | | | |
| 21 | jh | *j*oke | 42 | nx | wi*nn*er | | | |

Table 2. 61 TIMIT original phone set.

# Dataset 3: JSB Chorales

- JSB Chorales:  JSB Chorales is a collection of 382 four part harmonized chorales by J. S. Bach, the networks where trained to do next-step prediction.


1
2
3
4
C4
C3
C2


output
recurrent
SRN
unit
input
recurrent
output
recurrent
block output
LSTM block
y
o
output gate
recurrent
input
peepholes
recurrent
f
cell
c
forget gate
input
i
recurrent
input gate
input
z
block input
input
recurrent
Legend
unweighted connection
weighted connection
connection with time-lag
branching point
mutliplication
sum over all inputs
gate activation function
(always sigmoid)
input activation function
(usually tanh)
output activation function
(usually tanh)

$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z\mathbf{x}^t + \mathbf{R}_z\mathbf{y}^{t-1} + \mathbf{b}_z \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) && \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i\mathbf{x}^t + \mathbf{R}_i\mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) && \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f\mathbf{x}^t + \mathbf{R}_f\mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) && \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t && \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o\mathbf{x}^t + \mathbf{R}_o\mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) && \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t && \textit{block output}
\end{aligned}
$$

# Variants of the LSTM Block:

- NIG: No Input Gate
- NFG: No Forget Gate
- NOG: No Output Gate
- NIAF: No Input Activation Function
- NOAF: No Output Activation Function
- CIFG: Coupled Input and Forget Gate
- NP: No Peepholes
- FGR: Full Gate Recurrence

# NIG: No Input Gate

**NIG: No Input Gate: $\mathbf{i}^t = 1$**

$$\bar{\mathbf{z}}^t = \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z$$

$$\mathbf{z}^t = g(\bar{\mathbf{z}}^t) \qquad \textit{block input}$$

$$\bar{\mathbf{i}}^t = \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i$$

$$\mathbf{i}^t = \sigma(\bar{\mathbf{i}}^t) \qquad \textit{input gate}$$

$$\bar{\mathbf{f}}^t = \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f$$

$$\mathbf{f}^t = \sigma(\bar{\mathbf{f}}^t) \qquad \textit{forget gate}$$

$$\mathbf{c}^t = \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t \qquad \textit{cell}$$

$$\bar{\mathbf{o}}^t = \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o$$

$$\mathbf{o}^t = \sigma(\bar{\mathbf{o}}^t) \qquad \textit{output gate}$$

$$\mathbf{y}^t = h(\mathbf{c}^t) \odot \mathbf{o}^t \qquad \textit{block output}$$

# NFG: No Forget Gate

**NFG: No Forget Gate: $\mathbf{f}^t = 1$**

$$\bar{\mathbf{z}}^t = \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z$$

$$\mathbf{z}^t = g(\bar{\mathbf{z}}^t) \qquad \textit{block input}$$

$$\bar{\mathbf{i}}^t = \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i$$

$$\mathbf{i}^t = \sigma(\bar{\mathbf{i}}^t) \qquad \textit{input gate}$$

$$\bar{\mathbf{f}}^t = \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f$$

$$\mathbf{f}^t = \sigma(\bar{\mathbf{f}}^t) \qquad \textit{forget gate}$$

$$\mathbf{c}^t = \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t \qquad \textit{cell}$$

$$\bar{\mathbf{o}}^t = \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o$$

$$\mathbf{o}^t = \sigma(\bar{\mathbf{o}}^t) \qquad \textit{output gate}$$

$$\mathbf{y}^t = h(\mathbf{c}^t) \odot \mathbf{o}^t \qquad \textit{block output}$$

# NOG: No Output Gate

**NOG: No Output Gate: $\mathbf{o}^t = 1$**

$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z & \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) & \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i & \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) & \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f & \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) & \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t & \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o & \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) & \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t & \textit{block output}
\end{aligned}
$$

# NIAF: No Input Activation Function

**NIAF: No Input Activation Function:** $g(\mathbf{x}) = \mathbf{x}$

$$\bar{\mathbf{z}}^t = \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z$$

$$\mathbf{z}^t = g(\bar{\mathbf{z}}^t) \quad \textit{block input}$$

$$\bar{\mathbf{i}}^t = \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i$$

$$\mathbf{i}^t = \sigma(\bar{\mathbf{i}}^t) \quad \textit{input gate}$$

$$\bar{\mathbf{f}}^t = \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f$$

$$\mathbf{f}^t = \sigma(\bar{\mathbf{f}}^t) \quad \textit{forget gate}$$

$$\mathbf{c}^t = \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t \quad \textit{cell}$$

$$\bar{\mathbf{o}}^t = \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o$$

$$\mathbf{o}^t = \sigma(\bar{\mathbf{o}}^t) \quad \textit{output gate}$$

$$\mathbf{y}^t = h(\mathbf{c}^t) \odot \mathbf{o}^t \quad \textit{block output}$$

# NOAF: No Output Activation Function

**NOAF: No Output Activation Function:** $h(\mathbf{x}) = \mathbf{x}$

$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) && \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) && \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) && \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t && \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) && \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t && \textit{block output}
\end{aligned}
$$

# CIFG: Coupled Input and Forget Gate

**CIFG: Coupled Input and Forget Gate:** $\mathbf{f}^t = 1 - \mathbf{i}^t$

$$\bar{\mathbf{z}}^t = \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z$$

$$\mathbf{z}^t = g(\bar{\mathbf{z}}^t) \qquad \textit{block input}$$

$$\bar{\mathbf{i}}^t = \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i$$

$$\mathbf{i}^t = \sigma(\bar{\mathbf{i}}^t) \qquad \textit{input gate}$$

$$\bar{\mathbf{f}}^t = \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f$$

$$\mathbf{f}^t = \sigma(\bar{\mathbf{f}}^t) \qquad \textit{forget gate}$$

$$\mathbf{c}^t = \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t \qquad \textit{cell}$$

$$\bar{\mathbf{o}}^t = \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o$$

$$\mathbf{o}^t = \sigma(\bar{\mathbf{o}}^t) \qquad \textit{output gate}$$

$$\mathbf{y}^t = h(\mathbf{c}^t) \odot \mathbf{o}^t \qquad \textit{block output}$$

# NP: No Peepholes

**NP:** No Peepholes:

$$\bar{\mathbf{i}}^t = \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{b}_i$$
$$\bar{\mathbf{f}}^t = \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{b}_f$$
$$\bar{\mathbf{o}}^t = \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{b}_o$$

# NP: No Peepholes

$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) && \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) && \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) && \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t && \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) && \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t && \textit{block output}
\end{aligned}
$$

# FGR: Full Gate Recurrence

**FGR: Full Gate Recurrence:**

$$\begin{aligned}
\bar{\mathbf{i}}^t &= \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i \\
&\quad + \mathbf{R}_{ii} \mathbf{i}^{t-1} + \mathbf{R}_{fi} \mathbf{f}^{t-1} + \mathbf{R}_{oi} \mathbf{o}^{t-1} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f \\
&\quad + \mathbf{R}_{if} \mathbf{i}^{t-1} + \mathbf{R}_{ff} \mathbf{f}^{t-1} + \mathbf{R}_{of} \mathbf{o}^{t-1} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^{t-1} + \mathbf{b}_o \\
&\quad + \mathbf{R}_{io} \mathbf{i}^{t-1} + \mathbf{R}_{fo} \mathbf{f}^{t-1} + \mathbf{R}_{oo} \mathbf{o}^{t-1}
\end{aligned}$$

# FGR: Full Gate Recurrence

$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z & \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) & \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i & \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) & \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f & \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) & \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t & \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o & \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) & \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t & \textit{block output}
\end{aligned}
$$

# Hyperparameter Search

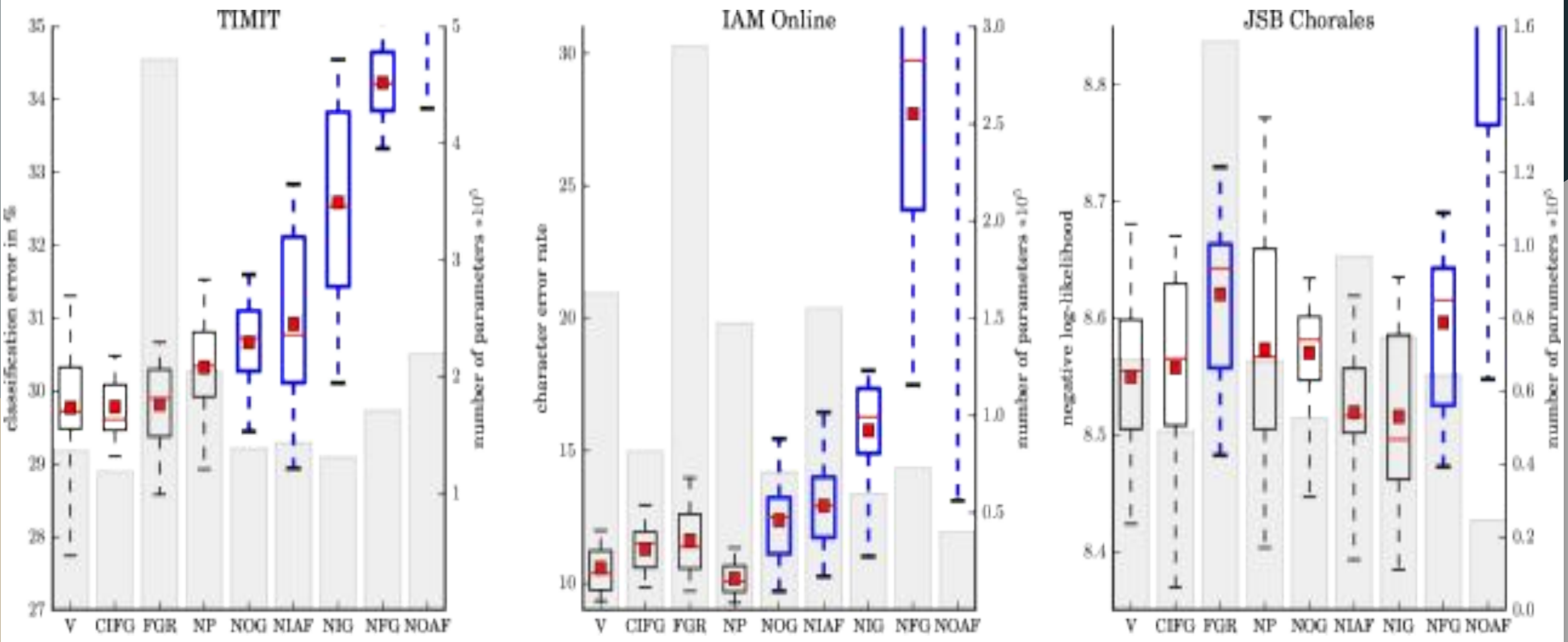
- While there are other methods to efficiently search for good hyperparameters, this paper uses random search has several advantages for our setting:
  - it is easy to implement
  - trivial to parallelize
  - covers the search space more uniformly, thereby improving the follow-up analysis of hyperparameter importance.

- The paper shows 27 random searches (one for each combination of the nine variants and three datasets). Each random search encompasses 200 trials for a total of 5400 trials of randomly sampling the hyperparameters.

- The hyperparameters and ranges are:
    - hidden layer size: log-uniform samples from [20; 200]
    - learning rate: log-uniform samples from [10^-6; 10^-2]
    - momentum: 1 - log-uniform samples from [0:01; 1:0]
    - standard deviation of Gaussian input noise: uniform samples from [0; 1].

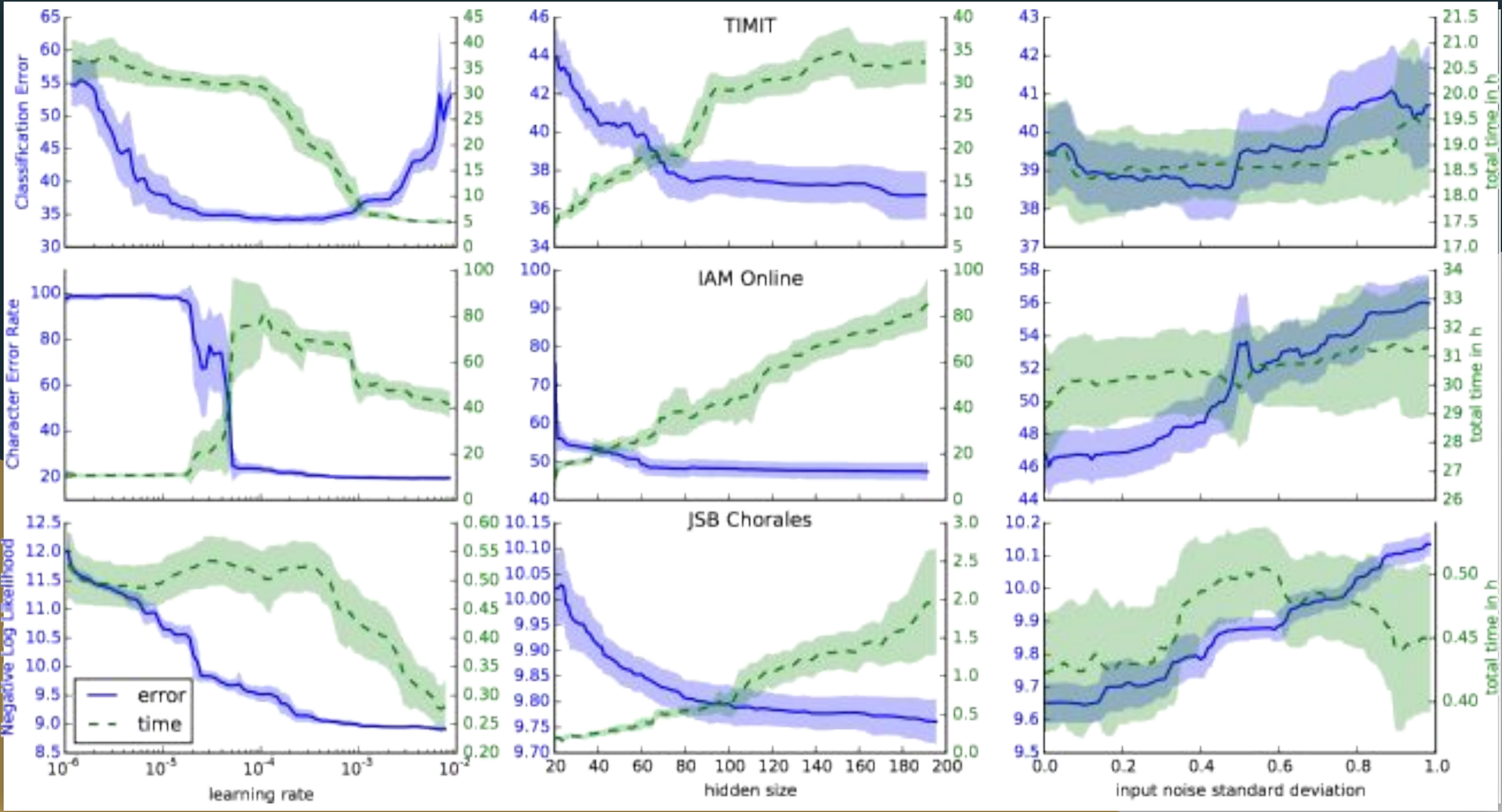
# Results and Discussions:

| Datasets: | State of the art: | Best result: |
|---|---|---|
| IAM Online | 26.9% (Best LSTM Result) | **9.26%** |
| TIMIT | **26.9%** | 29.6% |
| JSB Chorales | -5.56 | **-8.38** |


TIMIT
classification error in %
number of parameters ∗10^5
IAM Online
character error rate
JSB Chorales
negative log-likelihood
V
CIFG
FGR
NP
NOG
NIAF
NIG
NFG
NOAF

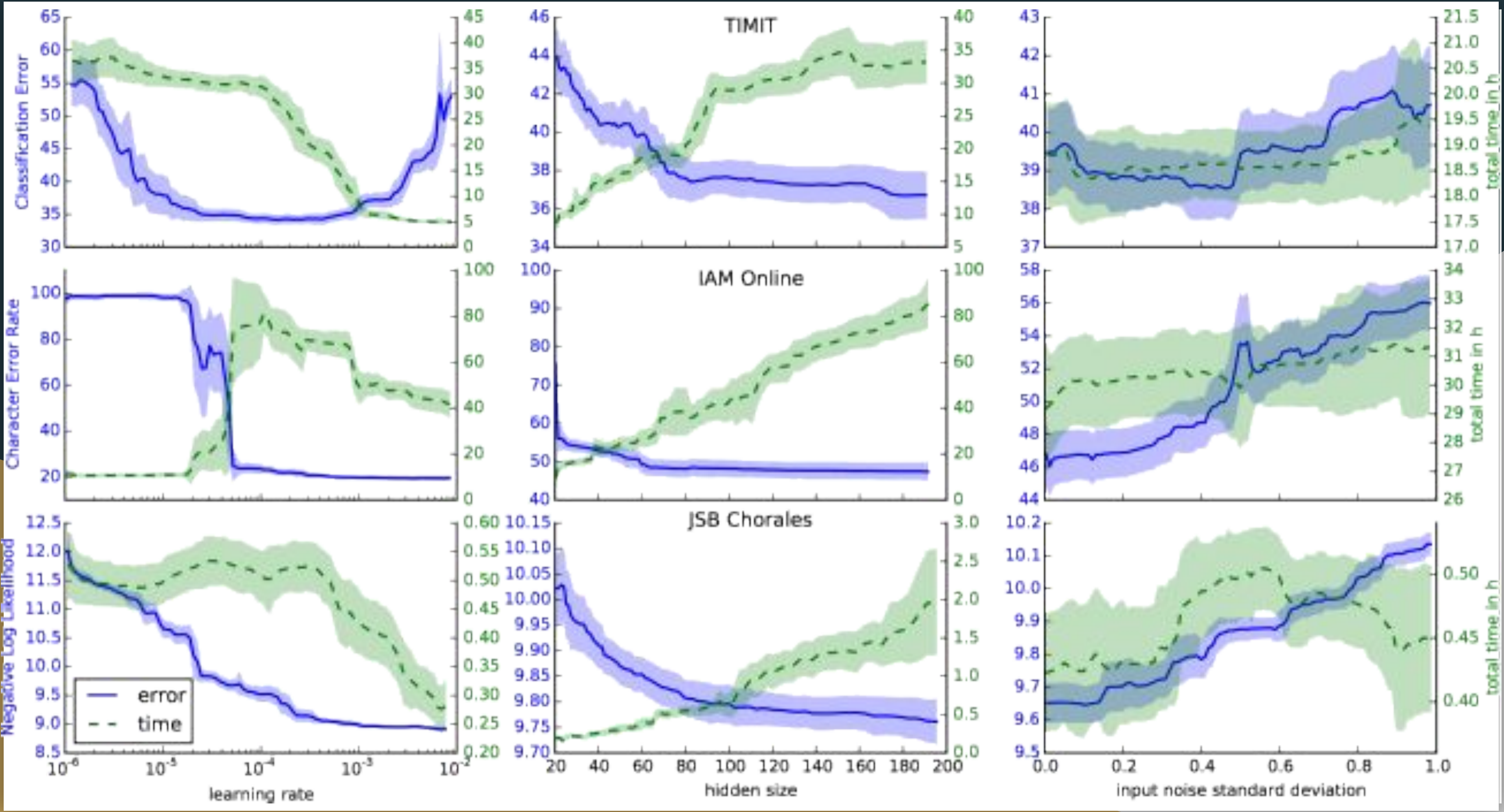
# Hyperparameter Analysis:

- Learning Rate: It is the most important hyperparameter and accounts for 67% of the variance on the test set performance.
- We observe there is a sweet-spot at the higher end of learning rate, where the performance is good and the training time is small.


TIMIT
IAM Online
JSB Chorales
Classification Error
Character Error Rate
Negative Log Likelihood
total time in h
learning rate
hidden size
input noise standard deviation
error
time

# Hyperparameter Analysis:

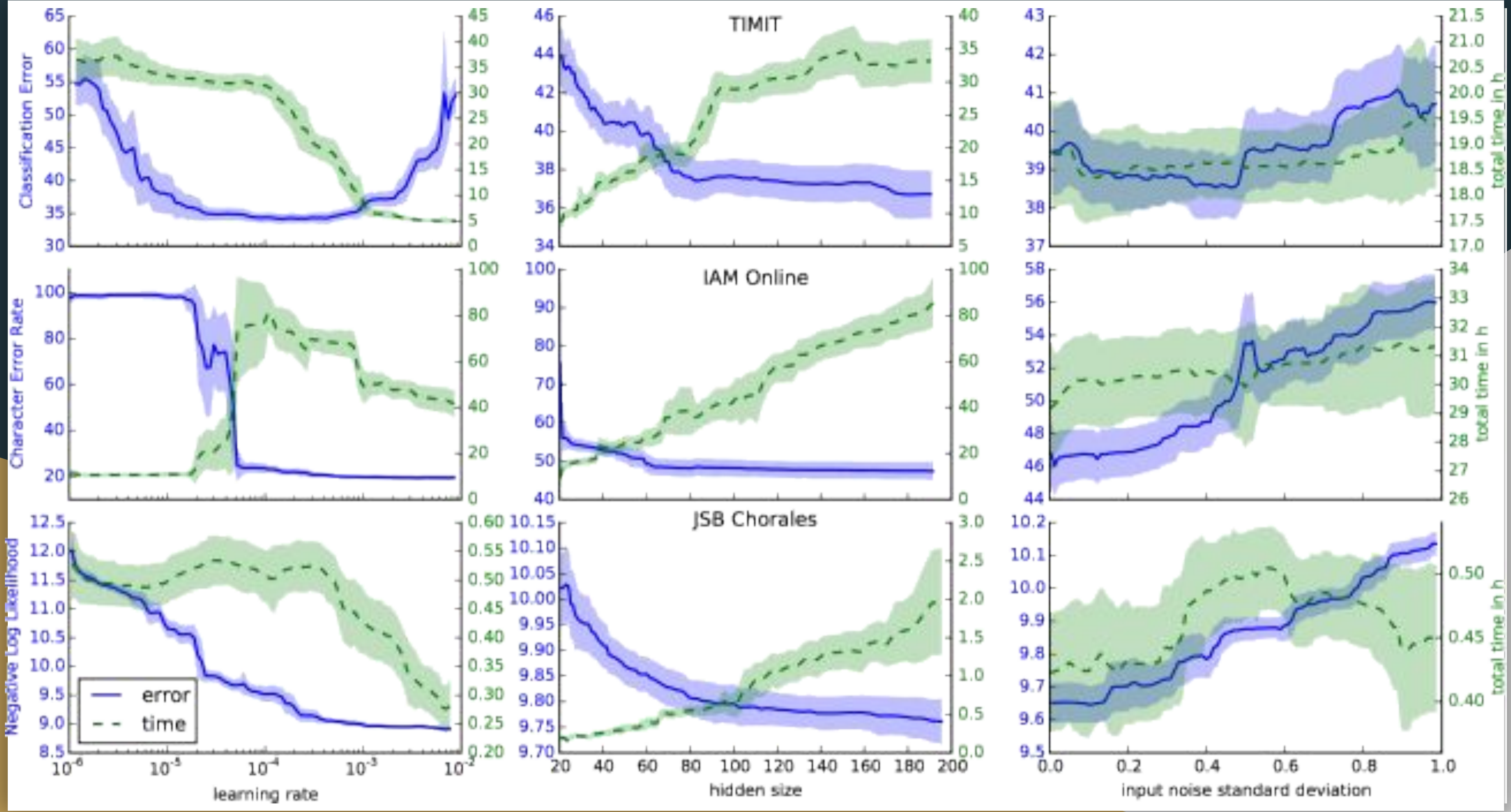
- Hidden Layer Size: Not surprisingly the hidden layer size is an important hyperparameter affecting the LSTM network performance. As expected, larger networks perform better.
- It can also be seen in the figure that the required training time increases with the network size.



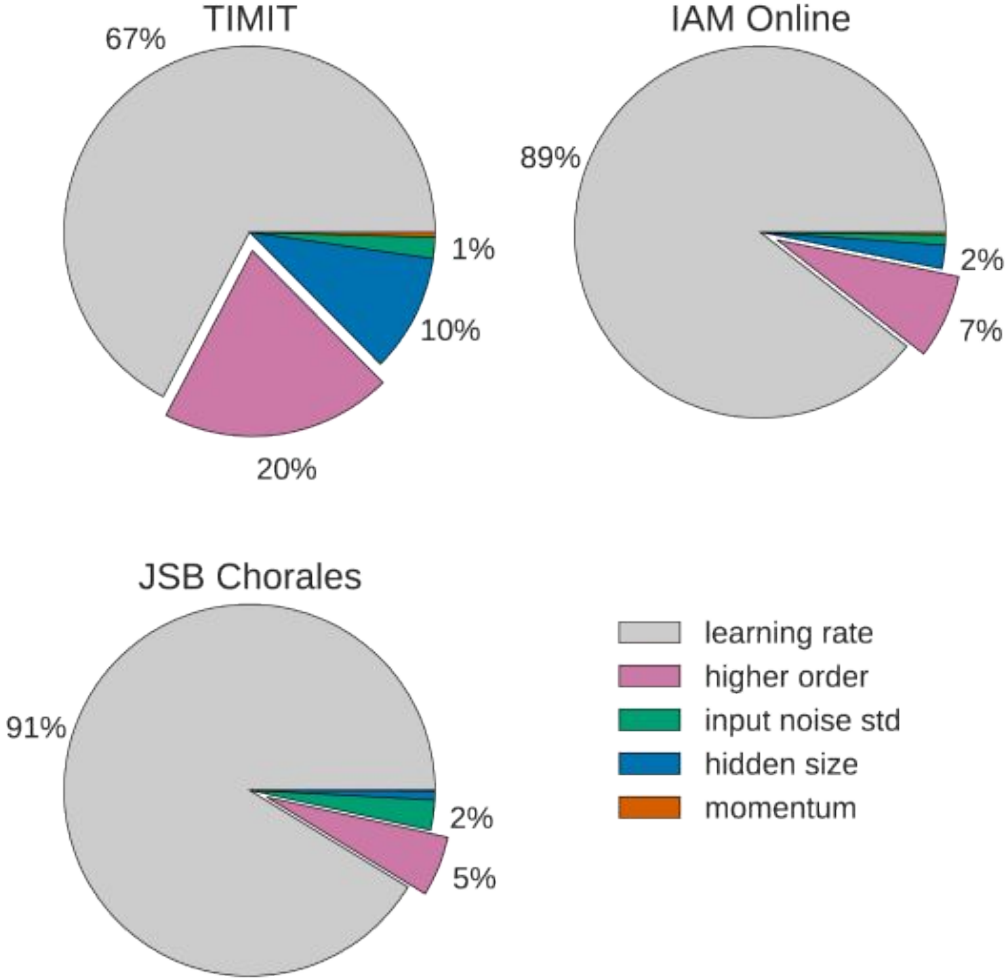
TIMIT
IAM Online
JSB Chorales
Classification Error
Character Error Rate
Negative Log Likelihood
total time in h
error
time
learning rate
hidden size
input noise standard deviation

# Hyperparameter Analysis:

- Input Noise: Additive Gaussian noise on the inputs, a traditional regularizer for neural networks, has been used for LSTM as well. However, we find that not only does it almost always hurt performance, it also slightly increases training times. The only exception is TIMIT, where a small dip in error for the range of [0:2; 0:5] is observed.



TIMIT
IAM Online
JSB Chorales
Classification Error
Character Error Rate
Negative Log Likelihood
total time in h
learning rate
hidden size
input noise standard deviation
error
time



TIMIT
67%
1%
10%
20%
IAM Online
89%
2%
7%
JSB Chorales
91%
2%
5%
learning rate
higher order
input noise std
hidden size
momentum

# Conclusion:

- We conclude that the most commonly used LSTM architecture (vanilla LSTM) performs reasonably well on various datasets.
- None of the eight investigated modifications significantly improves performance. However, certain modifications such as coupling the input and forget gates or removing peephole connections, simplified LSTMs in our experiments without significantly decreasing performance.

- The forget gate and the output activation function are the most critical components of the LSTM block. Removing any of them significantly impairs performance.
- The learning rate (range: log-uniform samples from [10^-6; 10^-2]) is the most crucial hyperparameter, followed by the hidden layer size( range: log-uniform samples from [20; 200]).
- The analysis of hyperparameter interactions revealed no apparent structure.

# THANK YOU